

Occupational Hygiene

Occupational Hygiene

EDITED BY

Kerry Gardiner

BSc, PhD, Dip. Occup. Hyg., FFOH

Professor of Occupational Health

The Medical School

The University of the Witwatersrand

Johannesburg

South Africa

and

International Occupational Health Ltd

Edgbaston

Birmingham, UK

J. Malcolm Harrington

CBE, BSc, MSc, MD, FRCP, FFOMI, MFPH, FMedSci

Emeritus Professor of Occupational Medicine

The University of Birmingham

Edgbaston

Birmingham, UK

THIRD EDITION



© 2005 by Blackwell Publishing Ltd
Blackwell Publishing, Inc., 350 Main Street, Malden, Massachusetts
02148-5020, USA
Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK
Blackwell Publishing Asia Pty Ltd, 550 Swanston Street, Carlton,
Victoria 3053, Australia

The right of the Authors to be identified as the Authors of this Work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

First published 1980
Second edition 1995
Reprinted 2002
Third edition 2005

Library of Congress Cataloging-in-Publication Data

Occupational hygiene / edited by K. Gardiner, J.M. Harrington.— 3rd ed.
p. ; cm.

Includes bibliographical references and index.

ISBN-10: 1-4051-0621-2

ISBN-13: 978-1-4051-0621-4

1. Industrial hygiene. 2. Industrial toxicology.

[DNLM: 1. Occupational Health. 2. Accidents, Occupational—prevention & control.

3. Occupational Diseases—prevention & control. WA 440 01497 2005] I. Gardiner,

K. II. Harrington, J. M. (John Malcolm)

RC963.0224 2005

613.6'2—dc22

2004025327

A catalogue record for this title is available from the British Library

Set in 9.5/12 pts Sabon by Kolam Information Services Pvt. Ltd, Pondicherry, India

Printed by Gopsons Papers, Noida, India

Commissioning Editor: Alison Brown

Editorial Assistant: Claire Bonnett

Production Editor: Fiona Pattison

Production Controller: Kate Charman

For further information on Blackwell Publishing, visit our website:

<http://www.blackwellpublishing.com>

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

Contents

List of contributors, vii

Preface, ix

Part 1: Introduction

- 1 Occupational hygiene 3
Kerry Gardiner
- 2 Global strategies and trends in occupational health: well-being at work in focus 6
Bengt Knave

Part 2: Organ structure and function and the adverse effects of work

- 3 The structure and function of the lungs 13
J. Malcolm Harrington and Anthony J. Newman-Taylor
- 4 Organ structure and function: the skin 25
Iain S. Foulds
- 5 Musculoskeletal disorders 36
Grahame Brown
- 6 The effects of inhaled materials on the lung and other target organs 47
Jon G. Ayres
- 7 The effects of some physical agents 59
Philip Raffaelli
- 8 Toxicology 67
Julian Delic, Steven Fairhurst and Maureen Meldrum

Part 3: Principles of occupational hygiene

- 9 The nature and properties of workplace airborne contaminants 85
Lisa M. Brosseau and Claudiu T. Lungu

- 10 Principles of risk assessment 105
Steven S. Sadhra
- 11 Design of exposure measurement surveys and their statistical analyses 124
Hans Kromhout, Martie van Tongeren and Igor Burstyn
- 12 Retrospective exposure assessment 145
Tom J. Smith, Patricia A. Stewart and Robert F. Herrick
- 13 Biological monitoring 160
Tar-Ching Aw
- 14 Epidemiology 170
J. Malcolm Harrington

Part 4: Environmental hazards: principles and methods of their assessment

- 15 The sampling of aerosols: principles and methods 185
David Mark
- 16 The sampling of gases and vapours: principles and methods 208
Richard H. Brown
- 17 Noise 222
Kerry Gardiner
- 18 Vibration 250
Michael J. Griffin
- 19 Light and lighting 268
N. Alan Smith
- 20 The thermal environment 286
Antony Youle
- 21 Non-ionizing radiation: electromagnetic fields and optical radiation 307
Philip Chadwick

22 Ionizing radiation: physics, measurement,
biological effects and control 328

Ronald F. Clayton

23 Biological agents 344

Julia M. Greig and Chris J. Ellis

24 Psychological issues 360

Anne Spurgeon

25 The development of ergonomics as a
scientific discipline 373

Joanne Crawford

26 Dermal exposure assessment 389

John W. Cherrie

Part 5: Allied and emerging issues

27 Occupational accident prevention 403

Richard T. Booth and Anthony J. Boyle

Part 6: Control

28 Work organization and work-related
stress 421

*Tom Cox, Amanda Griffiths and
Stavroula Leka*

29 Control philosophy 433

Kerry Gardiner

30 Ventilation 440

Frank Gill

31 Personal protective equipment 460

Robin M. Howie

32 Occupational health and hygiene
management 473

Lawrence Waterman and Karen Baxter

Contributors

Tar-Ching Aw

Division of Occupational Health
Kent Institute of Medicine and
Health Sciences
University of Kent
Canterbury, UK

Jon G. Ayres

Department of Environmental
and Occupational Medicine
University of Aberdeen
Aberdeen, UK

Karen Baxter

Sypol Limited
Aylesbury, UK

Richard T. Booth

Health and Safety Unit
School of Engineering
Applied Science
Aston University
Birmingham, UK

Anthony J. Boyle

Priors Marston
Warwickshire

Lisa M. Brosseau

University of Minnesota – School of
Public Health
Division of Environmental
Health Sciences
Minneapolis, MN
USA

Grahame Brown

Royal Orthopaedic Hospital NHS Trust
Northfield
Birmingham, UK

Richard Brown

Research & Laboratory
Services Division
HSE
Sheffield, UK

Igor Burstyn

Department of Public Health Sciences
University of Alberta
Edmonton
Canada

Philip Chadwick

Microwave Consultants Limited
Newbury
Berkshire, UK

John W. Cherie

Institute of Occupational Medicine
Riccarton
Edinburgh, UK

Ronald F. Clayton

Abingdon, UK

Tom Cox

Institute of Work, Health and Organisation
University of Nottingham
Nottingham, UK

Joanne Crawford

University of Birmingham
Institute of Occupational and
Environmental Medicine
Edgbaston
Birmingham, UK

Julian Delic

HSE
Magdalen House
Bootle, UK

Chris J. Ellis

Department of Infection
Heartlands Hospital
Bordesley Green
Birmingham, UK

Steven Fairhurst

HSE
Bootle, UK

Iain S. Foulds

Birmingham City Hospital
Birmingham Skin Centre
Birmingham, UK

Kerry Gardiner

The Medical School
The University of Witwatersrand
Johannesburg
South Africa
and
International Occupational Health Ltd
Edgbaston
Birmingham, UK

Frank Gill

Southsea
Portsmouth, UK

Julia M. Greig

Department of Infection
Heartlands Hospital
Bordesley Green
Birmingham, UK

Michael J. Griffin

Human Factors Research Unit
Institute of Sound and Vibration
Research
University of Southampton
Southampton, UK

Amanda Griffiths

Institute of Work, Health and
Organisation
University of Nottingham
Nottingham, UK

J. Malcolm Harrington

Emeritus Professor of Occupational
Medicine
The University of Birmingham
Edgbaston
Birmingham, UK

Robert F. Herrick

Department of Environmental
Health
Harvard School of Public Health
Boston, MA
USA

Robin M. Howie

Robin Howie Associates
Edinburgh, UK

Bengt Knave

National Institute for Working Life
Stockholm
Sweden

Hans Kromhout

Institute for Risk Assessment Sciences
Utrecht
Netherlands

Stavroula Leka

Health Psychology
Institute of Work, Health and
Organisation
University of Nottingham
Nottingham, UK

Claudiu T. Lungu

University of Minnesota – School of
Public Health
Division of Environmental
Health Sciences
Minnesota, MN
USA

David Mark

Health and Safety Laboratory
Exposure Co.
Sheffield, UK

Maureen Meldrum

HSE
Bootle, UK

Anthony J. Newman-Taylor

Royal Brompton Hospital
National Heart and Lung Institute
Imperial College School of Medicine
London, UK

Philip Raffaelli

Chief Executive
Headquarters Defence Medical
Education and Training Agency
Gosport, UK

Steven S. Sadhra

Institute of Occupational and
Environmental Medicine
University of Birmingham
Edgbaston
Birmingham, UK

Tom J. Smith

Environmental Health
Harvard School of Public Health
Boston, MA
USA

N. Alan Smith

Institute of Occupational and
Environmental Medicine
The Medical School
University of Birmingham
Birmingham, UK

Anne Spurgeon

University of Birmingham
Institute of Occupational and
Environmental Medicine
Edgbaston
Birmingham, UK

Patricia A. Stewart

Occupational Studies Section
National Cancer Institute
Rockville, MD
USA

Martie van Tongeren

Centre for Occupational and
Environmental Health
University of Manchester
Manchester, UK

Lawrence Waterman

Sypol Limited
Aylesbury, UK

Antony Youle

Robens Centre for Occupational
Health and Safety, EIHMS
University of Surrey
Guildford, UK

Preface

A quarter of a century has passed since Tony Waldron first suggested to one of us (JMH) that there was no available text for occupational hygienists. With much temerity and some trepidation, two physicians attempted to fill that gap – with some apparent success. A second edition added an occupational hygienist (KG) as editor. That edition sought to broaden the subjects covered as the discipline of occupational hygiene became more clearly defined and the need for skills in recognising, evaluating and controlling work environments became more pressing.

Major developments in the way in which occupational hygiene is taught and practised have followed in the past decade. This period of time has also witnessed dramatic changes in the world of work and, with it, the nature and type of health problems encountered at work. Whilst many of the classic occupational diseases unfortunately still exist in the world's poorer economies, in the developed world the main challenges now are musculoskeletal disorders and psychosocial problems. These appear to be consequent upon the rapid and continuing changes taking place in work organization.

At the same time, the boundaries between occupational medicine and occupational hygiene have become blurred as increasing emphasis has been placed on team solutions to the ever more complex work environment. We have attempted to include these newer issues in this edition and we hope that the work/health problems we, as professionals, attempt to control are appropriately and adequately covered in this revised and expanded text.

Kerry Gardiner, Birmingham
Malcolm Harrington, London
January 2005

Part 1

Introduction

Chapter 1

Occupational hygiene

Kerry Gardiner

Background
Challenges
Political/financial context

The future, our contribution
Reference

Background

The essential tenets of occupational hygiene as a discipline (i.e. separating people from unpleasant/deleterious situations/exposures) have been known for centuries, but, as a profession, occupational hygiene is in its relative infancy. There is much debate about the first recognition of the profession, whether it was first described in Agricola's treatise in the mid-sixteenth century or resulted from the fractious discourse between the Bureau of Labor Standards (which wanted to create good working conditions) and the US Public Health Service (which focused on the *post hoc* recognition of disease by means of clinical examination) (Carter, 2004). This difference in emphasis (a priori prevention/control versus *post hoc* diagnosis) has been pivotal in the progression (or otherwise) of occupational health/hygiene over the years.

Albeit in simplistic terms, definitions of occupational hygiene (all of which contain most, if not all, of the following: anticipation, recognition, evaluation and control) are little different in overall philosophy from any other part of occupational health. However, as alluded to above, what has changed in more recent years is a paradigm shift away from the belief that it is best practice to diagnose disease and try and identify a cause (often against a background of limited opportunity for treatment), towards a culture of risk mitigation/moderation by utilization of assessments of the working environment in advance of any harm being caused (Fig. 1.1).

Challenges

One area in which there is consensus amongst the relevant professions is that of the breadth of disciplines and/or skills required to address adequately modern working environments and the people who work within them. This is evident in terms of both the recognition of the various aspects that affect an individual's well-being outwith work (i.e. genetics, environment, housing, diet, pre-existing morbidity) and the complexity of modern working environments [i.e. electromagnetic fields, working hours/shift work, psychosocial hazards (physical verbal abuse at or during work), chemical risks] (Fig. 1.2). The resultant benefit is that there appears to be greater recognition that the professions must share their knowledge/expertise, that each of the disciplines can indeed address areas that historically were demarcated to be their sole preserve and that each, and all, have an essential role to play in the prevention/reduction of ill health at work.

An additional challenge to all occupational health professionals/disciplines (including occupational hygiene) is that there is a state of flux regarding the nature of work and the social contracts entered into by each individual/group. There is a marked move away from the historic manufacturing industries in the developed world towards very specialized/technical work employing few but generating many risks along with the proliferation of the 'service industry' and all that it encompasses. This goes some way to explaining the dynamic

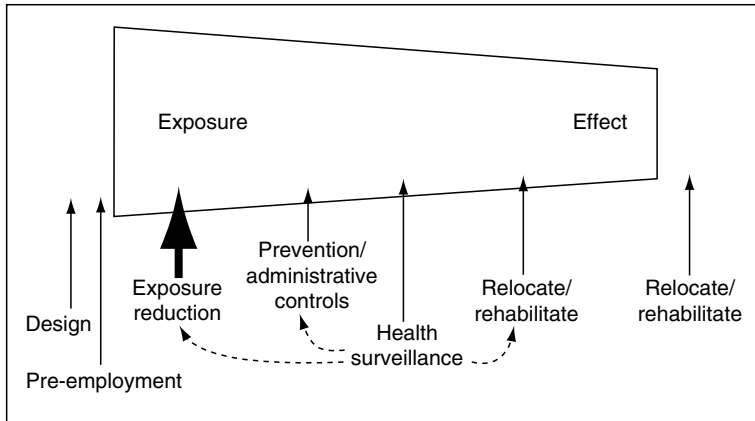


Figure 1.1 The importance of achieving control before the manifestation of disease.

shift of the hazard groups but implies so much more. In many societies, despite the absence of classic ‘chemical’ exposures, work still has a massive impact on people’s lives, from the individuals who regularly work 80-hour weeks to the employees in call centres, who are frequently abused verbally in the course of their work. The implied corollary is that of ‘technology export’, wherein processes deemed too hazardous, too restricted or simply too expensive are moved to countries where these factors are not or need not be a problem. The irony is that it is very common for such large-scale technology transportation to import standards of living, general or occupational health care and

social infrastructure not otherwise available or possible.

To complicate matters further, there is real concern in the Western world of the impact of such social phenomena as ageing workforces (differential impact of the hazards by age, protracted periods of exposure, the psychological desire to have retired at a reasonable age, etc.), peripatetic workforces, in which the expectation of both the employer and employee is that they will not remain employed in any one company/place for very long (thereby exacerbating the negative view of investing in training, the benefits of providing good occupational health care, the ability to undertake

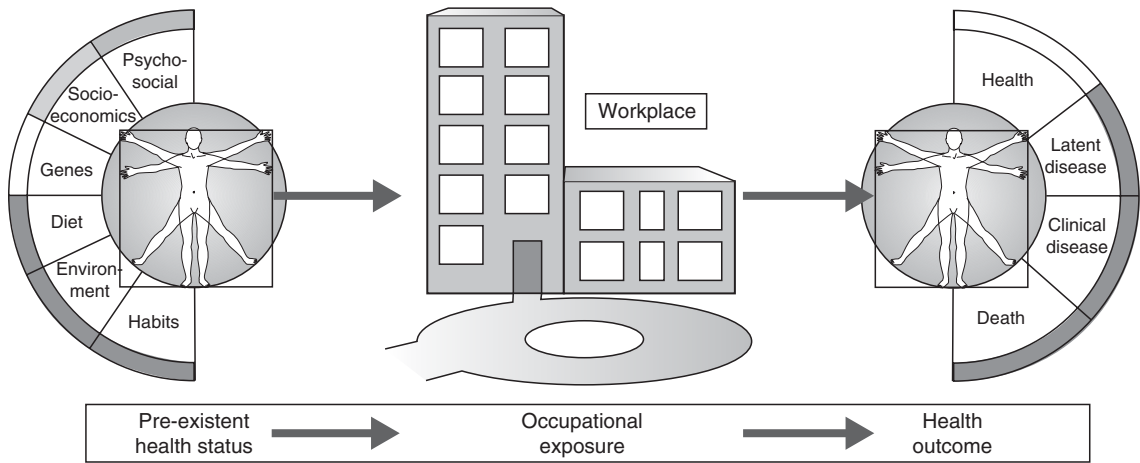


Figure 1.2 The multifactorial nature of occupational health/hygiene.

quality epidemiology, etc.), and migrant workforces (limited knowledge of previous occupational history).

It is exactly to address this ever-growing field that this book has grown not only in size since its first and second editions but also in the breadth of the subject areas included. Clearly, the chapters are not an exclusive list but are considered by the editors to be the more pivotal aspects of the discipline.

Political/financial context

In many ways, the effect of such ‘technical/professional’ issues is entirely dependent upon the political and financial context within which occupational health operates. Despite many governments around the world utilizing persuasive rhetoric to commit themselves to occupational health – in the form of well-being at work, valuing the country’s greatest asset, its people and so on – this stance is often not put into practice.

Notwithstanding the fact that non-compliance with occupational health statute is a breach of criminal law in many countries, it is often of great surprise to those not familiar with the discipline that it should be ‘policed’ in the way in which it is, with the enforcing authorities usually hugely underfunded and resourced, e.g. in the UK, at or around the millennium, the Health and Safety Executive received ~£200 million, whereas tax revenues exceeded £3.2 billion and the police force received £46 billion; no inference is made vis-à-vis the relative importance but their difference is profound. Despite the crudeness of these figures, they ably demonstrate the stark contrast between how society, through its elected representatives, views, and ultimately values, the well-being of its workforce and other sectors of the community. An

American colleague once recounted a wonderfully apt phrase in relation to the way in which organizations/agencies/government departments are provided with sufficient funds to *exist* but explicitly not to be *effective* – ‘*funded to fail*’.

The future, our contribution

The text above is meant to provide the reader solely with a context within which to read the following chapters and, also, if the reader is a student within the discipline or profession, some guidance regarding the context within which he or she will have to operate. It is *not* meant to be negative in any way but to create a backdrop to highlight quite how important ‘our’ field is; in fact, in many ways, those that utilize the skills/knowledge of occupational hygiene (whether a professional, sister-discipline or interested layperson) are as vital today as they were 100 years ago – it is just that the rules have changed.

The greatest contribution that occupational hygienists, and those that utilize their skill/knowledge base, can achieve is that they enable potentially ‘risky’ processes to be undertaken in relative safety. Society and individuals thrive on risk and risk-taking, and those practising good occupational hygiene principles facilitate this to continue; after all, the biggest societal impact concerning work is having no job at all. If we can ensure that the workforce is ‘Healthy, Happy and Here’ then we will have succeeded.

Reference

- Carter, T. (2004). British occupational hygiene practice 1720–1920. *Annals of Occupational Hygiene*, 48, 299–307.

Chapter 2

Global strategies and trends in occupational health: well-being at work in focus

Bengt Knave

The changing work life
Stress, downsizing of companies and unemployment
Lessons from the past – challenges for tomorrow
Men and women at work
Child labour
The integrated occupational health concept

But don't forget the 'old' and well-known hazards
New areas coming into focus
Occupational health – today and prospects for tomorrow

The changing work life

During the past decades, work life has undergone great changes. Not only work itself, but also our opinion of work, has changed. Now we know that a good work environment should not only be healthy and safe, but should also encourage personal and professional development, job satisfaction and personal fulfilment, all of which contribute to improved work quality and productivity. We know that the way work is organized is of importance, and that the situation in the labour market affects the work, the worker and the worker's health and well-being. So, there are different aspects of this development of work life. One of the most important and common occupational health problems today is 'stress'.

Stress, downsizing of companies and unemployment

In a recent study on stress symptoms, 'a feeling of general irritation' was found to be most common, followed by headaches, depression and sleep difficulties. On questioning workers on which were the eliciting stress factors, 'high workload' was the most common response, followed by 'short delivery times', 'no influence at work', 'no support from manager', 'long working hours' and 'worry about job security'.

During the 1990s, especially during the first half of the decade, general welfare in many countries was influenced by a severe, world-wide economic recession. Many companies had to downsize and unemployment rates increased dramatically.

It is well known that unemployment is traumatic. However, workers still in employment also suffer the effects of high rates of unemployment. Thus, as a consequence of the recession, enterprises downsized and work pace and work stress increased. Nobody complained because of the risk of job loss. Paradoxically, rates of sick leave diminished – when workers were sick, they nevertheless went to work because of fear of losing their job. During the late 1990s, the labour market in many countries recovered and the rate of unemployment fell steadily; unemployment has now reached a politically acceptable low level. However, the latest statistics on reported work-related injuries and diseases show a marked increase. Among the work-related diseases, those indicated to be caused by organizational and social factors have increased the most, and considerably more than one-half of these cases are attributed to stress.

Lessons from the past – challenges for tomorrow

It is interesting to note the observations of a historian on this subject (Johannisson, 2001).

Johannisson points out the similarities between the symptoms of what we see happening today among employees suffering from stress and what happened 100 years ago in the transition phase between the farming and industrial societies. Today we are living through transition: one between the industrial society and the information and computer technology (ICT) society. In both cases, large population groups have to adapt to work life skills and experiences that are quite new, which may be difficult for many of us without proper education and training, and which may result in stress reactions.

So, what is the remedy today for stress at work? Let me quote a former Director General of the Swedish Social Insurance Board (Sherman, 2002).

... the increased stress and sick leaves are a reaction of what happened in the 1990s when companies were 'slimmed', and work intensities increased. Nobody could expect me to believe that the 325 000 employees more on sick leave today than five years ago, are sick in an objective sense. This does not mean that they should be sent back to work, unless we change the work itself. Burn-out and stress are symptoms of a diseased society, where people have been pressed over their capacities.

I agree with Sherman and hope that society will take prompt measures to start recovering from its disease.

Men and women at work

We know today that women – more than men – are frequently affected by injuries and sickness caused by a poor work environment. In recent years, the prevalence of stress-related health problems has increased markedly among women in particular. This gender difference is linked to women's changing roles and greater participation in the paid workforce, without a corresponding reduction in unpaid work (household work, child care, etc.). Furthermore, regardless of country and continent, women are paid between 5% and 50% less than men are paid for the same work. The situation was summarized at the 'Work, Women and Health' Conference in June 2002 as follows (Lundby

Wedin, 2002): 'Everywhere you turn in the world you will find that it is the women who are the poorest, who have the lowest wages, who have the worst working environment and who get the worst pension deals'.

We need more data to describe women's and men's working conditions. Women often have a mixture of illnesses and so-called vague symptoms such as fatigue, reduced vitality, feelings of insufficiency, different pains and discomfort. These symptoms are associated with psychological causes and, up until now, according to some researchers, they have not been taken as seriously as other somatic symptoms. In turn, this means that women's health problems at work are not as visible as those of men, and, as a result, they are neglected both in research and in practice.

Child labour

Child labour has been, and still is, a world-wide problem. Involuntary underage workers typically forfeit the chances of developing knowledge and gaining education as other children do, and they risk their health and welfare, under duress, in the cause of commercial gain for others or simply for their own and their families' survival.

An example comes from Bangladesh, where child labour is a rapidly growing phenomenon of concern (Rahman, June 2002, Institute of Child and Mother Health, Bangladesh). The total number of children in labour approaches seven million, i.e. 20% of the total child population aged 5–14 years, of whom 96% are employed in informal sectors. Most of the children are compelled to engage themselves in dangerous and hazardous occupations, and many of the children suffer from injuries. In the Bangladesh study (covering the transport sector and small metallic manufacturing enterprises), the 1-month prevalence morbidity was almost 40%, and the proportional morbidity from injuries was almost 50%. Cutting of finger/hand constituted 52% of the injuries, fractures 8%, sprains 9% and bruising 14%. In total, 40% of these injured children did not receive any kind of treatment, 23% consulted doctors and 13% were sent to hospital.

This study is referred to in detail to show the magnitude of the problem, and to show how self-evident it must be for all international occupational health organizations to engage themselves wholeheartedly in the fight against child labour. Furthermore, it is important to understand how the meaning of the word 'well-being' varies according to whether you happen to be born in a rich or a poor country.

The integrated occupational health concept

At the beginning of the twentieth century, occupational health was a matter for physicians. By and by, however, new groups became involved: nurses, engineers and hygienists. Today, topics within ergonomics and work organization are included, and the 'integrated' occupational health area even extends to labour market issues.

It is easy to understand the inter-relationships between medicine, hygiene and ergonomics. Work organization defines the contents of the work and how it is distributed among the employees ('right person for the right work'). 'Theoretical' work organization encompasses, for instance, where and when the employee comes in the decision hierarchy; possibilities for lifelong learning; and evaluation of workloads and risks for ill health. More practical 'projects' may include 'stress and health', 'conditions for human service work', 'industry and the human resource', 'gender and work', 'lifelong learning' and 'work and culture'.

Labour market issues are somewhat more peripheral, however important for the employee. Examples of practical projects are 'job creation', 'labour law' and 'social economics'. Some of the labour market topics overlap with work organization, which in turn overlaps with ergonomics, which in turn overlaps with medicine, etc.

There are reasons to believe that the integrated occupational health concept will play a leading role in the future. The development in work life matters within the European Union (EU) points in this direction. Employability, entrepreneurship, adaptability and equal opportunities are the four 'pillars' in the EU 1998 guidelines.

But don't forget the 'old' and well-known hazards

It can readily be seen from the above that occupational health today is a growing field, covering a wide range of different topics, in which research, practice and prevention go hand in hand. Overall, in addition to stress, musculoskeletal diseases and asthma and other allergies are the most prevalent work-related diseases, while fatal accidents at work still claim a large number of lives each year world-wide.

However, we must not forget that there are one million different chemical compounds in our immediate environment; several hundred new chemicals are introduced each year into the environment; the long-term toxicity of extremely low concentrations of these chemicals is unknown. New work technologies mean exposure to 'new' physical factors, such as electromagnetic fields (EMFs), the full effects of which we do not know. Recent assessments by the International Agency for Research on Cancer (IARC) now classify low-frequency EMFs as possible risk factors for childhood leukaemia, and the health risks of using mobile phones are under evaluation.

Infections constitute another important health area, and special emphasis is directed, for self-evident reasons, towards the HIV/AIDS epidemic. Forty million people in the world are now infected ['what we now witness is a world historic epidemic larger than the Black Death and the Spanish flu' (Kallings, July 2002, Secretary General, International AIDS Society)], and in some parts of the world (sub-Saharan Africa, Eastern Europe and South-East Asia) the epidemic is still spreading. The consequences of the epidemic in the most affected countries are devastating; the workforce is dying, national production is falling, the level of education is declining and poverty is increasing. It is quite clear that one of the main missions of occupational health workers world-wide has to be the prevention of this epidemic. Promising workplace interventions including education about preventing transmission have been implemented in some of the sub-Saharan Africa countries, and services have also been extended to the wider community. But more has to be done.

New areas coming into focus

As mentioned, work life is undergoing a process of continuous change. New ‘problem’ areas can be expected come to the fore. Some of these may always have existed but, for various reasons, been ignored by society not out of any ill-will, but perhaps simply out of ignorance. Three such areas are:

- sexual harassment;
- physical violence; and
- bullying.

All of these are in some way related to each other. Recent studies conducted by the European Foundation for the Improvement of Living and Working Conditions (under the European Commission) have shown these problems to be relatively common; the prevalence of the first two has been estimated to be between 5% and 10%, and the prevalence of the last (bullying) between 10% and 15%.

Occupational health – today and prospects for tomorrow

Since the mid-1990s there has been a decline in occupational health activities in many ‘developed’

countries. The main reason has been a world-wide economic recession leading to ‘slimmed’ national budgets and increasing unemployment rates, with occupational health services (OHS) being one of the first areas to suffer. An additional contributing factor could be the increasing age of those who work within OHS and, consequently, an increasing number of people have retired and have not been replaced. Yet another factor of possible importance is ‘the changing world of work’, with a transition from an industrial society to an ICT society, with OSH still being largely considered a function of industry.

However, there are signs pointing in the right direction. In many countries, for instance, there is a political revival of positive interest in OHS. Furthermore, organizations such as the ICOH (International Commission for Occupational Health) are attracting new, young members from different parts of the world. With what we have today, there are reasons to be positive about the future. Work life is constantly changing, which is nothing but a challenge for OHS working for health and well-being at work.

Part 2

Organ structure and function and the adverse effects of work

Chapter 3

The structure and function of the lungs

J. Malcolm Harrington and Anthony J. Newman-Taylor

Introduction

Structure

- The airways
- The acinus
- The consequences of branching
- The alveolus
- The blood–gas barrier
- The lining of the airways
- The blood supply to the lungs
- Lymphatics
- The pleura

Function

- Ventilation
- Gas transfer
- Blood–gas transport
- Factors influencing lung function
 - Age
 - Smoking
- Other functions
 - Occupational lung disease and disordered function
- Further reading

Introduction

The primary function of the lungs is to secure the exchange of gas – oxygen and carbon dioxide – between air and blood. The structure of the lungs enables this function by conducting air through a series of branching tubes (bronchi) to the sites of gas exchange, the alveoli.

Structure

The lungs are composed of a number of topographical units called bronchopulmonary segments, which are roughly pyramidal in shape, with their apices directed inwards and their bases lying on the surface of the lung. Each lung is composed of 10 such segments grouped together into lobes. The right lung has three distinct lobes: upper, middle and lower; the left lung has two lobes, with a rudimentary third lobe (the lingula) incorporated into the upper lobe. Each lower lobe contains five segments: the right middle lobe (and the lingula), two; and the upper lobes, three.

The airways

Air is conducted into the lungs through a system of branching tubes. The first of these is the trachea, which is attached above to the larynx. It is a tube about 15 cm long, lying directly in front of the oesophagus, and held open by C-shaped rings of cartilage in its walls. At the back, it is joined together by a sheet of muscle fibres (the trachealis muscle), which, by contraction, reduce the diameter of the tube. The muscle can prevent overdistension of the trachea when the internal pressure is raised, for example during the act of coughing.

The trachea divides into the left and right main bronchi, which themselves branch into the segmental bronchi (Fig. 3.1) and further branching continues within each segment. A bronchus is defined as an airway that contains cartilage within its walls; divisions of the airways that contain no cartilage are termed bronchioles. The *smaller* bronchi contain less cartilage in their walls than do the trachea and the large bronchi. Their walls, and those of the bronchioles, are supported by elastic fibres and also by dense connective tissue. Moreover, they have two spiral layers of smooth muscle fibres surrounding their lumen, which

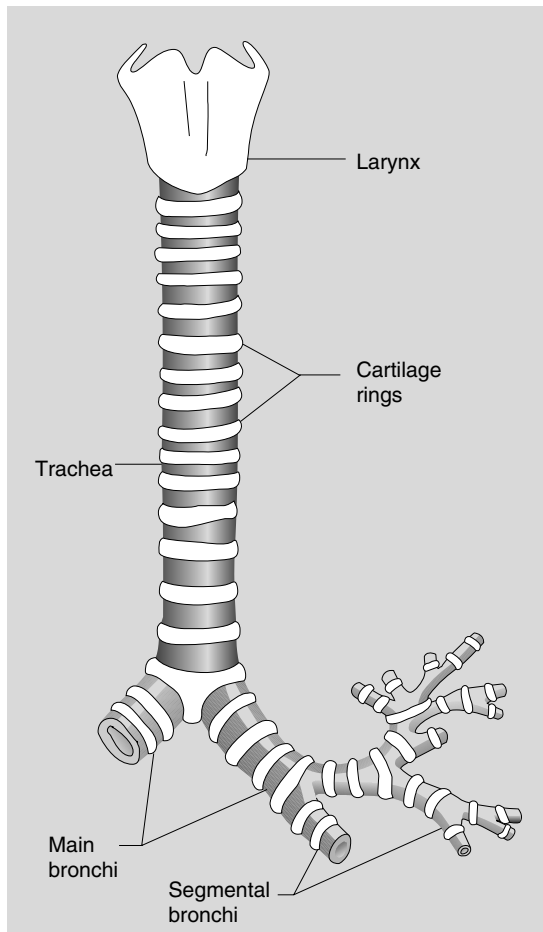


Figure 3.1 The main airways.

contract under the influence of nervous (or other) stimuli, thus lessening the calibre of the airway and altering the rate of flow of air into and out of the lungs.

It is usual to describe an airway in terms of the number of divisions, or generations, which separate it from the main bronchus. The segmental bronchus is counted as the first generation, and its first branch as the second generation, and so on. In this classification, the bronchi comprise about 15 generations, the first five of which are ‘large’ bronchi, having a plentiful supply of cartilage in their walls. The sixth to fifteenth generations are ‘small’ bronchi whose walls contain only a sparse amount of cartilage. There are some 10 generations of bronchioles, whose walls do not contain cartilage,

which ultimately open into the alveolar ducts from which the alveoli originate. The bronchiole that opens into the alveolar duct is defined as a respiratory bronchiole, the one immediately proximal to it as a terminal bronchiole.

The acinus

The acinus is that part of the lung distal to a terminal bronchiole. It includes up to eight generations of respiratory bronchioles and their associated alveoli. It is approximately 0.5–1 cm in diameter.

The lobule is the term used to describe the three to five terminal bronchioles, together with their acini, which cluster together at the end of an airway (Fig. 3.2).

The consequences of branching

The consequence of branching within the airways is to greatly increase the total cross-sectional area of the airways, decreasing the rate of airflow through them. By the time air has reached the terminal bronchioles, bulk flow of air has ceased, leaving diffusion down concentration gradients as

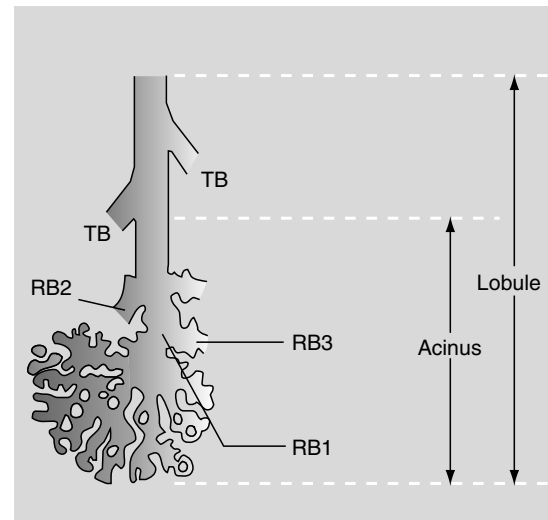


Figure 3.2 An acinus. TB, terminal bronchiolus; RB, respiratory bronchiolus. Note that RB has several generations (RB1, RB2, RB3, ...).

Table 3.1 Cross-sectional area of the airways and rate of airflow.

	Area (cm ²)	Flow rate (cm s ⁻¹)*
Trachea	2.0	50
Terminal bronchioles	80	1.25
Respiratory bronchioles	280	0.36
Alveoli	10–20 × 10 ⁵	Negligible

*For a volume flow of 100 ml s⁻¹.

the means of transport. This has relevance and importance for dust deposition. This is well illustrated by the data in Table 3.1.

The alveolus

The alveolus is the part of the lung in which gas exchange occurs. There are 200–600 million alveoli in the fully developed adult lung, which offer a surface area of some 100–200 m² over which gas can diffuse. The alveolus is about

250 μm in diameter and its walls are lined with two types of epithelial cells: the type 1 and type 2 pneumocytes. The alveolar wall is typically less than 0.5 μm thick, but the epithelial lining is continuous throughout, and is composed of the cytoplasm of the pneumocytes. The alveolar wall is surrounded by, and in intimate contact with, the endothelial cells of the capillaries of the pulmonary vascular bed (Fig. 3.3).

The type 1 pneumocyte is a large flat cell (like a fried egg!), covering a much greater area of the alveolar wall than the type 2 cell, although it is less numerous. The type 2 cell, which is usually found in corners of the alveolus, is a small cell containing characteristic lamellated inclusion bodies within its cytoplasm. These inclusion bodies are the origin of surfactant, a lipoprotein that lines the surface of the alveoli and acts to stabilize their size. The alveoli may be thought of as small bubbles, the surface tension of which is inversely related to the radius. The forces acting to keep the alveolar radius constant are less than one would predict, due

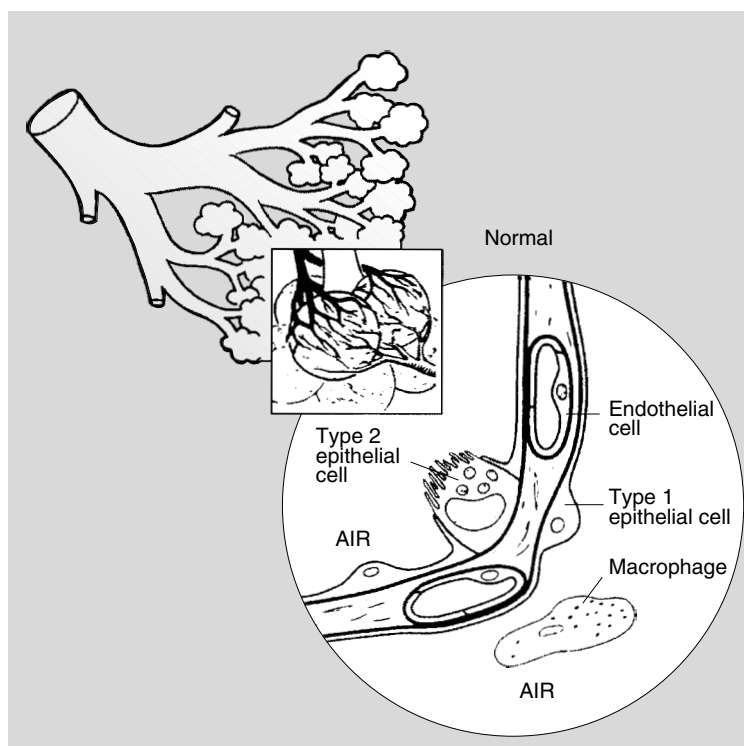


Figure 3.3 A section of alveoli, showing the relationship between the various cell types.

to the presence of surfactant, which reduces the surface tension within the alveolus. A number of other cells are also encountered within the alveoli, including connective tissue cells, blood cells and phagocytes, called alveolar macrophages, which are present in the alveolar space and may contain carbon pigment taken up from the smoke-laden air. The walls also contain elastic and collagen fibres, which provide support. Having taken up foreign material, macrophages travel up the airways aided by the cilia in the epithelial lining (see below) or are cleared via the lymphatics from the centre of the acinus. Upon reaching the larynx, they and their contents are usually swallowed but with mucus hypersecretion may be coughed out. Foreign material cleared from the lungs and swallowed may be subsequently absorbed from the gut.

The blood–gas barrier

The epithelial lining cells of the alveolar wall, together with the endothelial cells of the capillaries (each with their own basement membrane) and the tissue fluid in the spaces between, make up what is known as the blood–gas barrier. This is the ‘thin’ side of the alveolus where the basement membranes of type 1 epithelial cells and capillary endothelial cells are fused. It is the distance that gas molecules, or indeed any other material, must

cross when passing into or out of the alveolus. The total distance involved is probably less than 0.001 mm.

The lining of the airways

The airways are lined with a ciliated epithelium that contains a number of different types of cell (Fig. 3.4). In the large airways, the epithelium is pseudostratified, that is, it has the appearance of being composed of more than one layer. This is due to the fact that the basal cells do not reach the surface of the epithelium, although all the other cells do reach the basement membrane, which is a permeable layer of material synthesized by the cells, acting to stick the cells together. In the smaller airways there is only one layer of cells, each of which reaches the surface. The goblet cells secrete mucus, which spreads out to form a layer on the surface of the airway where it traps particulate matter. Mucus is also secreted by the submucosal glands and is constantly being moved up towards the larynx by the synchronous beating of the cilia, carrying with it material trapped in the airways, together with the alveolar macrophages containing material scavenged from the alveoli. This so-called mucociliary escalator is the most important mechanism through which the airways are cleared of particulate matter. Ciliated cells are found down

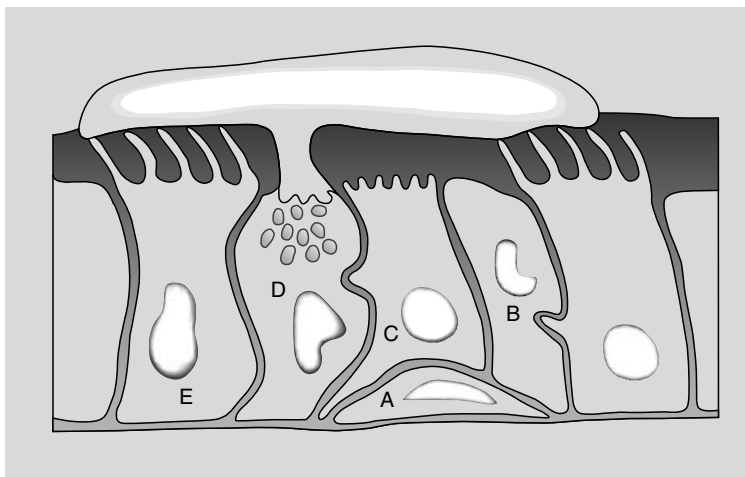


Figure 3.4 The epithelial lining of the trachea, with its various cell types. A, basal cell; B, non-ciliated cell; C, brush cell; D, goblet cell (secreting mucus); E, ciliated cell.

as far as the respiratory bronchioles so that this mechanism operates continuously from the bronchiolo-alveolar junction to the larynx.

The function of the brush cells is unclear at present, but the structure and dimensions of the microvilli on the surface of the cells, from which they take their name, suggest that they could be concerned with the absorption of fluids from the airways and hence with the control of fluid balance.

The blood supply to the lungs

The lungs have two arterial blood supplies and two sets of venous drainage. Their arterial supply is from the pulmonary and bronchial arteries, whereas venous blood is conducted away by the pulmonary and bronchial veins.

The left and right pulmonary arteries arise from the right ventricle of the heart and convey to the lungs blood that has been returned from the tissues into the right atrium. Thus, blood in the pulmonary arteries, unlike blood in other arteries, has a low partial pressure of oxygen. It is also under much less pressure than blood in the other arteries, the pressure in the pulmonary artery being about one-tenth of the pressure in the systemic arterial circulation.

Both airways and blood vessels are low-resistance systems that enable an equivalent flow of air and blood to either side of an alveolar capillary. The pulmonary artery supplies the capillary bed surrounding the alveoli and at any one time approximately 80 ml of blood is in contact with the alveolar air (out of the total of 400 ml in the pulmonary capillary circulation). The pulmonary artery divides with each airway divide but, in addition, other arteries arise to supply the alveoli around the airway. These additional arteries are called the supernumerary branches and outnumber the so-called conventional branches by about three to one.

The pulmonary veins run at the periphery of the bronchopulmonary segments and return blood with a high partial pressure of oxygen to the left side of the heart, where it is pumped into the systemic circulation.

The bronchial arteries are branches of the aorta and supply oxygenated blood to the capillary beds

in the walls of the airways. Blood from the capillaries drains back through the pulmonary veins, and the bronchial veins receive blood only from the large airways, the lymph nodes and the pleura, which they convey to the azygos vein for return to the right atrium.

Lymphatics

Lymph vessels convey interstitial fluid from the tissue spaces back to the thoracic duct, which opens into the left internal jugular and subclavian veins. After the mucociliary escalator, this drainage system from the centre of acini is the second major route by which alveolar macrophages, laden with dust particles, leave the lung. The lymph vessels drain at intervals through lymph nodes, which may be regarded, simply, as filters which trap particulate matter travelling in the lymph. Cells present in the lymph may also be arrested in the lymph nodes, and if these cells happen to have broken away from a tumour developing within the lung, they may divide and multiply within the nodes, causing them to enlarge.

Lymph vessels are abundant in the pleura, the connective tissue septa between the bronchopulmonary segments, and in the walls of the airways and blood vessels; they do not appear in the walls of the alveoli. The lymph drains from the periphery towards the hilum of the lung and through the nodes situated there (Fig. 3.5).

The pleura

The lungs are completely covered by a fibroelastic membrane called the visceral pleura. Lining the inside of the thoracic cavity is a complementary membrane, the parietal pleura. The opposing pleural faces are covered with a layer of cells that secretes a serous fluid into the space between them. This fluid acts as a lubricant between the two pleural layers and allows the lungs to move easily over the parietal pleura during respiration.

The two pleural layers are continuous around the root of the lungs (Fig. 3.6). The space between the two layers of pleura is known as the pleural cavity and under normal conditions it contains only a thin film of fluid. For this reason, it is only

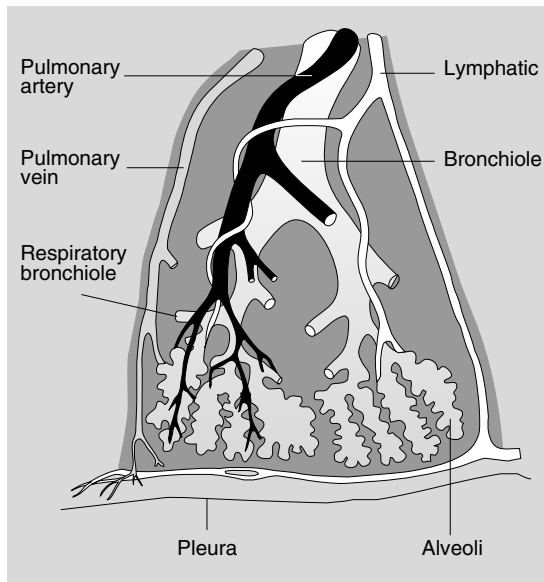


Figure 3.5 A bronchopulmonary segment. The sizes of the vessels and the airways are not to scale. Note that for clarity the lymphatics are shown only on the right and the blood vessels only on the left.

a potential space, although under some abnormal conditions it may become filled with fluid, and thus become a real space.

Function

Only a brief outline of lung function is provided here. Emphasis is placed on the main aspects of pulmonary function and on the more commonly used measures. Patterns of disordered function that are characteristic of the commoner occupational lung diseases are appended. More extensive reviews of the subject are listed at the end of the chapter.

The main function of the lungs is to allow exchange of gas between air and blood. Oxygenated blood is transported to tissue via the arterial system. Oxygen is carried in the red blood cell primarily by the haemoglobin, whereas carbon dioxide is transported from the cells using both the same mechanism and dissolved in the blood. The lungs, therefore, act like a bellows, transport-

ing oxygen in the air to the alveoli where it is exchanged for carbon dioxide. The blood leaving the lungs via the pulmonary veins has a higher oxygen concentration and a lower carbon dioxide concentration than the blood entering the pulmonary arteries. The rate of ventilation controls these concentrations within narrow limits in the presence of wide variations in the demand for oxygen and for the removal of carbon dioxide by responding to and maintaining the partial pressure of carbon dioxide (P_{CO_2}) in blood.

The pathway from air to tissue involves several steps. First, the air has to reach the alveolar-capillary membrane, where oxygen crosses to reach the red blood cell. The final step is the uptake of oxygen by tissue cells from the circulating blood. Carbon dioxide is transported in the opposite direction. This chapter is mainly concerned with the first two steps in the pathway.

The propelling force for the first step is the cyclical pressure changes produced in the lungs by ventilation, which is itself controlled by the respiratory muscles. Inspiration is an active process, achieved primarily by the contraction of the diaphragm and the intercostal muscles, which overcomes the resistance of the lungs. Resistance to these pressure changes comes from the conducting airways themselves and from the elasticity of the lung and chest wall tissues. Expiration occurs passively. Disordered ventilatory function can be due to respiratory muscle weakness, stiff lungs (decreased lung compliance) or narrowed airways.

The second step, diffusion across the alveolar-capillary membrane, is short compared with the first step (300 μm compared with 50 cm). However, gas transport can be disrupted at this stage by thickening of the alveolar-capillary membrane as occurs in diffuse pulmonary fibrosis. Such thickening impedes the flow of oxygen and carbon dioxide, which is governed by the partial pressure differences of these two gases across the membrane. As carbon dioxide is much more water soluble than oxygen, the main effect of such fibrosis is failure to transport oxygen. Carbon dioxide retention in the body is much more likely to be due to ventilation unmatched with perfusion caused by ventilatory failure, more usually caused by airflow limitation than pulmonary fibrosis.

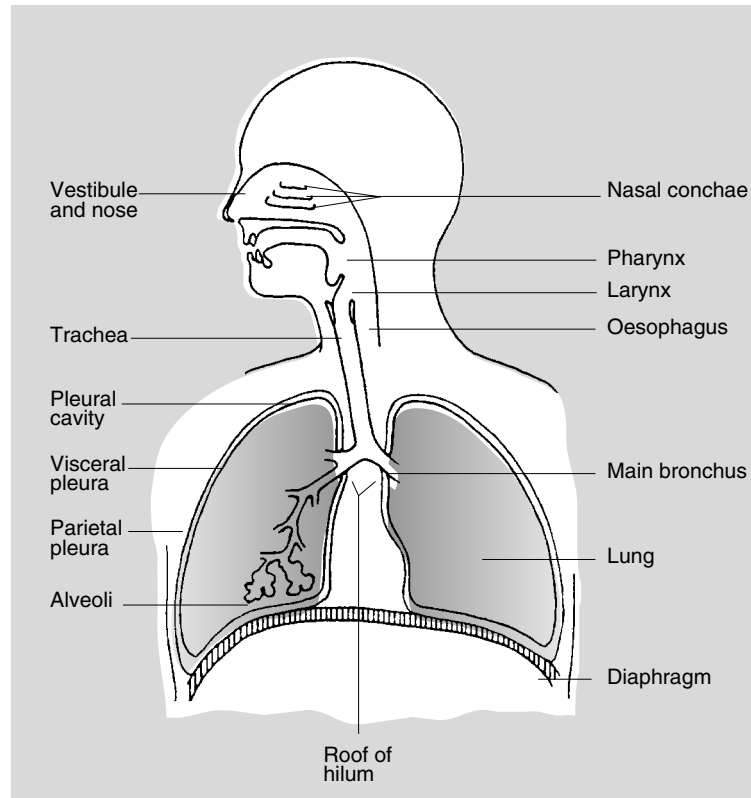


Figure 3.6 The respiratory tract, showing the relation of the pleura to the other parts.

Efficient ventilation and undamaged alveolar-capillary membranes are of no value unless ventilation and perfusion of the lung with blood are matched. Ventilation-perfusion imbalance is the major cause of impaired gas exchange. This can be illustrated, *reductio ad absurdum*, by stating that death rapidly supervenes if the right main bronchus is blocked and the left pulmonary artery is occluded! Under normal circumstances about 4 l of air are exposed to about 5 l of blood in the lungs, giving an overall ventilation-perfusion (V/Q) ratio of around 0.9. Deviation in either direction can cause serious deficiencies in blood and therefore tissue cell oxygenation.

The processes involved in gas exchange between air and tissues can, therefore, be subdivided into:

- 1 ventilation;
- 2 diffusion;
- 3 blood-gas transport.

These subdivisions can also be used to outline lung function tests in common usage. None of

these measures of lung function is diagnostic in itself, and the degree of accuracy and sophistication used in measuring lung function depends on many factors, including the facilities available (e.g. a cardiothoracic unit) and the needs of the patient and the investigator. In this chapter, emphasis is placed on the simpler tests used in occupational health practice, either to monitor an individual worker or as a measure of lung function in a clinical epidemiological survey of a factory population. It is worth noting, however, that although tests of lung volume, ventilation, gas distribution and gas transfer may be useful in differing circumstances, the basic spirometric measures of ventilatory capacity described below are indispensable in all investigations.

Ventilation

The rhythmic contraction of the inspiratory muscle produces expansion of the thorax and

lungs. This ventilation is controlled by the brain stem to maintain the partial pressure of carbon dioxide. In order to produce lung expansion, the inspiratory muscles must be capable of overcoming the inherent elastic recoil of the lung tissues and the resistance of the airways to the flow of air consequent upon these pressure changes.

The *ventilatory capacity* of the lungs is frequently assessed by using either a peak flow meter or a spirometer. Peak flow readings are not equivalent to spirometer readings, nor are they as valuable. If a single test of lung function is to be used, most authorities would advocate the use of a spirometer to measure forced vital capacity (FVC) and forced expiratory volume in 1 s (FEV_1).

Typical tracings obtained by such a machine are illustrated in Fig. 3.7. As these measures are frequently carried out by non-physiologists, it is essential that the techniques used are standardized to minimize inter- and intra-observer error. All such tests should be accompanied by information on the subject's age, height, sex, ethnic group and smoking habit. Prediction formulae incorporating the first four of these variables are available. Results

may be expressed as a percentage of the predicted value or as number of standard deviations from the mean predicted value.

Total lung capacity at full inspiration is governed to a large extent by body size. In normal subjects, this correlates well with height. Chest deformities, inspiratory muscle weakness and increased lung stiffness (loss of compliance) will reduce this volume. During maximal expiration, some air is left in the lungs – the residual volume. The physiological event that limits expiration is the closure of the intrathoracic airways. Their patency depends not only on the elastic tissues of the lungs holding them open, but also on the strength and integrity of their walls and on the presence or absence of fluid (oedema) or muscle spasm that may narrow the airways. For example, in asthma and bronchitis, the airways tend to close prematurely because of thickening of the walls of the airways, increasing resistance to airflow.

In normal subjects, 75% or more of the vital capacity can be expired in 1 s and the remaining 25% takes a further 2–3 s. Diffuse airways obstruction, as in asthma, bronchitis or emphysema, causes

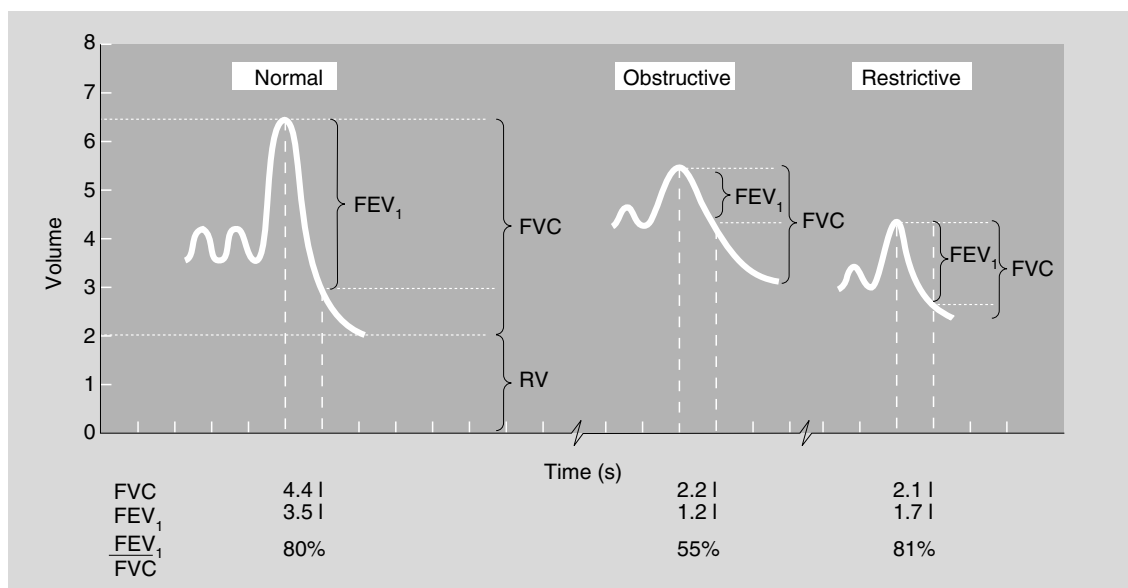


Figure 3.7 Spirograms to illustrate the differences in forced expiratory volume in 1 s (FEV_1) and forced vital capacity (FVC) in health, airways obstruction and restrictive defects (such as diffuse lung fibrosis or severe spinal deformity). Inspiration is upwards and expiration downwards. Although the vital capacity is reduced in both obstructive and restrictive lung disease, the proportion expired in 1 s shows considerable differences.

irregular and premature airways closure during expiration. Therefore, not only do these patients frequently have a reduced vital capacity (VC), but they may also be unable to expire a large portion of this air in the first second (Fig. 3.7). FEV_1 is reduced in all lung diseases that reduce the VC, but the ratio FEV_1/VC is reduced only in airways obstruction. When the subject has restrictive lung disease, for example in diffuse pulmonary fibrosis, although the VC and FEV_1 are reduced, the ratio is normal or increased. Variable airways obstruction is characteristic of asthma and can be assessed by measuring the FEV_1/FVC ratio before and after the administration of bronchodilator drugs.

Increased attention has been paid to so-called 'small airways' disease, which seems to be an early sign of impaired ventilatory capacity. Frequently, it is not distinguishable on the standard spirometer measurement recounted above. It can, however, be estimated by measuring the rate of expiratory flow during the middle of expiration. Consequently, some respiratory physiologists have turned to the forced mid-expiratory flow rate (FMF) to reveal early airways disease. This measure is sometimes referred to as the FEF_{25-75} . The fashion has diminished somewhat recently because of the low 'signal-noise' ratio. The standard spirometric reading can be used to calculate this value and is illustrated in Fig. 3.8. A more accurate assessment can be achieved using a flow-volume pneumotachograph.

The *mechanical properties* of the lung may also need to be studied. This includes not only the ability of the respiratory muscles and rib cage to perform their tasks, but also the physical (viscoelastic) properties of the lungs themselves. 'Elasticity' in this sense strictly means the ability of a structure or substance to return to its original shape and dimensions after a deforming force has been removed. It is not synonymous with 'stretchability'. 'Compliance' is the term often used to describe this property of lung tissue, but it is not the only elastic force in operation. Surface tension of the liquid lining the alveolus, for example, is estimated to provide about one-half of the lung's elastic recoil. In addition, there is impedance to airflow through the airways associated with viscous (non-elastic resistance) properties of the lung.

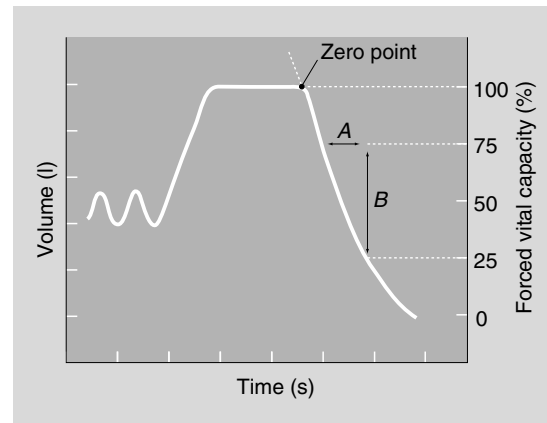


Figure 3.8 Spirogram showing the measurement of the forced vital capacity (FVC) and the derivation of the forced mid-expiratory flow rate (FMF). This is calculated from the rate of flow between the 75% and 25% points on the FVC scale and can be calculated from A/B and corrected to 1 BTPS s^{-1} (body temperature and pressure, saturated).

The estimation of compliance and non-elastic resistance of the lungs requires the measurement of the pressure differences between the alveoli and the lung surface, and between the alveoli and the mouth. The first of these parameters requires the measurement of intra-oesophageal pressure and is beyond the scope of this book. However, an apparently satisfactory measure of non-elastic resistance is the pressure difference between mouth and alveoli per unit rate of airflow. This can be achieved using the body plethysmograph. The subject sits in an airtight box and measurements are made of the subject's rate of airflow and the mouth pressures with and without blockage of the airway at the mouth by means of a shutter. Although the body plethysmograph is unsuitable for most field surveys, it does have the advantage of measuring residual volume and total lung capacity as well as non-elastic resistance. Similar measurements can also be achieved using chest radiographs.

Gas transfer

The effective exchange of oxygen and carbon dioxide in the lung depends upon three processes: 1 the correct distribution of ventilated lung to blood-perfused lung;

2 the efficient diffusion of gases across the alveolar capillary membrane;

3 the appropriate uptake and release of the gases by the red cell.

Ventilation–perfusion measures are primarily research tools and frequently involve the inhalation of inert gases and the subsequent measurement of the exhaled concentrations. Alternatively, radioactive gases can be used and lung scanning techniques employed to show the distribution of the inhaled gas in the lung.

Diffusion of gases across the alveolar capillary membrane is, likewise, a laboratory technique. However, the measurement of the gas transfer factor (TL) using carbon monoxide is so widely employed in assessing diffusion defects (as may occur in asbestosis and other fibrotic diseases of the lung) that it is worthy of note here.

TL_{CO} is defined as the quantity of pure carbon monoxide that crosses the alveolar capillary membrane in 1 min when the difference between the concentration of carbon monoxide in the lung and the blood is 1 mmHg of tension. Three measurements are required: the volume of carbon monoxide taken up by the pulmonary capillary blood, the partial pressure of carbon monoxide in the alveolar air and the partial pressure of carbon monoxide in the pulmonary capillary blood. The procedure involves breathing a known concentration of carbon monoxide and air or oxygen (usually with helium in order to assess the dilution effect of the residual volume).

It must be noted, however, that the TL_{CO} is not solely governed by the thickness of the alveolar capillary membrane. Many other factors are involved – ventilation–perfusion inequalities and haemoglobin concentration in the blood being among the more important. Nevertheless, the TL_{CO} remains the most useful, convenient and widely quoted measure of diffusing capacity.

Blood–gas transport

The final step in getting oxygen to the tissues and transporting the carbon dioxide in the opposite direction involves the blood. An estimate of the efficiency of blood–gas transport requires the measurement of the concentration of alveolar car-

bon dioxide as well as arterial oxygen, carbon dioxide and pH. Blood–gas tensions and pH are rarely measured in occupational health practice, although they are frequently vital in the management of respiratory failure in hospital practice.

Factors influencing lung function

Before outlining some of the lung function abnormalities present in the common occupationally induced lung diseases, it is pertinent to consider non-occupational factors that can influence lung function. Of these, the most important are age, height, gender and smoking habits.

Age

As a person gets older, and especially past middle age, the alveolar volume decreases and the airway volume increases. In addition, the respiratory muscles weaken and the elastic recoil of the lung becomes reduced. This results in a rise in the residual volume–total lung capacity (RV/TLC) ratio and a fall in the TL_{CO}. It is essential, therefore, that age is taken into account when comparing lung function results in survey populations. The effect on FEV₁, for example, is a loss of about 25–30 ml per year from the third decade onwards.

Smoking

In so-called ‘susceptible’ individuals, tobacco smoke, particularly cigarette smoke, has a serious adverse effect on lung function, causing an accelerated decline in FEV₁, as great as 90–100 ml per year (cf. age effect of 25–30 ml per year). The acute effect of smoking is increased airways resistance, whereas habitual smoking leads to a chronic airways obstruction in both sexes and all age groups. It causes a narrowing of small airways (obstructive bronchiolitis) and alveolar destruction (emphysema) as well as small airway narrowing. In lung function terms, this leads to a lowered FEV₁, FVC and FEV₁/FVC ratio, and a raised TL and RV. The TL_{CO} is normal unless emphysema supervenes when it falls.

The effects of smoking are so marked that failure to consider this in estimating occupational

factors that might be influencing lung function can vitiate the whole investigation. In many circumstances, those most exposed to a potential hazard have been those who have smoked most heavily (i.e. cigarette smoking *confounds* the association between hazard and lung disease). It should also be remembered that the carcinogenic effect of some inhaled occupational hazards may be greatly enhanced by smoking (a modifying effect). In some cases this effect may be multiplicative.

Other functions

Lung size, sex, ethnic group, height (rather than weight), skeletal deformity, posture, exercise tolerance, observer and instrument error, diurnal variation and ambient temperatures can all influence the results obtained from lung function tests.

Many of these can be either controlled or allowed for using standard techniques and consulting nomograms of normal (predicted) values.

Occupational lung disease and disordered function

Lung disease can be arbitrarily divided into sub-groups on the basis of the nature of the inhaled noxious agent. These are:

- 1 mineral dusts;
- 2 organic dusts;
- 3 irritant gases and vapours;
- 4 radiations.

Table 3.2 summarizes the common occupational pulmonary diseases and their pattern of functional disorder. A detailed description of the deposition of materials in the lung and the effects of inhaled materials appears in Chapter 6.

Table 3.2 Occupational lung diseases and their functional impairment.

<i>Agent</i>	<i>Type of respiratory impairment</i>	<i>Pathology</i>
Dusts		
<i>Mineral</i>		
Iron, barium, tin	None	Dust accumulation
Coal (early effects)	None (usually)	Simple pneumoconiosis (focal dust accumulation)
Coal (late effects)	a) Restrictive, diffusion; b) obstructive	Progressive massive fibroses, emphysema
Silica	Restrictive, diffusion	Nodular fibrosis can conglomerate
Beryllium	Restrictive, diffusion	Interstitial fibrosis (granulomas)
Cobalt	a) Restrictive, diffusion; b) obstructive	Interstitial fibrosis, asthma
Asbestos	Restrictive, diffusion	Interstitial fibrosis, diffuse pleural thickening
Talc	Restrictive, diffusion	Interstitial fibrosis (usually due to silica or asbestos contamination)
<i>Organic</i>		
Cotton, hemp, sisal,	Obstructive	Byssinosis
mouldy hay, barley, bagasse and straw, pigeon droppings, certain insects, animal hairs, bacterial products and drugs	Restrictive, diffusion (chronic) (immunological tests may be helpful)	Granulomatous alveolitis, interstitial fibrosis
Gases and vapours*		
Nitrogen dioxide	Obstructive	Obstructive bronchiolitis, pulmonary oedema
Sulphur dioxide, ammonia, chlorine, phosgene	Obstructive	Acute bronchitis, acute pulmonary oedema
Cadmium oxide	Acute restrictive diffusion, chronic obstructive diffusion	Emphysema, acute pulmonary oedema
Isocyanates, platinum salts	Obstructive	Asthma
Radiation		
Ionizing radiation	Restrictive, diffusion	Pulmonary oedema (early), intestinal fibrosis (late)

*Obstruction or restriction depends on solubility of gas and on site of injury.

Further reading

Newman-Taylor, A.J. (1991). Occupational aspects of pulmonary disease. In *Recent Advances in Respiratory Medicine*, 5th edn (ed. D.M. Mitchell). Churchill Livingstone, Edinburgh.

Parkes, W.R. (ed.) (1994). *Occupational Lung Disorders*, 3rd edn. Butterworth, London.

West, J.B. (1990). *Respiratory Physiology – the Essentials*, 4th edn. Williams & Wilkins, Baltimore.

Chapter 4

Organ structure and function: the skin

Iain S. Foulds

Introduction

Epidermis

Rate of maturation

Dermis

Subcutaneous layer

Derivatives of the skin

Hair

Nails

Sebaceous glands

Sweat glands

Eccrine sweat glands

Apocrine glands

Other structures in the skin

Nerve supply

Blood and lymphatic vessels

Melanocyte function

Thermoregulation

Blood flow

Sweat

Defence mechanisms of the skin

The skin as a barrier

Irritant contact dermatitis

The clinical course of irritant contact dermatitis

Immunology of the skin

Hypersensitivity reactions and the skin

Type I (immediate)

Type II (antibody-dependent cytotoxicity)

Type III (immune complex disease)

Type IV (cell mediated or delayed)

Allergic contact dermatitis

Barrier creams

Skin cleansing

After-work creams

Conclusions

Reference

Further reading

Introduction

The skin is one of the largest organs in the body, having a surface area of 1.8 m² in an adult, making up approximately 16% of the body weight. It has many functions, the most important of which is as a barrier to protect the body from external factors and to keep the internal body systems intact.

The skin is composed of three layers: the epidermis, the dermis and the subcutis (Fig. 4.1). There are two main kinds of human skin: glabrous skin (found on the palms and soles) and hairy skin.

The skin is a metabolically active organ with vital functions including the protection and homeostasis of the body (Table 4.1).

Epidermis

The epidermis is defined as a stratified squamous epithelium which is about 0.1 mm thick, although

the thickness is greater (0.4–1.4 mm) on the glabrous skin of the palms and soles. Its main function is to act as a protective barrier. The main cell of the epidermis is the keratinocyte, which produces the protein keratin. The four layers of the epidermis – the basal, prickle, granular cell layers and the horny layer (stratum corneum) (Fig. 4.1) – represent the stages of maturation of the keratin by keratinocytes.

The differentiation of basal cells into dead, but functionally important, corneocytes is a unique feature of the skin. The horny layer is important in preventing all manner of agents from entering the skin, including micro-organisms, water and particulate matter. The epidermis also prevents the body's fluids from getting out.

Epidermal cells undergo the following sequence during keratinocyte maturation.

1 Undifferentiated cells in the basal layer (stratum basale) and the layer immediately above divide continuously; one-half of these cells remain in

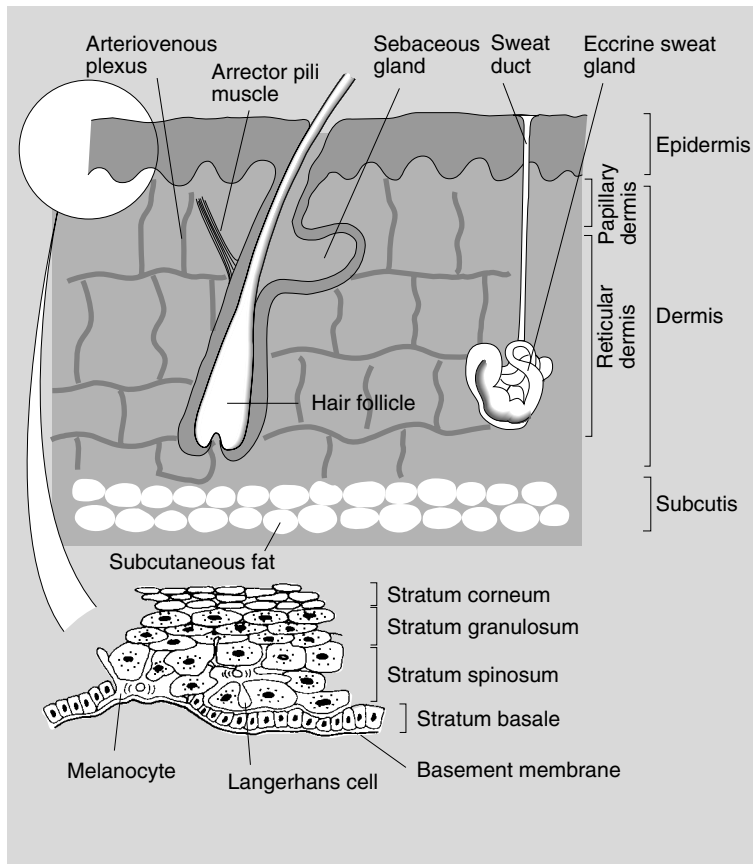


Figure 4.1 A cross-section of normal skin.

place and one-half progress upwards and differentiate.

2 In the prickle cell layer (stratum spinosum), cells change from being columnar to polygonal. Differentiating keratinocytes synthesize keratins, which

Table 4.1 Functions of the skin.

Presents barrier to physical agents
Protects against mechanical injury
Prevents loss of body fluids
Reduces penetration of ultraviolet radiation
Helps regulate body temperature
Acts as a sensory organ
Affords a surface for grip
Plays a role in vitamin production
Acts as an outpost for immune surveillance
Cosmetic association

aggregate to form tonofilaments. The desmosomes are the connections between keratinocytes that are condensations of tonofilaments. Desmosomes distribute structural stresses throughout the epidermis and maintain a distance of 20 nm between adjacent cells.

3 Enzymes in the granular cell layer (stratum granulosum) induce degradation of nuclei and organelles. Keratohyalin granules mature the keratin and provide an amorphous protein matrix for the tonofilaments. Membrane-coating granules attached to the cell membrane release an impervious lipid-containing cement that contributes to cell adhesion and to the horny layer barrier.

4 In the horny layer (stratum corneum), the dead, flattened corneocytes have developed thickened cell envelopes encasing a matrix of keratin tonofibrils. The strong disulphide bonds of the keratin

provide strength to the stratum corneum but the layer is also flexible and can absorb up to three times its own weight in water. However, if it dries out with the water content falling below 10%, pliability fails.

The corneocytes are eventually shed from the skin surface.

Rate of maturation

Kinetic studies show that, on average, the dividing basal cells replicate every 200–400 h and the resultant differentiating cells take about 14 days to reach the stratum corneum and a further 14 days to be shed. The cell turnover time is considerably shortened in keratinization disorders such as psoriasis.

Dermis

The dermis is defined as a tough, supportive connective tissue matrix, containing specialized structures, which is found immediately below, and intermittently connected with, the epidermis. It varies in thickness, being thin (0.6 mm) on the eyelids and thicker (3 mm or more) on the back, palms and soles. Collagen fibres make up 70% of the dermis and impart a toughness in strength to the structure. Elastin fibres are loosely arranged in all directions in the dermis and provide elasticity to the skin. They are more numerous near hair follicles and sweat glands and less so in the papillary dermis. The ground substance of the dermis is a semisolid matrix of glycosaminoglycans (GAGs) that allows dermal structures some movement.

The dermis contains fibroblasts that synthesize collagen, elastin and other connective tissue, and GAG. In addition, the dermis contains dermal dendrocytes (dendritic cells) with a probable immune function, mast cells, macrophages and lymphocytes.

Subcutaneous layer

The subcutis consists of loose connective tissue and fat (up to 3 cm thick on the abdomen).

Derivatives of the skin

Hair

Hairs are found over the entire surface of the skin, with the exception of the glabrous skin of the palms, soles, glans penis and vulva. The density of the follicles is greatest on the face. Embryologically, the hair follicle has an input from the epidermis, which is responsible for the matrix cells and the hair shaft, and the dermis, which contributes the papilla with its blood vessels and nerves.

There are three types of hair and these are listed below.

1 *Lanugo* hairs are fine and long and are formed in the fetus at 20 weeks' gestation. They are normally shed before birth but may be seen in premature babies.

2 *Vellus* hairs are the short, fine, light-coloured hairs that cover most of the body surfaces.

3 *Terminal* hairs are longer, thicker and darker and are found on the scalp, eyebrows, and eyelashes, and also on the pubic, axillary and beard areas. They originate as vellus hair; differentiation is stimulated at puberty by androgens.

In most mammals, hair or fur plays an essential role in survival, especially in the conservation of heat; this is not the case in nude man. Scalp hair in humans does function as a protection against the cancer-inducing effects of ultraviolet radiation; it also protects against minor injury. However, the main role of hair in human society is as an organ of sexual attraction and therein lies its importance to the cosmetics industry.

Nails

The nail is a phylogenetic remnant of the mammalian claw and consists of a plate of hardened and densely packed keratin. It protects the fingertips and facilitates grasping and tactile sensitivity in the finger pulp.

Sebaceous glands

Sebaceous glands are found associated with hair follicles, especially those of the scalp, face, chest

and back, and are not found on non-hairy skin. They are formed from epidermis-derived cells and produce an oily sebum, the function of which is uncertain. The glands are small in the child but become large and active at puberty, being sensitive to androgens. Sebum is produced by holocrine secretion in which the cells disintegrate to release their lipid cytoplasm.

Sweat glands

Sweat glands are like coiled tubes, located within the epidermis, which produce a watery secretion. There are two separate types: eccrine and apocrine.

Eccrine sweat glands

Eccrine sweat glands develop from down-budding of the epidermis. The secretory portion is a coiled structure in the deep reticular dermis; the excretory duct spirals upwards to open onto the skin surface. An estimated 2.5 million sweat ducts are present on the skin surface. They are universally distributed but are most profuse on the palms, soles, axillae and forehead, where the glands are under both psychological and thermal control (those elsewhere being under thermal control only). Eccrine sweat glands are innervated by sympathetic (cholinergic) nerve fibres.

Apocrine glands

These are also derived from the epidermis. Apocrine sweat glands open into hair follicles and are larger than eccrine glands. They are most numerous around the axillae, perineum and areolae. The secretion is odourless when produced, although an odour develops after the action of skin bacteria. Sweating is controlled by sympathetic (adrenergic) innervation. The apocrine glands represent a phylogenetic remnant of the mammalian sexual scent gland.

Other structures in the skin

Nerve supply

The skin is richly innervated, with the highest density of nerves being found in areas such as the hands, face and genitalia. All nerve supplies in the skin have their cell bodies in the dorsal root ganglia. Both myelinated and non-myelinated fibres are found. Free sensory nerve endings occur in the dermis and also encroaching on the epidermis where they may abut onto Merkel cells. These nerve endings detect pain, irritation and temperature. Specialized corpuscular receptors are distributed in the dermis, such as Pacini's corpuscles detecting pressure and vibration, and touch-sensitive Meissner's corpuscles which are mainly seen in the dermal papillae of the feet and hands.

Autonomic nerves supply the blood vessels, sweat glands and arrector pili muscles. The nerve supply is dermatomal with some overlap.

Blood and lymphatic vessels

The skin also has a rich and adaptive blood supply. Arteries in the subcutis branch upwards, forming a superficial plexus at the papillary-reticular dermal boundary. Branches extend to the dermal papillae, each of which has a single loop of capillary vessels, one arterial and one venous. Veins drain from the venous side of this loop to form the mid-dermal and subcutaneous venous networks. In the reticular and papillary dermis there are arteriovenous anastomoses that are well innervated and are concerned with thermoregulation.

The lymphatic drainage of the skin is important. Abundant meshes of lymphatics originate in the papillary dermis and assemble into larger vessels that ultimately drain into the regional lymph nodes.

Melanocyte function

Melanocytes (located in the basal layer) produce the pigment melanin in elongated, membrane-

bound organelles known as melanosomes. These are packaged into granules that are moved down dendritic processes and transferred by phagocytosis to adjacent keratinocytes. Melanin granules form a protective cap over the outer part of keratinocyte nuclei in the inner layers of the epidermis. In the stratum corneum, they are uniformly distributed to form an ultraviolet-absorbing blanket that reduces the amount of radiation penetrating the skin.

Ultraviolet radiation, mainly the wavelengths of 290–320 nm (ultraviolet B), darkens the skin first by immediate photo-oxidation of preformed melanin, and second over a period of days by stimulating melanocytes to produce more melanin. Ultraviolet radiation also induces keratinocyte proliferation, resulting in thickening of the epidermis.

Variations in racial pigmentation are not due to differences in melanocyte numbers, but to the number and size of melanosomes produced. Red-haired people have pheomelanin, not the more usual eumelanin, and their melanosomes are spherical rather than oblong.

Thermoregulation

The maintenance of a near constant body core temperature of 37°C is a great advantage to humans, allowing a constancy to many biochemical reactions which would otherwise fluctuate widely with temperature changes. Thermoregulation depends on several factors, including metabolism and exercise, but the skin plays an important part in control through the evaporation of sweat and by direct heat loss from the surface.

Blood flow

Skin temperature is highly responsive to skin blood flow. Dilatation or contraction of the dermal blood vessels results in vast changes in blood flow, which can vary from 1 to 100 ml min⁻¹ per 100 g of skin for the fingers and forearms. Arteriovenous anastomoses under the control of the sympathetic nervous system shunt blood to or from the superficial venous plexus, affecting skin temperature. Local

factors, both chemical and physical, can also have an effect.

Sweat

The production of sweat cools the skin through evaporation. The minimum secretion per day is 0.5 l and the maximum is 10 l, with a maximum output of about 2 l h⁻¹. Men sweat more than women.

Watery isotonic sweat produced in the sweat gland is modified in the excretory portion of the duct so that the fluid delivered to the skin surface has:

- 1 a pH of between 4 and 6.8;
- 2 a low concentration of sodium (32–70 mequiv. l⁻¹) and chlorine (30–70 mequiv. l⁻¹);
- 3 a high concentration of potassium (up to 5 mequiv. l⁻¹), lactate (4–14 mequiv. l⁻¹), urea, ammonia and some amino acids.

Only small quantities of toxic substances are lost. Sweating may also occur in response to emotion and after eating spicy foods. In addition to thermoregulation, sweat also helps to maintain the hydration of the horny layer and improves grip on the palms and soles.

Defence mechanisms of the skin

The skin achieves protection through a variety of mechanisms. The outermost layer of the skin, known as the horny layer or stratum corneum, acts as a barrier against chemicals. Therefore, with the stratum corneum being relatively thin on the backs of the hands compared with the palms, it is often the backs of the hands that are initially affected by dermatitis. The presence of pigment cells (melanocytes) provides protection against the damaging effects of ultraviolet light, and the sweat glands and sebaceous glands help maintain hydration and suppleness of the skin. The skin is able to resist shearing stresses owing to elastic tissue and collagen. A continual upwards movement of cells in the epidermis provides continual

replacement from wear and tear, and at the same time discourages growth of bacteria on the surface of the skin. The skin is able to resist (buffer) the effects of mild acids but does not buffer alkalis effectively and does not tolerate strong acids, alkalis or solvents.

The skin as a barrier

Penetration of the skin is a passive process that occurs unaided by the cells within the skin. With different chemical substances the penetration rate may vary fourfold. In addition, damage or disease may affect the barrier and result in increased percutaneous absorption. Percutaneous absorption involves a series of processes. Molecules of the chemical must be absorbed at the surface of the stratum corneum and then diffuse through the flattened cell layers. After that, the compound must enter the viable epidermis and then the dermis until it reaches a capillary where it can enter the systemic circulation. Water can penetrate the skin at an average of $0.2\text{--}0.4\text{ mg cm}^{-2}\text{ h}^{-1}$ at 30°C . Hydrophilic chemicals generally penetrate the skin more slowly than does water.

Cell membranes and intercellular spaces contain lipids, and therefore lipophilic compounds can also penetrate the skin. When chemical compounds penetrate the skin, local toxicity may occur, for example with caustic, allergenic or phototoxic agents or compounds that can cause skin cancer.

Chemicals absorbed through the skin rather than through the oral route may be more toxic to the body as detoxification by the liver may be bypassed. However, the skin also contains many different enzyme systems, some of which are similar to the liver and therefore some detoxification may be possible. An example of this is when organophosphorous pesticides are applied to the skin and a large proportion are metabolized during their passage through the skin. Conversely, aryl hydrocarbon hydroxylase in the skin may convert non-carcinogenic benzo-alpha-pyrene into a potent carcinogen that could cause effects within the skin or elsewhere. Therefore, first-pass metabolism may increase or decrease the systemic bioavailability of a compound.

In theory compounds may also penetrate through the appendages of the skin (sweat, sebaceous and apocrine glands and beside hair follicles). In practice, this route accounts for less than 6% of chemical penetration.

Irritant contact dermatitis

Over 90% of industrial dermatitis is due to irritant contact factors. Irritants are substances that damage the skin by direct toxic action. Their effect is proportional to the nature of the chemical, the length of exposure and the individual's skin protection and tolerance.

Irritants can be divided into absolute and relative irritants. Absolute irritants are chemicals which produce irritation in everyone by direct tissue destruction if the concentration and exposure time are adequate. Concentrated acids or alkalis are examples of absolute irritants.

Relative irritants are milder substances that will produce inflammatory changes in most individuals provided there is repeated exposure; for example, paraffin and solvents. A fuller list of potential irritants in certain occupations can be found in standard dermatology texts (see Further reading). Moreover, certain factors can increase the susceptibility of the skin to irritants. Dermatitis that has healed, or a burn, may remain more susceptible to irritants for several months. This is due to a lower threshold of resistance, which is also one of the reasons why those with previous atopic eczema are more susceptible to the effects of irritants. Extremes of humidity and temperature, friction, pressure, sweating and occlusion may also contribute to irritant damage by allowing relatively bland substances to cause irritation.

This can be more easily understood in simple graphical terms (Fig. 4.2). Exposure to an irritant may cause some damage to the skin (point A), but no clinical abnormality is present. Repeated exposure to this irritant or to other irritants over a period of time (B, C, D) will produce further damage to the skin, but still no clinical abnormality. However, with further exposure to an irritant (E), dermatitis (inflammation of the skin) develops once the threshold is crossed, as a result of the

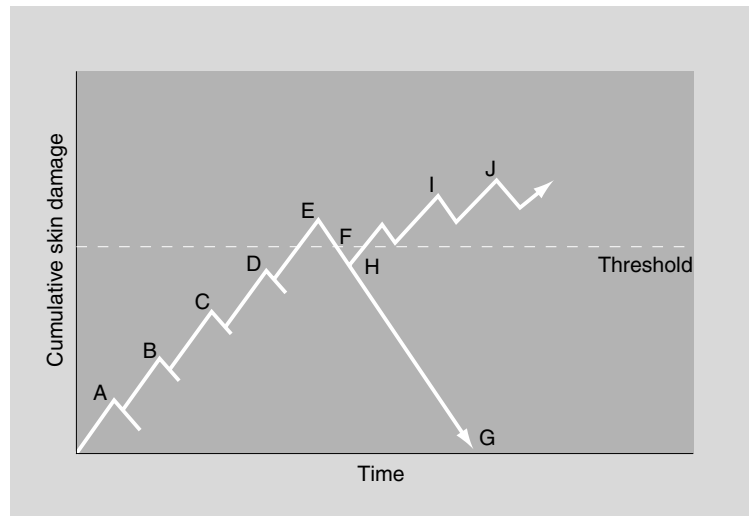


Figure 4.2 A model of irritant contact dermatitis. After Malten (1981).

cumulative effect of damage by irritants. This process may take a considerable time to develop. Many workers will state that they have worked with a particular substance for years with no trouble until the present time. This is a typical story of an individual suffering from irritant contact dermatitis due to repeated insults occurring to the skin.

In practice, the threshold for the development of dermatitis varies in individuals, and those with a pre-existing skin disease or previous history of atopic eczema will have a lower threshold of susceptibility. Once the dermatitis has started it can appear to improve clinically (F), but to achieve full recovery (G) may take months or possibly even years. In practice, once dermatitis appears to have improved, individuals tend to revert to exposing themselves to irritants, and at this stage (H) a minimum amount of irritation will result in a recurrence of the dermatitis. With continuing exposure to irritants (I, J, etc.) the dermatitis may never heal.

The clinical course of irritant contact dermatitis

Once irritant contact dermatitis develops then all exposure to irritant factors must be reduced, not only at work, but also at home if clinical clearance

is to be achieved and maintained. All too often an employee is signed off work with no particular advice for irritant avoidance at home, the dermatitis clears, a return to work follows with re-exposure to irritants and the dermatitis recurs. When recurrence occurs repeatedly on returning to work the employees may be forced to stop working, but if all potential irritants were carefully identified, clearance of dermatitis could be achieved and maintained. Every effort must therefore be made to reduce overall exposure to irritants, which should enable an individual to remain in gainful employment.

However, repeated daily exposure to irritants may induce toughening and resistance, allowing repeated contact without further evidence of irritation. This process is called 'hardening' and is an individually acquired resistance, providing only local protection, with short periods away from work resulting in decreased resistance. This cannot, therefore, be relied upon as a means of protection against irritants.

Water is a potential irritant. It penetrates relatively easily through the stratum corneum and prolonged exposure is a well recognized cause of dermatitis among housewives, bartenders, nurses and mechanics. The protective layer of the skin is diminished, and warmth, occlusion, bacteria, fungi and chemical exposure can further

contribute to skin damage. Few claims are made for occupational dermatitis caused by water, as many affected are self-treated with no loss of time from work.

The stratum corneum accounts for the major diffusional resistance of the skin. Initially, on contact with water, a diffusion pathway exists through the hair follicles and sweat glands for a short period. With continued exposure, a steady state is reached across the whole stratum corneum, with diffusion across this layer being the predominant pattern. In environments of relative humidities of less than 60%, water binds directly to keratin fibrils. At relative humidities from 60% to 94%, water interacts with the fibrillar bound water. At humidities greater than 94%, water content increases rapidly as 'free water' not bound to keratin fibrils, and the stratum corneum begins to break down mechanically.

Solvents, which are mixtures of fluids capable of dissolving substances to produce other compounds, are estimated to be responsible for up to 20% of industrial dermatitis. The physical properties of solvents influence the injurious effects they have on the skin. In general, the more poorly absorbed solvents cause the most skin damage and the least systemic symptoms. For example, saturated hydrocarbon solvents and the paraffin series of solvents are stronger skin irritants than those derived from the aromatic series. Solvents with a higher boiling point tend to have a less irritant effect.

The most frequent cause of solvent-induced industrial dermatitis is the practice of washing with solvents. As solvents are frequently used in industry as degreasing agents to remove oil, grease and stains from manufactured products, they are commonly used as hand cleansers because they are quick and effective. With repeated use cumulative irritation occurs, with the risk of irritant contact dermatitis developing. Painters use thinners and turpentine, printers use type-wash, plastic workers use acetone, mechanics use petrol or paraffin and dry-cleaners use trichloroethylene or perchloroethylene as soap substitutes.

Solvents dissolve and remove surface lipids, the lipid material within the stratum corneum, and the fatty fraction of cell membranes. With defatting of

the skin there is increased percutaneous absorption of water and other substances.

Soaps and detergents are the major predisposing or perpetuating factors in the majority of cases of hand dermatitis. The irritancy of soaps is due to the combination of alkalinity, degreasing action and direct irritancy of fatty acids, which with repeated water exposure damage the epidermal layers. In addition to this abrasive contact, the presence of additives may also contribute to irritation. Most bar soaps have alkaline builders such as sodium bicarbonate, sodium phosphate, ash, borax or silicate added to increase cleansing. These may additionally irritate the skin. Perfumes and colours increase the appeal, and germicidal agents help to prevent deterioration, but these may cause irritancy or sensitization. Abrasive agents are often incorporated into industrial cleansers to increase the mechanical cleansing action. These include inorganic agents such as pumice, chalk, sand or borax, and organic agents such as groundnut shells, cornmeal or wood flours. In summary, the early recognition and appropriate institution of measures to reduce the overall exposure to irritants by an individual with irritant contact dermatitis will often enable continuation of employment.

Immunology of the skin

The skin is an important immunological organ and normally contains nearly all of the elements of cellular immunity, with the exception of B cells. Much of the original research into immunology was undertaken using the skin as a model.

Hypersensitivity reactions and the skin

'Hypersensitivity' is the term applied when an adaptive immune response is inappropriate or exaggerated to the degree that tissue damage results. The skin can exhibit all the main types of hypersensitivity response.

Type I (immediate)

Immunoglobulin E (IgE) is bound to the surface of mast cells by Fc receptors. On encountering an

antigen (e.g. a house dust mite, food or pollen) the IgE molecules become crosslinked, causing degranulation and the release of the inflammatory mediators. These include preformed mediators (such as histamine) and newly formed ones (e.g. prostaglandins or leukotrienes). The result in the skin is urticaria, although massive histamine release can cause anaphylaxis – a life-threatening condition. The response occurs within minutes and individuals may react to many different chemicals coming in contact with the skin. This results in a widespread itchy rash similar to a giant nettle rash (urticaria or hives). This may be caused by diverse substances, from rubber latex to amniotic fluid (Table 4.2). If the swellings affect the mouth, then the throat may also become affected and this can lead to suffocation.

Type II (antibody-dependent cytotoxicity)

Antibodies directed against an antigen on target skin cells or structures induce cytotoxicity by killer T cells or by complement activation. This type of hypersensitivity is not thought to be of significance in occupational health.

Type III (immune complex disease)

Immune complexes are formed by the combination of an antigen and antibodies in the blood and are deposited in the walls of small vessels, often those of the skin. Complement activation, platelet aggregation and the release of lysosomal enzymes from

polymorphs cause vascular damage. Again this type of hypersensitivity is not thought to be of significance in occupational health.

Type IV (cell mediated or delayed)

Specifically sensitized T lymphocytes have secondary contact with the antigen when it is presented on the surface of the antigen-presenting cells. Cytokine release causes T-cell activation and amplifies the reaction by recruiting other T cells and macrophages to the site. Tissue damage results, which is maximal at 48–72 h after contact with T-lymphocyte antigen. Allergic contact dermatitis and the tuberculin reaction to intradermally administered antigen are both forms of type IV reaction. This is the most important immunological reaction in the development of allergic dermatitis.

Allergic contact dermatitis

Clinically, this form of dermatitis looks identical to irritant contact dermatitis, but it is caused by an individual developing a specific allergy to a substance. Whereas irritant contact dermatitis has the potential to affect all people, allergic contact dermatitis will only affect a small proportion of people exposed to the substance. Unfortunately, it is not possible to predict which individuals may develop this problem.

For allergy to develop, repeated exposure to the substance over a period of time is required, usually months or years, until the skin becomes sensitized. For sensitization to occur, a potential allergen needs to penetrate the epidermis and therefore its molecular weight must be less than 500. More typically, this falls in the range of 200–300. The chemical is then picked up by Langerhans cells and transported to the regional lymph nodes, where it is processed. In some individuals, a pool of sensitized cells (lymphocytes) is then formed. In practice, it usually takes months or even years of exposure for sensitization to occur.

Once sensitized, further exposure to the substance, even at low concentrations, and at any skin site will result in an outpouring of these cells from the lymph nodes to the site of exposure,

Table 4.2 Agents causing contact urticaria.

Foodstuffs and body fluids
Latex
Ammonium persulphate
Platinum salts
Cobalt chloride
Ammonia
Aliphatic polyamines
Sulphur dioxide
Aminothiazole
Lindane
Acrylic monomers
Exotic woods

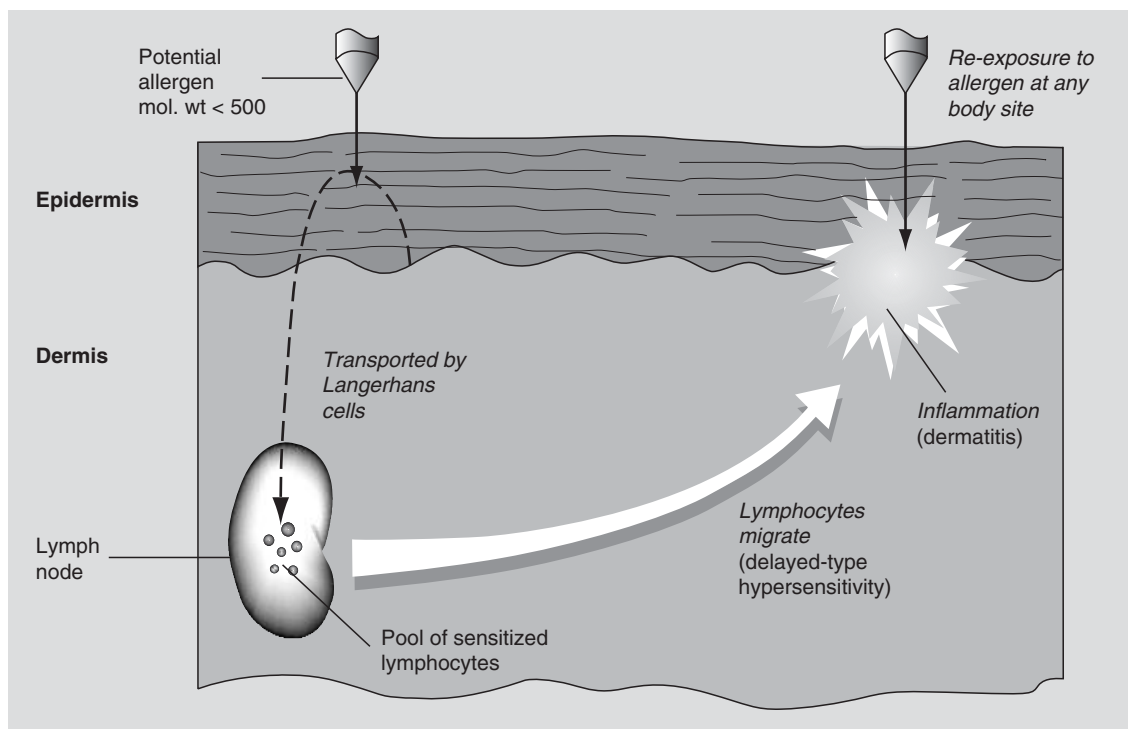


Figure 4.3 A model of allergic contact dermatitis.

causing inflammation (dermatitis) (Fig. 4.3). However, the time of onset of dermatitis from re-exposure may take from 1 to 5 days as it takes time for the allergen to penetrate and for the sensitized cells to respond. Hence the name ‘delayed-type hypersensitivity’. For this reason, a cause and effect may not be realized by the sufferer. Once individuals become sensitized, they remain sensitized for the rest of their lives; however, a small proportion may lose their sensitivity with the passage of time.

If allergic sensitization is suspected then patch testing may be undertaken by dermatologists to identify the allergen. This involves trying to reproduce what is happening in the skin by applying the allergen to the skin at an appropriate concentration to try to induce an inflammatory reaction. The allergens are usually applied to the back and are left in place. After 48 h an initial reading is taken and a final reading is taken after 96 h.

Once an allergen has been identified for an individual then avoidance has to be instituted to pre-

vent recurrences of the dermatitis. Allergic contact dermatitis accounts for about 5% of all industrial dermatitis, whereas irritant contact dermatitis accounts for the rest.

Barrier creams

It should be recognized that there is no cream that actually provides a barrier preventing the penetration of substances into the skin. In fact, in some situations they may actually enhance penetration. There are numerous formulations available and these are intended for either dry or wet work. Those for dry work are water soluble and often contain polyethylene glycols. Those for wet work are water insoluble and are often based on lanolins, paraffins or silicones.

In practice, the main benefit they offer is due to their bases, which may help to improve the hydration (suppleness) of the skin, with the result that,

when cleansers are used, less degreasing of the skin occurs. In theory, this may help to reduce irritant contact dermatitis from repeated hand washing. There is subjective evidence among those using barrier creams that the skin is easier to cleanse.

There is no evidence that barrier creams protect against sensitizers and occasionally sensitization to some of the constituents of the cream may occur. However, the use of a barrier cream may give an employee a false sense of security and lead to increased abuse of the skin.

Provided their limitations are recognized, barrier creams are of benefit overall by increasing the hydration of the skin, which may result in less harsh cleansers being required.

Skin cleansing

If substances remain on the skin after the working day, the risk of irritation or sensitization is increased. However, the most efficient skin cleansers are often the most irritant of substances because of their solvent or detergent content. If cleansers are too mild for the task, workers will often use degreasing agents used at work for industrial purposes, for example solvents or paraffins, to obtain adequate cleansing. Although these substances will clean, they are potentially very irritant if used repeatedly. It is often not appropriate to provide one type of cleanser for different jobs. These agents should be chosen to provide adequate cleaning in a short period of time, without being too strong a degreasing agent.

Drying of the skin after cleansing is equally important, and disposable paper towels or a pull-down roller towel are preferable to a roller towel or rag which will become – and remain – wet or dirty.

After-work creams

Many companies now produce after-work creams, which in essence are moisturizers, but which have the benefit of increasing the hydration of the skin at the end of the day. They are of particular benefit

in occupations where excessive drying of the skin may occur. In addition, their use should be encouraged where hot-air driers are used as these tend to dry the skin excessively.

Conclusions

In the workplace, some skin contact with chemicals is almost inevitable. This cutaneous exposure may not necessarily constitute a health hazard, but seemingly minimal contact with certain substances may lead to toxicity. Many factors play an important role with regard to the rate and extent of absorption through the skin. As a result of the complexities, individual percutaneous absorption cannot be predicted, even from working habits and methods. This situation is therefore very different from the more familiar respiratory reactions to airborne chemical exposure. Therefore, for safe working practices and for preventive purposes, appropriate safeguards must be defined for each specific exposure.

Reference

Malten, K.E. (1981). Thoughts on irritant contact dermatitis. *Contact Dermatitis*, 7, 238–42.

Further reading

- Adams, R.M. (1999). *Occupational Skin Disease*, 3rd edn. W.B. Saunders, Philadelphia.
- Fisher, A.A. (2001). *Contact Dermatitis*, 5th edn. Lippincott Williams, Philadelphia.
- Foulds, I.S. (1987). Occupational skin disease. In *Textbook of Occupational Medicine*, (eds J.K. Howard and F.H. Tyner). Churchill Livingstone, Edinburgh.
- Fregert, S. (1981). *Manual of Contact Dermatitis*, 2nd edn. Munksgaard, Copenhagen.
- Kanerva, L. (2000). *Handbook of Occupational Dermatology*. Springer, Berlin.
- Wilkinson, J.D. and Rycroft, R.J.G. (1998). The principal irritants and sensitizers. In *Textbook of Dermatology*, 6th edn (eds A. Rook, D.S. Wilkinson, F.J.G. Ebling, R.H. Champion and J.L. Burton), pp. 709–821. Blackwell Scientific Publications, Oxford.

Chapter 5

Musculoskeletal disorders

Grahame Brown

Introduction
Terminology
Historical perspective
 Upper limb pain
 Back pain
A biopsychosocial approach
Essential epidemiology

Occupational management
 Primary prevention
 Secondary prevention
 Physical treatments
 Tertiary prevention
Hand–arm vibration syndrome
References
Further reading

Introduction

Musculoskeletal symptoms of various types (regional pain disorders of the neck, limb, or low back, joint pain, chronic widespread pain) are a major reason for consultation in primary care and, together with mental health problems, are the major reasons for long-term sickness absence from work. There is a substantial body of evidence emerging for the management of low back pain: the principles apply to the management of other regional pain disorders. This chapter will focus on these regional pain disorders that contribute to the vast majority of work disability. Serious diseases of the musculoskeletal system such as bone infections or tumours, systemic diseases such as rheumatoid arthritis and major trauma such as fractures are separate and will not be discussed: these usually clear-cut pathological disorders have well-defined pathways of medical management that are not disputed. Hand–arm vibration syndrome is a unique occupational disease and will be considered separately.

The increasing prevalence of disability attributable to musculoskeletal disorders in industrialized countries has been described as an epidemic, and the management of back pain in particular as a twentieth century health-care disaster [1]. Pain complaints are usually self-limiting but, if they

become chronic, the consequences are serious. These include the distress of patients and their families and consequences for employers in terms of sickness absence and for society as a whole in terms of welfare benefits and lost productivity. Many causes for musculoskeletal pain have been identified. Psychological and social factors have been shown to play a major role in exacerbating the peripheral tissue source of pain by influencing pain perception and the development of chronic pain and disability. This new understanding has led to a ‘biopsychosocial’ model for pain and functional disorders.

There are many different reasons for patients to consult health-care professionals with pain: seeking cure or symptomatic relief, diagnostic clarification, reassurance, ‘legitimization’ of symptoms, medical certification for work absence or to express distress, frustration or anger. Health-care professionals need to clarify which of these apply to an individual and to respond appropriately.

Terminology

Much confusion arises as a result of the lack of consensus on naming musculoskeletal disorders that are by nature heterogeneous. This is one of the reasons why good quality research into

epidemiology and management is sadly lacking, especially with regard to upper limb pain. Terms such as ‘cumulative trauma disorder’, ‘cervico-brachial syndrome’, ‘work-related upper limb pain’, ‘repetitive strain injury’, ‘repetitive stress injury’ and ‘writer’s cramp’ are all referring to the same problem of chronic (long-term) diffuse arm pain with neuromuscular dysfunction for which there is no simple orthodox medical explanation. In this text I shall refer to these and other similar pain disorders of the low back simply as *regional pain disorders*. This term avoids implying causation, as so often the contributing factors are multifactorial. So often the word *injury* is used when referring to these disorders, for example *back injury*, *repetitive strain injury*. The term implies that something has been injured and often sets the sufferer on a long and counterproductive search for this injury when most often all the high-tech investigations show is normal age-related changes or no changes at all. It is rather like calling a common headache a *head injury*. I prefer to reserve the use of the term injury to unambiguous tissue damage usually caused by a significant trauma such as a fracture or musculo-tendinous rupture.

It is important to clarify as far as possible other terms commonly used in texts and patient assessment to avoid misunderstandings:

- *Nociception*. A noxious stimulus in the tissues with the potential to generate the perception of pain, for example a bruise, fracture, inflamed joint or nerve, a trigger point.
- *Pain*. This is a subjective experience; every individual will have a different level of tolerance. The perception and modulation of pain in the nervous system is exceedingly complex and is readily influenced by cognitive and emotional factors. The nervous system’s response to pain is described as ‘plastic’; it will change over time, with central sensitization and morphological changes occurring in the dorsal horn of the spinal cord leading to ‘neuropathic’ pain. In neuropathic pain, the sensitized dorsal horn of the spinal cord is sending pain messages up to the brain when there is no, or only trivial, sensory input from the periphery.
- *Suffering*. A patient will often describe their pain in terms of suffering which is subject to emotional influence.

- *Disability*. The inability to perform a task owing to the presence of pain. This is known to be strongly influenced by emotional, cognitive, behavioural, motivation and socio-economic as well as biomedical factors.

- *Impairment*. The loss of function of a part of the body, either temporary or permanent, for example a foot drop due to paralysis of the muscles supplied by the fifth lumbar nerve following a severe episode of sciatica.

Historical perspective

Upper limb pain

Bernadino Ramazzini (1633–1714), the founder of occupational medicine, described the problems experienced by scribes (who wrote with a quill) in his treatise *De morbis artificum* in 1713:

Furthermore, incessant driving of the pen over paper causes intense fatigue of the hand and the whole arm because of the continuous and almost tonic strain on the muscles and tendons, which in course of time results in failure of power in the right hand . . . what tortures these workers most acutely is the intense and incessant application of their mind, for in work such as this the whole brain, its nerves and fibres, must be constantly on the stretch: hence ensues loss of tonus.

It is clear that Ramazzini was aware of the interaction between mental stress and musculotendinous pain 300 years ago. In the 1830s there was an outbreak of writer’s cramp in male clerks working in the British Civil Service, which was attributed to the introduction of the steel nib in preference to the goose-quill pen. During the next 50 years, various authors described this condition of ‘scrivener’s palsy’ or ‘writer’s cramp’ in other occupations, with such names as ‘musician’s cramp’ and ‘shoemaker’s cramp’. At the end of the nineteenth century, Gowers (1845–1915) coined the term ‘occupational neurosis’ for this group of disorders. In the early part of the twentieth century there was an outbreak of cramp in telegraphists in Britain. A government committee that investigated this concluded that it was due to a combination of two factors: one, a nervous ‘instability’ on the part of the operator, and the

other, repeated fatigue during the complicated movements required for sending messages. At that time, it was clear that the association between peripheral pathology and a nervous disposition was recognized. At about the same time, there were descriptions of ‘tenosynovitis’, painful conditions affecting tendons. In 1904, Gowers described a painful inflammatory condition of the muscles and coined the term ‘fibrositis’. He noticed that musculotendinous structures in affected individuals were symptomatic when firmly palpated. This concept was developed by Janet Travell (1901–1997) by defining the ‘myofascial pain syndrome’, in which hypertonic trigger points (‘jump and shout points’) were the essential feature. Interestingly, careful research has established that 70% of known myofascial trigger points correlate precisely with ancient Chinese ‘ah shih’ or acu-points, known to have been described 2000 years ago.

In Japan there was sufficient concern regarding work-related arm pain problems for the Japanese Ministry of Labour to introduce work guidelines in 1964. These included rest periods and maximal work intensities. The Japanese Association of Industrial Health introduced the term ‘occupational cervico-brachial disorder’. The number of workers compensated for these disorders rose rapidly between 1970 and 1975.

In Australia in the 1970s there was an alarming increase in incidence of incapacitating arm pain, subsequently named ‘repetitive or repetition strain injury’ (RSI) by Stone in 1983. There was considerable geographical variation throughout Australia even within the same organization and the same types of work. Much controversy arose to explain this problem. The absence of a histopathological diagnosis was consistent: polarized views were put forward, ranging from psychiatric diagnoses to social and medical iatrogenesis (created by social or health-care policies and practice). In the UK, an influential group of hand surgeons made a review of the RSI problem and made recommendations to the Industrial Injuries Advisory Council in 1990. Their view was that the absence of clear-cut histopathological (‘organic’) disease indicates that there is no physical basis for these conditions, and the disorders that accounted for ‘non-specific diffuse’ arm pain were dismissed. All of these views repre-

sent examples of the duality of orthodox medicine that has persisted from the seventeenth century: i.e. disease can be explained as arising from the body or the mind; the two are quite separate. Western medicine has had great difficulty accepting the interaction of physical and psychological factors in aetiology.

It is clear that the orthodox biomedical model of illness is incapable of explaining these regional pain disorders. Chronic fatigue syndrome is another example of a similar functional disorder around which a great deal of controversy arises and polarized opinions are firmly held by either the ‘in the mind’ or the ‘in the body’ camps.

Back pain

It can be difficult to believe that all through history neither doctors nor patients thought that back pain was due to injury. This idea that it might be came in the second half of the nineteenth century. The industrial revolution, and in particular the building of the railways, led to a spate of serious injuries. Violent trauma could cause spinal fractures and paralysis, so perhaps less serious injuries to the spine might be the cause of backache. There might be cumulative or repetitive trauma. In 1866 the condition of ‘railway spine’ was named: a condition of subjective weakness and disability. For the first time, back pain was linked to injury. Most health-care professionals, patients and lawyers still regard back pain (erroneously) as an ‘injury’, despite the absence of evidence for any tissue damage having occurred in the vast majority of cases.

The discovery of X-rays opened up a whole new perspective on back pain. Soon every incidental radiographic finding became a cause of back pain and sciatica. The same has occurred with modern high-tech investigations such as magnetic resonance imaging (MRI) and computerized tomography (CT).

In Britain, the father of orthopaedics, Hugh Owen Thomas (1834–1891), proposed rest as one of the main orthopaedic principles for the treatment of fractures, tuberculosis and arthritis. This was reasonable in the days before antibiotics and internal fixation; however, these ideas became

important in the management of simple backache and sciatica and persist to this day.

Sciatica has been described since ancient Greek times. In 1934, Mixter and Barr discovered the 'ruptured disc' as the cause of sciatica. This led to the age of the 'dynasty of the disc', where nearly every backache was attributed to 'discs out' or diseased. From the 1950s there was an explosion of disc surgery, closely related to the growth of orthopaedics and neurosurgery. There is to this day evidence that the number of spinal operations for degenerative conditions in different countries is directly proportional to the numbers of spinal surgeons in the country. Major limitations were recognized and, by 1970, spinal surgery was accused of 'leaving more tragic human wreckage in its wake than any other operation in history'. Ignoring the normal age-related changes on radiographs and their poor correlation with symptoms, the disc was blamed for most back pain. The answer was spinal fusion, reinforcing the influence of orthopaedics in the management of simple backache. This approach, still highly prevalent in western medicine, has gravely distorted health care for the 99% of people with back trouble who do not have or need surgery. It caused us to see backache as a purely mechanical or structural problem, and therefore patients expect to be 'fixed'.

The evidence is that back pain has been a twentieth century health-care disaster. Despite vast expenditure on biomedical investigations and treatments, we still do not know the exact cause of most back pain. Worse still is that disability attributable to low back pain in all industrialized countries is getting steadily worse (Box 5.1). In western society, simple back strains disable many

more people than all the serious spinal diseases put together. As with upper limb pain disorders, the orthodox biomedical model of medicine for the treatment of back pain has been seriously limited; indeed, there is much evidence that this approach so frequently contributes in converting a person troubled by simple backache into a back-cripple patient with multiple failed treatments and who has social and occupational disability, but no identifiable pathology to account for it. The Manual Handling Regulations 1992 introduced in the UK were an attempt to reduce the incidence of back pain in the workplace (primary prevention), based on the injury model of understanding back pain disability. Very reasonable in principle, but there is no research evidence yet that this legislation has made any difference to back disability and sickness absence. There is a view that these regulations may be indirectly confounding the problem by focusing the minds of those affected on injury and adding to fear avoidance beliefs and behaviours.

Fortunately, there is a much-needed movement for change; indeed it is a paradigm shift. Waddell, among others, has influenced this. Scientific research has identified psychosocial factors as extremely important in influencing the presentation and course of back pain in an individual [2]. This has been supported by clinical guidelines published: in the UK, the Clinical Standards Advisory Group (1994), the Royal College of General Practitioners (1995) and the Faculty of Occupational Medicine (2000). The importance of adopting a biopsychosocial approach to back pain is emphasized. The adoption of clinical 'red flags' for possible serious pathology (Box 5.2), triage of back pain at presentation (Box 5.1) and the 'yellow

Box 5.1 Back pain triage

- 1 *'simple back pain'*: presenting between the ages of 20 to 55 years, mechanical in nature, lumbosacral, buttocks and thighs, patient well, prognosis good;
- 2 *nerve root pain*: unilateral leg pain worse than low back pain, pain generally refers to foot, numbness or paraesthesia in same distribution, nerve irritation signs, motor sensory or reflex signs, limited to one nerve root: prognosis reasonably good, 50% recover from an attack within 6 weeks;
- 3 possible serious pathology: 'red flags'.

Box 5.2 'Red flags' for possible serious pathology

- age of onset of less than 20 years or greater than 55 years;
- violent trauma, e.g. fall from a height;
- constant, progressive, non-mechanical pain;
- thoracic pain;
- past medical history of carcinoma, systemic steroids, drug abuse, HIV;
- systemically unwell, weight loss or fever;
- widespread or progressive neurological deficit;
- structural deformity;
- sphincter disturbance.

flags' that are risk factors for chronicity (Box 5.3) have all helped to improve the management of back pain.

The neurophysiology and musculoskeletal dysfunction causing 'simple' back pain is far from simple: the term is used to distinguish from serious

pathology. The cause of most cases of simple back pain is altered movement patterns in the spinal motion segments: sometimes segmental stiffness, sometimes segmental instability and usually a combination of both, affecting various levels of the spine. This is called joint dysfunction. Pain

Box 5.3 Psychosocial 'yellow flags'

The risk factors for chronicity and disability for back pain with or without established pathology:

Attitudes and beliefs about back pain

- pain is always harmful;
- pain must be abolished before return to activity;
- catastrophizing, thinking the worst, misinterpreting bodily symptoms;
- belief that pain is uncontrollable;
- passive attitude to rehabilitation.

Behaviours

- withdrawal from normal activities, substituted by down, non-productive time;
- activity intolerance and avoidance;
- poor compliance with exercise, 'all or nothing' approach to exercise;
- reliance on aids or appliances;
- substance abuse: smoking and alcohol especially.

Emotions

- fear of pain;
- depression;
- anxiety, irritability, distress, post-traumatic stress;
- fear of moving (kinesiophobia);
- learned helplessness and hopelessness;
- anger.

Box 5.3 (Continued)*Diagnosis and treatment (iatrogenics)*

- health professionals sanctioning disability;
- conflicting opinions and advice, accepting opinions as fact, unhelpful labelling, e.g. ‘crushed discs’, ‘arthritis in the spine’;
- behaviour of health professionals, excessive unnecessary investigations, dependency on treatments, over-controlling therapists;
- prolonged courses of passive treatments that clearly are not working;
- advice to give up work and avoid pleasurable activities.

Family

- overprotective partner;
- solicitous behaviour from spouse;
- socially punitive responses from spouse, e.g. ignoring;
- lack of support;
- cultural beliefs and behaviours.

Compensation issues

- lack of incentive to return to work;
- history of compensation claim for other health problems;
- disputes over eligibility for benefits, internal conflict: ‘how can you get better if you have to prove you are ill’;
- persistent focus on ‘diagnosis’ and ‘cause’ rather than restoration of function and health;
- ill health retirement benefit issues;
- previous experience of ineffective case management.

Work

- poor job satisfaction, feels unsupported, frequent job changes;
- poor relationship with managers, supervisors, co-workers;
- belief that work is harmful;
- minimal availability of selected or alternative duties, or a graduated return to work: ‘do not come back until you are totally better’;
- low socio-economic status;
- job involves significant biomechanical demands;
- stress at work: e.g. interpersonal, disciplinary, bullying.

can be generated by a number of different tissues maintained by reflexes in the neuromuscular system and biomechanical stresses due to altered spinal mechanics.

A biopsychosocial approach

Also known as a person-centred rather than a disease-based approach. It recognizes the interaction

of biomedical or physical factors with psychological and behavioural influences in addition to social factors, which include family, culture and occupation (Boxes 5.2 and 5.3).

When considering a person who is troubled by a regional pain problem of the upper limb or back, it is helpful to think in terms of *predisposing factors*, *precipitating factors* and *perpetuating factors*. This assists problem solving by identifying clinical, emotional, cognitive, social or environmental

factors that may be influenced by intervention to promote restoration of function and symptom relief.

For example, in a case of chronic regional pain in the upper part of the body, *predisposing factors* may include:

- biomechanical: habitual poor posture, short upper arms, heavy breasts;
- psychosocial: history of anxiety or depression; job dissatisfaction, emotional or physical trauma in childhood (between 30% and 50% of hospital patients with chronic pain and multiple somatic functional complaints have a history of very unhappy childhood experiences).

Precipitating factors may include:

- biomechanical: minor trauma, change of equipment, change of work or recreational routines, excessively long hours of work leading to muscle fatigue;
- psychosocial: poor job satisfaction, interpersonal difficulties, personal problems and negative life events.

Perpetuating factors may include:

- biomechanical: poor posture and workplace ergonomics using display screen equipment, habitual elevation of the arms above shoulder height to operate machinery, tool design and gripping tools excessively hard;
- biomedical: untreated spinal joint or myofascial dysfunctions (so often overlooked), poor health due to infection or coexisting disease, deficiency conditions such as thyroid, vitamin B₁₂, folate and iron, and physical deconditioning due to lack of exercise;
- psychosocial: chronic anxiety, depression, attitudes and beliefs, fatigue, chronic interpersonal problems, low job satisfaction, financial and compensation issues, family influences, fear and avoidance and illness behaviours;
- iatrogenic: health-care professionals sanctioning disability, overinvestigation, conflicting opinions and failed treatments.

Successful management of a person with these disorders demands that the clinician makes a confident diagnosis and identifies perpetuating factors and obstacles to recovery. Communicating with other professionals such as occupational health staff is very important for optimal outcome.

There is no doubt that these disorders are easier to treat in the early stages.

Essential epidemiology

Almost all of us will experience back pain at some time in our life: it is normal and a fact of life. All epidemiology studies indicate that up to 90% of persons between the ages of 18–55 (i.e. of working age) will recall an episode of low back pain that interfered with their ability to function for at least 24 hours at some time. All social and occupational groups are the same. Approximately 40% of us will experience recurring problems with our backs. Most primary care patients who seek treatment for back pain will improve considerably over the first 4 weeks, but only 30% will be pain free. At 1 year 70–80% will still report some recurring back symptoms, one third will have intermittent or persistent pain of at least moderate intensity, and about 15–20% will have a poor functional outcome [3].

The period prevalence of neck and arm pain in the population is similar to low back pain but not as frequently disabling.

Occupational management

There is a greater body of evidence for the management of low back pain at work. However, the same principles apply to the occupational management of regional pain disorders of the upper limb, with the exception that interventions to identify and modify ergonomic perpetuating factors in the workers affected by arm pain are likely to be more important for optimal outcome. Occupational health is primarily about prevention. It is therefore helpful to look at strategies for primary (preventing the problem occurring in the first place), secondary (limiting the consequences) and tertiary (reducing the risk of recurrence after remission) prevention. Given current evidence, secondary and tertiary prevention strategies provide the most effective means of reducing the risk of costly chronicity and disability developing.

Primary prevention

A recent evidence review on back pain at work [4] highlights our current state of knowledge and makes recommendations: essential reading for any occupational health care professional. It is clear that attempts over the past few decades (supported by legislation) to prevent back pain occurring in the workplace have been unsuccessful. The reason for this might be difficulties in research methodology or it might be that the thrust of manual handling policies is based on the injury (bio-medical) model of low back pain causation. The only evidence to date that intervention in the workplace aimed at primary prevention has had any effect on important outcomes such as sickness absence is the promulgation of information-challenging attitudes and beliefs based on a cognitive behavioural model [5–7]. The primary prevention of back pain, and upper limb pain as a symptom, is an unrealistic goal. At present, the most effective strategies for primary prevention in the workplace appear to be good working relationships and probably good ergonomic and job design factors for upper limb pain, although evidence for this assumption remains scarce.

Secondary prevention

The goals are *to prevent disability and chronicity developing*. The occupational health professional is in an ideal position to see workers having difficulty with back pain or upper limb pain early in the course of events, and hence to positively influence the outcome. The first consultation that a person troubled by these symptoms has with a health-care professional is probably the most important and may set the person on the road to recovery and restoration of function, or (as happens all too often) it can precipitate despair, depression and disability. For back pain, know the ‘red flags’ and the principles of triage; for back pain and upper limb disorders, be aware of the ‘yellow flags’ that are risk factors for chronicity and poor outcomes. Aim to influence these ‘yellow flags’ at every opportunity, especially early on, even in the presence of identifiable pathology. Be mindful of the data on return to work after sickness absence for back

pain: after only 3 months just 75% of workers will ever return (Box 5.4).

Physical treatments

Occupational health departments do not normally provide such treatment services, but some do with in house physiotherapists or osteopaths. Many treatments are available but there is a poor evidence base for most clinical interventions, largely for methodology reasons. There is good evidence for spinal manipulation in the first 6 weeks of an episode of back pain, and very strong evidence for activity rather than rest at all stages for back pain in particular. It is important for many reasons to target this treatment service where it is most likely to be effective for the patient and economically for the organization.

When the goal is to reduce long-term sickness absence, it appears that the group of workers to whom available resources are best targeted are those who are off work for between 4 and 12 weeks. Interventions provided to those who are at work but struggling, or beginning to accumulate short spells of work absence, are arguably no less important to help them remain functioning. Those at work and coping with nuisance symptoms that are not interfering with their ability to work are a low priority. Some form of priority has to be given when resources will always be insufficient to meet demand.

Whatever the course of treatment for low back pain, or any other regional pain problem, it is worth remembering that if it is not beginning to make any useful difference to the patient by six treatments, as reflected in improved function, it is not working. Reassess, review the obstacles to recovery and do something different. To prolong ineffective treatments is very damaging to the psychological well-being of the patient.

Tertiary prevention

Much can be done and begins at the first consultation. Some important points are:

- The precipitating and perpetuating factors for regional pain disorders are many, therefore management needs to be multimodal and may involve a number of professionals.

Box 5.4 Upper limb pain: clinical types*Those with usually clear-cut histopathological diagnosis*

- cervical nerve root pain;
- tendinopathies, e.g. stenosing tenosynovitis of the first dorsal compartment of the wrist* or the flexor tendons of the fingers;
- nerve entrapment in the carpal tunnel of the wrist[†];
- arthropathies, e.g. osteoarthritis of the carpo-metacarpal joint at the base of the thumb;
- enthesopathies, e.g. degenerative tendon anchor points on bone in some types of ‘tennis elbow’.

Those without clear-cut histopathological diagnoses, i.e. neuromuscular and joint dysfunctions

- myofascial pain: trigger point activation in muscle groups exposed to habitual awkward postures and tensions, usually isometric muscle contraction, extremely common[‡];
- joint dysfunctions in the neck and thoracic spine with somatic referred pain;
- focal dystonias, e.g. ‘writer’s cramp’[‡];
- thoracic outlet syndrome[‡];
- complex regional pain syndrome type 1 (formerly called ‘reflex sympathetic dystrophy’), a neuropathic pain disorder;
- whole-body pain and fatigue, often called ‘fibromyalgia’.

*There is good evidence that this is attributable to occupations that require repeated forceful twisting and gripping movements.

[†]There is good evidence that this is occupationally acquired in workers who habitually are, or have been, exposed to hand–arm vibration from tools.

[‡]These disorders have, beyond reasonable doubt, occupational precipitating and perpetuating factors; unfortunately, good research evidence is lacking because of reproducible diagnostic difficulties and methodological problems.

- Empower the person to participate in their active rehabilitation and to take responsibility for maintaining any programme of management.
- Return to work as soon as possible; there is no need to wait until all the pain has gone.
- Return to *normal* work must be the goal: this reinforces the patient’s belief that normality can be achieved. It also, and vitally, reduces fear avoidance beliefs and behaviour.
- If the person is off work, a fixed period of return to work, gradually increasing activity and responsibility is desirable to assist the return to normal activities. This must be time limited, with goals set and reviews arranged. Some form of temporary restrictions may be helpful, but these must also be time limited. It is a mistake to allow restrictions to depend on ‘how the patient feels’: this encourages pain and illness behaviour and only creates more

problems in the future, which are even more difficult to solve.

- Treatments must facilitate active rehabilitation and not interfere with it.
- Do a workplace assessment. Interventions on ergonomic and work practices may be particularly important in cases of upper limb pain. Encourage changes of position and opportunities to stretch and move muscles that have to work in static positions for long periods of time.
- Consider short spells in functional restoration programmes for those who have demonstrated commitment to work hard to improve their functional capacity, but are having difficulty. These programmes are not suitable to send people to in the hope that they can magically motivate a person who has learned helplessness, is depressed, is focused on compensation issues of whatever nature

and has no belief that their life quality or occupational status can be improved.

- Support and encourage the person through the difficulties and setbacks that will inevitably occur. Occupational health staff are in an ideal position to provide this.
- Consider redeployment or severance of employment contract only when all reasonable attempts have been made to rehabilitate to normal work. Case law following the introduction of the Disability Discrimination Act 1995 in the UK indicates that some persons with chronic musculoskeletal symptoms may have a disability under the provisions of this Act.
- Liaise with all health-care professionals involved in the case. Be prepared to take a lead in case management. Seek other opinions if you believe the patient will benefit.

Hand–arm vibration syndrome

Prolonged and regular exposure of the fingers or the hands to vibrating tools can give rise to various signs and symptoms of this disorder. The disorder may comprise vascular or neurosensory effects, or a combination of both.

The vascular effects (formally known as vibration white finger) are characterized by episodic

blanching of the fingers. Attacks of blanching are usually precipitated by cold and continue until the fingers are warmed. One standardized means of recording the clinical severity is the Stockholm scale. There are various ways to provoke symptoms to help verify the extent of disease, such as cold provocation.

The neurological effects are typically numbness, tingling, elevated sensory thresholds for touch, vibration, temperature and pain, and reduced nerve conduction velocity. There are currently no specific objective tests to stage the disease.

There is a roughly linear relationship between exposure to the hazard and development of disease. There are other medical factors such as smoking and others unknown that might predispose to developing the problem. Once the symptoms start to appear there is no treatment for the disease. Progress can be limited by reducing or avoiding exposure to hand–arm vibration. Primary prevention must be the goal by reducing exposure to the hazard as far as is reasonably possible along standard occupational hygiene principles, and by health surveillance for early case detection.

In the UK, the disease of hand–arm vibration syndrome entitles the sufferer to apply for no-fault compensation from the government through the industrial injuries prescribed disease system (PD A11) (Box 5.5).

Box 5.5 Prescribed industrial diseases in the UK (musculoskeletal)

- A4. Cramp of the hand or forearm due to repetitive movements in any occupation involving prolonged periods of handwriting, typing or other repetitive movements of the hand or forearm.
- A8. Traumatic inflammation of the tendons of the hands or forearm, or of the associated tendon sheaths in manual labourers or those exposed to repeated or prolonged forceful movements of the wrist.
- A11. Episodic blanching occurring throughout the year, affecting the middle or proximal phalanges or, in the case of the thumb, the proximal phalanx. In occupations exposed to habitual use of hand-held vibrating tools.
- A12. Carpal tunnel syndrome in occupations exposed to hand-held vibrating tools.

Conditions A4 and A8 continue to be sources of much controversy, mainly because of the difficulty establishing valid and reliable diagnostic criteria that are free from subjective interpretation.

References

- 1 Waddell, G. (1998). *The Back Pain Revolution*. Churchill Livingstone, London.
- 2 Linton, S.J. (2000). A review of psychological risk factors in back and neck pain. *Spine*, **25**, 1148–56.
- 3 Croft, P.R., Macfarlane, G.J., Papageorgiou, A.C., Thomas, E. and Silman A.J. (1998). Outcome of low back pain in general practice: a prospective study. *BMJ*, **316**, 1356–9.
- 4 Faculty of Occupational Medicine (2000). *Occupational Health Guidelines for the Management of Low Back Pain at Work: Evidence Review and Recommendations*. Faculty of Occupational Medicine, London.
- 5 Symonds, T.L., Burton, A.K., Tillotson, K.M. and Main, C.J. (1995). Absence resulting from low back trouble can be reduced by psychosocial interventions at the workplace. *Spine*, **20**, 2738–45.
- 6 Buchbinder, R., Jolley, D. and Wyatt, M. (2001). Population based intervention to change back pain beliefs and disability: three part evaluation. *BMJ*, **322**, 1516–20.
- 7 Roland, M., Waddell, G., Moffett, J.K., Burton, A.K., Main, C.J. and Cantrell, E. (1997). *The Back Book*. The Stationery Office, Norwich.

Further reading

- Baldry, P.E. (2001). *Myofascial Pain and Fibromyalgia Syndromes: a Clinical Guide to Diagnosis and Management*. Churchill Livingstone, London.
- Clinical Standards Advisory Group (CSAG) committee (1994). *Back Pain Report*. HMSO, London.
- Hadler, N.M. (1999). *Occupational Musculoskeletal Disorders*, 2nd edn. Lippincott Williams & Wilkins, Baltimore.
- Hutson, M.A. (1997). *Work-Related Upper Limb Disorders Recognition and Management*. Butterworth Heinemann, London.
- Main, C.J. and Spanswick, C.C. (2000). *Pain Management: an Interdisciplinary Approach*. Churchill Livingstone, London.
- Simons, D.G., Travell, J.G. and Simons, L.S. (1999). *Myofascial Pain and Dysfunction: The Trigger Point Manual*, Vol. 1. *Upper Half of Body*, 2nd edn. Williams & Wilkins, Baltimore.

Chapter 6

The effects of inhaled materials on the lung and other target organs

Jon G. Ayres

Introduction

Deposition of particles and gases in the respiratory tract

Airflow in the respiratory tract

Mechanisms of particle deposition

Impaction

Sedimentation

Diffusion

Alveolar clearance

Deposition and uptake of gases

The lung

Irritation of the airways

Diseases of the airways

Asthma

Chronic obstructive pulmonary disease and bronchitis

Byssinosis

Pneumoconiosis

Silicosis

Coal-miners' pneumoconiosis

Benign pneumoconiosis

Asbestosis

Extrinsic allergic alveolitis

Malignant disease

Asbestos

Other carcinogens

Effects on other target organs

The nervous system

Peripheral neuropathy

Narcosis

Mental changes

Defects in balance and vision

The liver

The kidney

Heavy metals

Solvents

The cardiovascular system

The blood

Further reading

Introduction

Many hazardous substances encountered in the workplace gain entry to the body through inhalation. Some may be absorbed through the skin or ingested, although ingestion is rarely an important route of entry in occupational health. Inhaled materials may affect the lung directly, or may be absorbed from the lung and affect other parts of the body. Consequently, a large section of this chapter is devoted to the lung as it is the target organ of greatest importance.

Deposition of particles and gases in the respiratory tract

Particle deposition throughout the respiratory tract depends on a variety of factors, including particle size (Fig. 6.1), ventilation rate and the presence of disease.

At rest, most respiration is nasal. The nose warms and humidifies inspired air, clears larger particles and absorbs the more soluble gases (e.g. sulphur dioxide). Once past the nose, air passes through the glottis (comprising the false and true

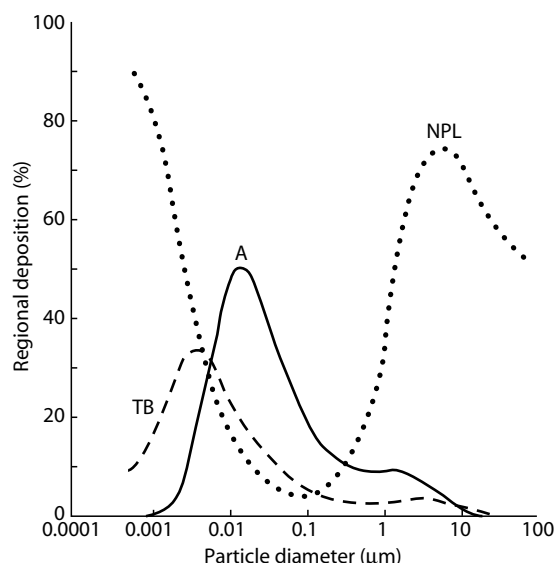


Figure 6.1 International Committee on Radiological Protection (ICRP) model for lung deposition of particles (1994). A, alveolar deposition; TB, tracheobronchial deposition; and NPL, deposition in nose, pharynx and larynx.

vocal cords) and into the tracheobronchial tree. This system consists of 25 generations of branching airways, reaching a diameter of around 200 μm , terminating in the alveoli, where oxygen uptake and carbon dioxide excretion occur. The alveolar region provides an enormous surface area for the deposition of particles and the uptake of gases.

Airflow in the respiratory tract

During quiet breathing, tidal volumes of around 550 cm^3 are inhaled at a rate of around 10 breaths min^{-1} , a minute ventilation of 5–6 l min^{-1} . At low flow rates, the majority of a tidal volume breath goes to the base of the lung but, as flow rates increase (e.g. during exercise), all regions of the lung are near equally ventilated. Consequently, there is a very wide range of flow conditions throughout the lung at any one time. The presence of airflow limitation due to disease [e.g. asthma, chronic obstructive pulmonary disease (COPD)] will affect deposition, particularly regionally as these disease states do not affect airflow homogeneously.

Mechanisms of particle deposition

Three main processes affect particle deposition in the lung – impaction, sedimentation and diffusion – but other physical characteristics of the particle, notably density, size, charge and water solubility, also have an effect. The denser a particle, the more likely it is to impact or sediment; smaller particles will deposit deeper, whereas charged particles can influence deposition by altering the charge in the mucosa and lining fluid. If the inhaled material is hydrophilic, particles can grow in size as they pass down the airways unless they are extremely small. The net effect of this growth on overall deposition is highly complex and will depend strongly on the original size of the inhaled particles.

Impaction

Impaction is a function of the inertia of the inhaled particle. The most important sites for impaction are in the nose, the glottis and the larger bronchial bifurcations (Fig. 6.2).

Sedimentation

Particles smaller than about 1 μm in diameter sediment rapidly because they tend to slip between the air molecules as the particle diameter gets smaller. Particles larger than about 30 μm sediment more slowly because their higher inertia sets up complex eddies in the surrounding gas. The aerodynamic diameter is a measure of the inhalability of a particle or fibre and is defined as the diameter of a sphere of unit density that would sediment at the same velocity as the particle. For instance, a fibre of unit density, whose length is at least ten times its diameter, will have an aerodynamic diameter only approximately three times greater than its physical diameter. Consequently, fibres can persist in the air very much longer than spherical particles of the same mass and therefore have a much greater likelihood of being inhaled. Only particles with diameters less than 5 μm reach the alveoli, with the exception of fibrous particles that can travel in a streamlined way along the conducting airways.

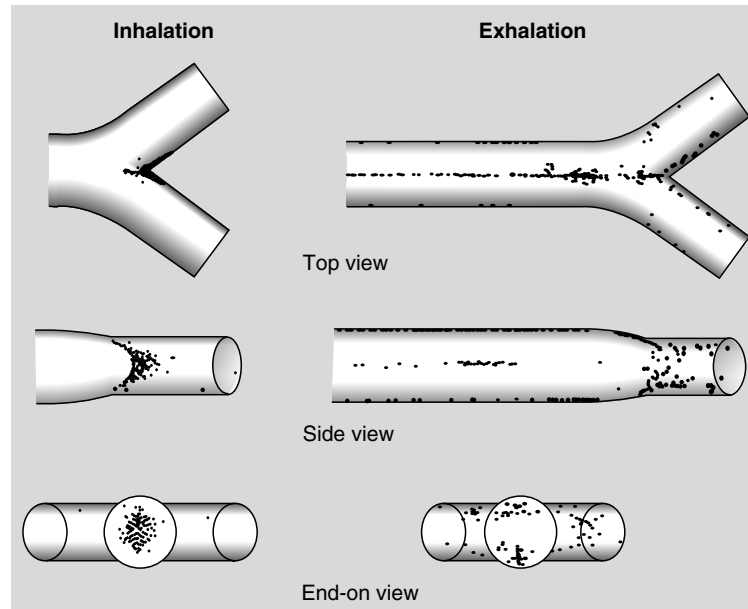


Figure 6.2 The spatial distribution of the deposition of 10- μm particles in a bifurcation. During inhalation, deposition occurs predominantly on the carina, but during exhalation it occurs in a different distribution (adapted from Balashazy, I. and Hoffman, W. (1993). Particle deposition in airway bifurcations I. Inspiratory flow, II. Expiratory flow. *Journal of Aerosol Science*, 24, 745–86).

Diffusion

Very small particles diffuse by Brownian motion, depositing on any surface. For instance, a 0.1- μm -diameter particle can diffuse up to 150 μm during a normal 4-s breath (i.e. the radius of a small airway), but a 1- μm -diameter particle can only diffuse up to 30 μm in the same period.

Alveolar clearance

Clearance of particles from the alveolar region by macrophages occurs only slowly, some thus remaining in the lung for long periods. For particles of less than 0.1 μm , alveolar deposition normally exceeds 50% of that inhaled.

Deposition and uptake of gases

Diffusion, convection, the partial pressure and tissue solubility of the gas and pulmonary blood flow all influence gas deposition and uptake in the lung. With persistent exposure, an equilibrium will become established between the rate of uptake of gas and the rate at which it is cleared by the blood, although this varies considerably according to the gas. Vapours like benzene are relatively insoluble in

blood so, even although enough vapour dissolves to bring the pulmonary blood into equilibrium with the gas phase, only a small fraction of that which is inhaled is absorbed. In contrast, when a gas like carbon monoxide, which reacts with haemoglobin, enters the alveoli a very large proportion of the material will be lost from the gas phase. Hence, alveolar concentration will fall markedly during the course of a breath. High ventilation rates will enhance alveolar uptake of soluble gases while some will be taken up through the bronchial mucosa, which may mean that transient exposure could result in significant uptake of such gases.

The lung

The lung, in common with the other organs of the body, has a limited capacity to respond to toxic materials. The ways in which the lung is affected by environmental exposures can be broadly divided into four categories.

Irritation of the airways

A number of gases and fumes produce intense irritation of the airways (Table 6.1). Symptoms

Table 6.1 Effects of inhalation of selected irritant gases or fumes.

<i>Substance</i>	<i>Sources</i>	<i>Acute effects</i>	<i>Chronic effects</i>
Ammonia	Production of fertilizers and explosives, refrigeration, manufacture of plastics	Pain in eyes, mouth and throat, oedema of mucous membranes, conjunctivitis, pulmonary oedema	Airways obstruction, usually clears in about a year
Chlorine	Manufacture of alkali, bleaches and disinfectants	Chest pain, cough, pulmonary oedema	Usually none, occasionally causes airways obstruction
Sulphur dioxide	Paper production, oil refining, atmospheric pollutant	As for ammonia	Chronic bronchitis
Nitrogen oxides	Silo filling, arc welding, combustion of nitrogen-containing materials	Pulmonary oedema after lag of 1 or 2 h; obliteration of bronchioles in severe cases after 2–3 weeks	Permanent lung damage with repeated exposure
Phosgene	Chemical industry, First World War gas	Pulmonary oedema after a lag of several hours	Chronic productive bronchitis
Ozone	Argon-shielded welding	Cough, tightness in the chest, pulmonary oedema in severe cases	None
Mercury	Chemical and metal industries	Cough and chest pain after lag of 3–4 h; acute pneumonia	Usually none, pulmonary fibrosis rarely
Osmium tetroxide	Chemical and metal industries, laboratories	Tracheitis, bronchitis, conjunctivitis	None
Solvents (e.g. styrene)	Many different industries, cleaning processes	Bronchitis, exacerbation of asthma, conjunctivitis, laryngitis	None
Vanadium pentoxide	Ash and soot from oil	Nasal irritation, chest pain, cough	Bronchitis, bronchopneumonia
Zinc chloride	Manufacture of dry cells, galvanizing	Tracheobronchitis	None

vary depending upon which parts of the airways are affected, partly a function of the solubility of the material. For example, highly soluble gases, such as ammonia, produce immediate effects on the upper respiratory tract (and the eyes), causing pain in the mouth, throat and eyes, due to swelling and ulceration of the mucous membranes. These symptoms are so intensely unpleasant that affected individuals will immediately try to remove themselves from exposure. Continual exposure, or a single exposure to a very high concentration, will result in the smaller airways becoming affected, leading to inflammation and oedema in the bronchiolar and alveolar walls. The presence of fluid in the alveoli (pulmonary oedema) seriously interferes with gas exchange and can be fatal if not

adequately treated. By contrast, a relatively insoluble gas such as phosgene produces no immediate effects on the upper respiratory tract, but does induce profound pulmonary oedema after a delay of several hours. Some of the pulmonary irritants will also cause permanent lung damage if exposure is particularly high or frequently repeated, whereas others appear to predispose individuals to other conditions such as pneumonia or chronic airway inflammation.

Diseases of the airways

Diseases affecting the airways reduce flow through them because of airway narrowing. In lung function terms, this means that while the forced vital

capacity (FVC) is either maintained or somewhat reduced, the forced expired volume in one second (FEV₁) is reduced. In other words, airway narrowing results in a smaller proportion of the FVC being expelled in the first second of expiration, an obstructive ventilatory defect.

Asthma

Occupational asthma is common but often remains undiagnosed. There are over 200 recognized causes of occupational asthma, some of which are shown in Table 6.2, but the list is continually en-

larging. True occupational asthma where the agent causes asthma *de novo* must be separated from work-related asthma, when someone with asthma finds their condition exacerbated by their work conditions. The underlying mechanism of occupational asthma is either a hypersensitivity to the agent, usually mediated through immunoglobulin E (IgE), or is irritant induced (once known as RADS, reactive airways dysfunction syndrome), when the molecular mechanism is unclear. Some agents can induce very strong allergic responses mediated through IgE (such as platinum asthma and laboratory workers' asthma), whereas in

Table 6.2 Examples of agents producing occupational asthma.

Groups of agents	Industry/occupation	Comments
Aldehydes, e.g. formaldehyde, glutaraldehyde	Various industrial processes	Glutaraldehyde sensitivity is increasingly common in the health sector
Aluminium	Health-care workers Aluminium smelting	Aluminium pot-room asthma
Antibiotics: antibiotic itself or constituent such as gum acacia	Antibiotic manufacturing industry	The most important of the drugs known to cause occupational asthma
Flour and amylase	Baking	Commonest cause of occupational asthma in Germany
Colophony: resin acids within solder flux	Solderers	Much reduced since colophony removed from fluxes
Enzymes, e.g. alcalase	Manufacturers of biological detergents	IgE mediated
Epoxy resins: acid anhydrides, triethylene tetramine	Anywhere where resins are used	Antibody formation is more marked in cigarette smokers
Grain dust: probably mostly due to moulds in the dust	Farmers, grain workers, etc.	Can also result in COPD, allergic alveolitis and rhinitis
Isocyanates: toluene, diphenylmethane and hexamethylene di-isocyanates (TDI, MDI, HDI)	Automotive industry (paint spraying using two-pack polyurethane paints)	Used as hardeners in paints by cross-linking. Once sensitized, some are exquisitely sensitive. After removal from exposure long-term deterioration in lung function can occur
Laboratory animals and arthropods	Laboratory workers	IgE-mediated reactions to proteins in animal urine
Latex	Health-care workers	Increasingly common, IgE mediated, cross-sensitization to bananas and mangoes, etc.
Platinum salts, e.g. ammonium hexa- and tetra-chlorplatinate	Platinum refining and reprocessing industries	Highly allergenic metal, IgE mediated
Wood, e.g. western red cedar	Timber industry, wood processing and manufacturing industries	Other hard woods also a problem but not soft woods

For a complete list see Hendrick *et al.* (2002).

others the mechanisms are not easily explained by a single mechanism (e.g. Baker's asthma, isocyanate asthma).

Classically, symptoms of occupational asthma are worse during the working week and improve at the weekend or when away from exposure for a longer period such as on holiday. However, although symptoms may follow immediately upon exposure to the antigenic material, there is often a delay of several hours, symptoms developing during the evening or at night. When a history is suggestive of occupational asthma, confirmation should be attempted initially by using domiciliary peak flow readings (a simple form of self-recorded lung function test). Final confirmation may be necessary in the laboratory, using bronchial challenge testing to the offending substance.

Chronic obstructive pulmonary disease and bronchitis

Occupational COPD is more common than had once been thought, although in many cases specific causes are not able to be identified. The exposures that result in COPD are mining, silica, cadmium fume (an important cause of emphysema), welding fume, and cotton, grain and wood dusts. The fact that welding fume and grain and wood dusts can also result in occupational asthma raises the possibility that these conditions may share common causal pathways. Workers exposed to so-called nuisance dusts can also show loss of lung function over time.

Byssinosis

This condition, due to inhalation of cotton dust, is the best recognized example of an obstructive airways disease of occupational origin but perhaps the least understood. The condition is now rare in developed countries, although it remains common in the developing world. Characteristically, symptoms are first noted on Monday mornings on returning to exposure. There may be an interval of several years from first exposure to the onset of symptoms, which, in the early stages, may only last a day or two but become continuous with prolonged exposure. Symptoms will disappear if ex-

posure to cotton dust is discontinued early enough, which will not happen if the disease has progressed to the later, more fibrotic stage.

Pneumoconiosis

The term pneumoconiosis literally means 'dusty lungs' and, as such, conveys no connotation of harm. For medical purposes, however, the term should be confined to mean permanent alteration of lung structure following the inhalation of mineral dust, and the tissue reactions of the lung to its presence. The dusts that are most harmful to the lungs are silica (or quartz), coal dust and asbestos. This broad group of conditions generally affects the lung interstitium rather than the airways and therefore results in a restrictive defect on lung function testing, with both FEV₁ and FVC being proportionately reduced.

Silicosis

This is probably the oldest of all the occupational diseases and follows exposure to fine crystalline silicon dioxide or quartz. In the past, the disease was common in many industries including mining, quarrying, the pottery industry, iron and steel foundries and sand blasting. In recent years, the number of new cases has fallen sharply in the UK as the result of improved working practices and from the substitution of safer materials for silica wherever possible.

Inhalation of silica leads to the formation of small nodules of fibrotic tissue (around 1 mm in diameter), which increase in size and coalesce as the disease progresses. These nodules are seen on a chest radiograph as small, round opacities scattered predominantly in the upper parts of the lung, and may be present before any symptoms of breathlessness appear.

Early diagnosis in cases of silicosis is essential, as the symptoms may progress even after exposure has ceased if a sufficiently large quantity of dust has been inhaled. Progression is marked by increasing difficulty in breathing, and death frequently results from combined heart and lung failure. An unexpectedly large number of patients with silicosis also develop tuberculosis.

Coal-miners' pneumoconiosis

This condition produces a less severe picture than silicosis and is less aggressively fibrotic. It may be subdivided into simple pneumoconiosis and progressive massive fibrosis (PMF).

The diagnosis of simple pneumoconiosis is made on the finding of small, round opacities in the lung and the degree of radiographic change is closely related to the total amount of dust inhaled. Beyond a slight cough, which produces blackish sputum, simple pneumoconiosis causes virtually no symptoms; its importance lies in the fact that in some individuals it is the precursor of PMF. Why simple pneumoconiosis progresses to PMF in some miners is not clear but, in PMF, increased production of various growth factors, such as fibroblast growth factor, may play a part.

Radiographically, PMF is characterized by large, often irregular opacities, usually in the upper part of the lung. In the lung itself, these areas appear as hard, black masses, often with a central cavity filled with jet black fluid, which may be coughed up as inky black sputum. Moderate or severe forms of PMF can cause significant disability and lead to premature death.

Benign pneumoconiosis

The diagnosis of benign pneumoconiosis is entirely dependent upon radiological findings as, by definition, the patient is symptom-free. Examples are tin (stannosis), barium (barytosis), iron (siderosis) and antimony.

Asbestosis

This is associated with fibrotic changes in the lung (which may become extremely severe) and with the development of cancer of the bronchus or mesothelioma (see below).

The development of pulmonary fibrosis, or asbestosis, mainly follows exposure to white asbestos (chrysotile), although blue (crocidolite) and brown (amosite) asbestos are also fibrogenic. The patient complains of increasing shortness of breath and often a dry cough. The radiographic changes are usually confined to the lower parts of the lung

and show up as linear shadows that become larger and more irregular as the disease progresses.

Asbestos bodies are commonly found in the sputum of individuals exposed to asbestos. They are rod-like structures, 20–150 mm in length, often with a beaded appearance. The presence of asbestos bodies in sputum can be taken as evidence of exposure to asbestos but it does not imply disease. Similarly, the calcified pleural plaques seen on a chest radiograph are a sign of exposure to asbestos, but are not by themselves indicative of ill effects.

Extrinsic allergic alveolitis

Inhaled organic compounds may either cause asthma (as described above) or cause inflammation in the alveoli, due to an extrinsic (external) allergen. A wide range of substances may cause this condition (Table 6.3), the most common being fungal spores. Symptoms and the clinical picture are similar across these conditions, which are usually identified by the occupation (e.g. farmer's lung, bird fancier's lung). Symptoms typically start a few hours after exposure, with fever, fatigue and shivering. As the condition becomes more entrenched, the patient complains of breathlessness and cough. Removal from exposure usually results in clearing of symptoms over a day or two, although persistent or repeated exposure can lead to fibrosis and chronic symptoms. Lung function tests show a reduction in lung volumes and gas transfer. Chest radiograph shows generalized shadowing.

Malignant disease

Asbestos

Many individuals exposed to asbestos die from lung cancer. Smokers who are exposed to asbestos are at a very much greater risk of developing lung cancer than are people who only smoke or only have asbestos exposure. Pleural tumours (mesotheliomas) are rare and seem to occur almost entirely in those who have been exposed to blue asbestos (crocidolite). The tumour spreads from the pleura into the underlying tissues and is inevitably fatal. There may be a delay of up to 50 years between the

Table 6.3 Some types and causes of extrinsic allergic alveolitis.

<i>Clinical condition</i>	<i>Due to exposure to</i>	<i>Allergen</i>
Farmer's lung	Mouldy hay	Thermophilic actinomycetes
Bird fancier's lung	Bird droppings	Protein in the droppings
Bagassosis	Mouldy sugar-cane	<i>Thermoactinomyces vulgaris</i> and <i>T. sacchari</i>
Malt worker's lung	Mouldy malt or barley	<i>Aspergillus clavatus</i>
Suberosis	Mouldy cork dust	<i>Penicillium frequentans</i>
Maple bark stripper's lung	Infected maple dust	<i>Cryptostroma corticale</i>
Cheese washer's lung	Mouldy cheese	<i>Penicillium casei</i>
Wood pulp worker's lung	Wood pulp	<i>Alternaria</i> species
Wheat weevil disease	Wheat flour	<i>Sitophilus granarius</i>
Mushroom worker's lung	Mushroom compost	<i>Micropolysporon faeni</i>
Animal handler's lung	Dander, dried rodent urine	Serum and urine proteins
Pituitary snuff-taker's lung	Therapeutic pituitary snuff	Pig or ox protein
Air-conditioner disease	Dust or mist	<i>T. vulgaris</i> , <i>T. thapophilus</i> and amoebas (various)

first exposure and the onset of symptoms. Mesothelioma of the abdominal cavity (peritoneum), cancer of the larynx, and perhaps cancer of the ovary are also linked to asbestos exposure.

Other carcinogens

Coke oven workers who have been exposed to high levels of polycyclic aromatic hydrocarbons have a near threefold increased risk of lung cancer and miners subjected to ionizing radiation are also at significant risk. The radioactive source in mines is radon or its daughter products (polonium-218, -214 and -210). They all emit alpha particles that have a penetration range in tissue cells of between 40 and 70 nm, which is just sufficient to enable them to damage the basal cells of the bronchial epithelium, resulting in malignant transformation (see also Chapter 22).

Other carcinogens are arsenic, cadmium, chrome, nickel and paint spray, exposures that occur across a wide range of occupations. Exposure to certain hexavalent, but not trivalent, chrome salts increases the likelihood of contracting lung cancer.

Adenocarcinoma of the nasal sinuses was first described in woodworkers in the furniture industry in High Wycombe and in leather workers in the Northampton shoe trade. This is an otherwise rare tumour, and came to light when clinicians became

aware that they were seeing an unusually large number of patients with uncommon diseases. Subsequent epidemiological studies were able to confirm the original clinical observations and identify the exposure sources but not the specific carcinogenic agents.

Recent epidemiological studies have shown a consistent excess of lung cancer in patients with silicosis. Whether silica alone or in combination with fibrosis and/or smoking is to blame remains unresolved.

Effects on other target organs

Materials that are inhaled and absorbed from the lungs may be distributed to other organs whose function may be adversely affected either directly or after metabolic transformation, normally in the liver. Most of the metallic poisons (e.g. lead, cadmium and metallic mercury) are directly toxic, whereas, for example, carbon tetrachloride becomes toxic only when it has been hydroxylated to the highly reactive CCl_3 radical. Only cells capable of hydroxylation are subject to the toxic effects of carbon tetrachloride. Similarly, polycyclic aromatic hydrocarbons all require metabolic transformation by the enzyme aryl hydrocarbon hydroxylase before their carcinogenic potential can be realized.

Many toxic materials must be metabolized in the liver before they or their metabolites can be excreted in the urine, as the kidney only excretes water-soluble molecules. The liver transforms insoluble molecules into soluble molecules, frequently by conjugation with glucuronic acid. The metabolism of trichloroethylene (TRI) is an example. Following absorption, TRI is converted to chloral hydrate, which is further metabolized via two routes. One of these involves a rapid reduction to trichloroethanol (TCE) and the other involves a slow oxidation to trichloroacetic acid (TCA). TCE is not water soluble and any that is not metabolized to TCA is conjugated with glucuronic acid and the conjugate is excreted by the kidney. TCA, being water soluble, is excreted directly. Exposure to TRI vapour can be monitored by the estimation of the rate of excretion of TCA and TCE in the urine.

The body's responses to toxin exposure are limited, because the various organs and organ systems have a finite capacity to respond to damage. Thus, we have considered the toxic effects by individual target organ, rather than listing the effects of individual toxins, as the individual affected worker will complain of target organ-based symptoms.

The nervous system

That part of the nervous system over which we have some control is conventionally divided into the central nervous system (CNS), comprising the brain and spinal cord, and the peripheral nervous system, consisting of sensory and motor nerves that convey information to and from the brain. Involuntary activity is controlled by the autonomic nervous system, which also has central and peripheral components. Materials damaging the peripheral nervous system can produce defects in motor function, manifested by a partial or total decrease in muscular activity, or loss of sensation, or both. The number of neurotoxic materials that are likely to be encountered in industry is significant (Table 6.4) and, with the exception of lead, each produces a mixed motor and sensory loss. Some of the neural symptoms due to specific causative agents are shown in Table 6.5.

Table 6.4 The most common neurotoxins.

Organophosphate pesticides
Carbamate pesticides
Triorthocresyl phosphate
<i>n</i> -Hexane
Methylbutylketone
Acrylamide and/or dimethylaminopropionitrile (DMAPN)
Carbon disulphide
Mercury compounds (inorganic and organic)
Lead and its compounds (inorganic)
Arsenic
Thallium
Antimony

Peripheral neuropathy

Lead is exceptional among the occupational neurotoxins in that it produces a pure motor neuropathy, caused largely through structural damage, which interferes with nerve impulse conduction. The organophosphates produce their effects, however, not by damaging the integrity of the nerve cells but by prolonging the effect of acetylcholine, thus preventing the normal passage of the nerve

Table 6.5 Changes in central nervous system function due to industrial toxins.

Symptom	Caused by
Narcosis	Organic solvents
	Trichloroethylene
	Carbon tetrachloride
	Chloroform
	Benzene
	Carbon monoxide
	Carbon dioxide
Mental changes	Hydrogen sulphide
	Carbon disulphide
	Manganese
	Mercury
Epilepsy	Tetraethyl lead
	Chlorinated naphthalenes, e.g. aldrin, dieldrin, endrin
	Methyl mercury
Defects in balance and vision	Solvents (e.g. styrene)
	Manganese
Parkinsonism	Carbon disulphide
	Carbon monoxide

impulses. There have been some reports that lead and carbon disulphide may produce subclinical nerve damage, which means that impairment in function occurs of which the worker is unaware. It is not known whether subclinical changes progress to produce frank clinical effects with continued exposure, nor how far performance is affected. In the case of lead, there is no clear correlation between subclinical neuropathy and impairment of physiological or psychological tests and so its importance is open to doubt.

Narcosis

Volatile organic solvents produce a range of symptoms ranging from headache and dizziness to unconsciousness or death. Other asphyxiant agents of importance in industry, such as the chemical asphyxiant carbon monoxide or the simple asphyxiant carbon dioxide, are in some ways more insidious than the organic compounds because they cannot be detected by smell and so give no warning of impending danger. The converse is true of hydrogen sulphide as the foul smell of this gas ensures that exposure is minimized, although 'smell fatigue' can set in.

Mental changes

The mental changes induced by carbon disulphide became apparent soon after the compound was introduced for the cold curing of rubber. Workers heavily exposed began to demonstrate bizarre behaviour even to throwing themselves through windows. Today, levels of exposure are minimized and frank psychotic syndromes are unknown in carbon disulphide workers.

During the nineteenth century, pelters exposed to hot mercuric nitrate were frequently affected by a condition known as *erethism*. They became quarrelsome, easily upset by even the mildest criticism, and liable to verbal and physical outbursts. One of the best descriptions of a patient with this syndrome is the Mad Hatter in *Alice in Wonderland* but this condition is no longer seen.

Psychotic signs may precede other signs in manganese poisoning and manganese madness occurs

relatively frequently among miners in Chile. Cases of tetraethyl lead poisoning in this country are unheard of now, and cases reported from abroad have usually come about from the inappropriate use of leaded petrol as a solvent. In the USA, a number of cases of lead encephalopathy have been found in children sniffing petrol for kicks, some of whom have died as a result. The chlorinated naphthalenes are used as insecticides and do not produce toxic symptoms in normal working concentrations. Epileptic fits have been induced, however, in men who have been heavily exposed during the manufacture of these compounds.

Whether chronic low-dose exposure to various organic solvents can cause organic psychosis remains unresolved. The 'Danish painters' syndrome'—much vaunted in the early 1980s—has not been corroborated in recent American or British studies. The best of these studies, however, show a decremental change in psychological tests following long occupational exposure to organic solvents. The clinical and epidemiological significance of such results remain to be evaluated.

Defects in balance and vision

Methyl mercury may damage the cerebellum and the occipital cortex, resulting in problems with balance and vision. Poisoning is rare in industry and most cases have resulted from environmental accidents. The Minamata Bay disaster was caused by the discharge of industrial waste containing inorganic mercury into the sea where it was methylated by micro-organisms in the sea bed. The methyl mercury so formed entered the food chain, concentrating in the fish that formed a substantial part of the diet of the population living around the bay. Other outbreaks of methyl mercury poisoning have occurred when seed grain dressed with the compound as a fungicide was eaten instead of sown. These episodes, particularly those that occurred in Iraq, affected many hundreds of people, large numbers of whom died.

Recently, a specific syndrome due to solvent exposure, comprising tremor, memory loss and colour blindness, has been described. The long-term outlook for these individuals is not yet known.

Toxin-related parkinsonism is very similar to Parkinson's disease, a common affliction in the elderly. The 'naturally' occurring form results from a depletion of the neurotransmitter dopamine in the basal ganglia. Manganese is the most important of the toxic agents that cause Parkinson-like symptoms, but the means by which the disorder is induced is not clear. As treatment with drugs that control the symptoms in the 'natural' variety are also effective in the toxic states, the underlying mechanism is presumably similar in both.

The liver

Liver function is adversely affected by a relatively small number of organic solvents, prominent among which is carbon tetrachloride. Massive exposure to carbon tetrachloride results in the death of large numbers of liver cells and the patient becomes deeply jaundiced and may die. Because of the regenerative capacity of the liver, liver function normalizes provided that the patient can be helped over the acute phase, and that any concomitant kidney damage does not prove fatal. Other industrial chemicals that can produce jaundice are derivatives of benzene (such as trinitrotoluene, dinitrophenol and toluene), yellow phosphorus and very large doses of DDT. An outbreak of jaundice in Epping, which occurred in 1965, was traced to wholemeal flour contaminated with 4,4-diaminodiphenylmethane, an aromatic amine, which had been spilt on the floor of the van in which the flour had been transported. All the affected patients recovered.

The development of angiosarcoma (a malignant tumour of the blood vessels) in men heavily exposed to vinyl chloride monomer, although very rare, is often fatal. The occurrence of several cases in one factory in the USA promptly alerted the occupational physician to the possibility that the cases were related to some exposure at work. Since the first report, about 50 cases have been reported, only a handful of which have come from Britain, and all in men exposed to very high concentrations of vinyl chloride monomer. It seems unlikely that new cases will arise in men with exposures to current low concentrations.

The kidney

The kidney is very vulnerable to damage from toxic materials because of its rich blood supply and because chemical compounds may be concentrated in its tissue when they are being excreted. Two groups of compounds encountered in occupational practice are most likely to produce kidney damage: heavy metals and organic solvents.

Heavy metals

In the cases of both mercury and cadmium, renal damage results in the appearance of protein in the urine from leaky renal tubules. In some patients with mercury poisoning, this may be so great as to interfere with normal fluid balance, producing the nephrotic syndrome, characterized by albuminuria, hypoalbuminaemia and oedema. This condition is fully reversible once exposure is discontinued. Cadmium poisoning leads to the appearance in the urine of a low-molecular-weight protein, β_2 -microglobulin.

Acute lead poisoning does not cause protein to appear in the urine, but can give rise to the abnormal excretion of glucose, phosphate and amino acids, a combination referred to as Fanconi syndrome. In lead poisoning, characteristic inclusion bodies are found in the nuclei of the cells of the proximal tubule. Chronic exposure to lead may induce structural deformation in the kidney, with impairment of the blood supply, which may result in hypertension. In the early part of the twentieth century, lead workers were noted to be unduly likely to die from cerebrovascular disease secondary to hypertension.

Solvents

Carbon tetrachloride is the most dangerous solvent affecting the kidney. Acute exposure can cause renal tubular death and complete cessation of kidney function. Treatment is by renal dialysis and recovery is usual once the acute phase is passed. Ethylene glycol, which is widely used in antifreeze solutions, is occasionally drunk by alcoholics when other forms of 'booze' are unavailable. Most of the ethylene glycol is metabolized to oxalic acid, which deposits in the renal tissue in the

form of calcium oxalate crystals that obstruct the renal tubules and lead to renal failure.

The cardiovascular system

Workplace hazards rarely have a direct effect on the cardiovascular system. However, certain organic solvents are thought to be capable of inducing cardiac arrhythmias, and vinyl chloride can cause peripheral arterial spasm akin to the clinical features of hand–arm vibration syndrome. Methylene chloride is metabolized in part to carbon monoxide, and carbon disulphide appears to have a direct atherogenic potential.

The blood

Only three chemicals have an important effect on the blood of people at work: lead, benzene and arsine. Each produces an anaemia, but by different mechanisms.

Lead

Lead is able to inhibit the activity of many of the enzymes in the body, especially those that contain active sulphhydryl groups. A number of such enzymes are required for the synthesis of haem, which, combined with the protein globin, forms haemoglobin, the pigmented complex that transports oxygen around the body in the red blood cells. Exposure must be unduly great or prolonged to produce a serious degree of anaemia. It should be remembered that the main source of lead exposure in humans is by ingestion rather than inhalation.

Benzene

Benzene, apart from being a genotoxic carcinogen, is unique among industrial poisons in being able to

depress bone marrow function. There may be several months or years between the onset (or cessation) of exposure and the development of anaemia. The severity of the anaemia is variable but in the worst cases, cell function in the marrow is completely destroyed. This so-called aplastic anaemia has a poor prognosis, although prolonged survival can be achieved with regular blood transfusions. Benzene can also lead to leukaemia.

Arsine

Arsine (AsH_3) is a hydride gas of arsenic. Exposure results in a haemolytic anaemia, perhaps related to an effect on red cell membrane sulphhydryl groups. One consequence of the haemolysis of red cells is that the renal tubules become blocked with destroyed red cells, which is fatal without artificial support. In the most severe cases, the patient may have no intact red cells, oxygen being transported in solution in the plasma. Prompt transfusion and renal dialysis are usually effective in producing a recovery. Hydrides of phosphorus (phosphine) and antimony (stibine) can produce haemolysis of red cells similarly to arsine.

Further reading

- Dodgson, J., McCallum, R.I., Bailey, M.R. and Fisher, D.R. (1988). *Inhaled Particles VI*. Pergamon Press, Oxford.
- Harrington, J.M. and Gill, F.S. (1998). *Occupational Health*, 4th edn. Blackwell Scientific Publications, Oxford.
- Hendrick, D.J., Burge, P.S., Beckett, W.S. and Churg, A. (2002). *Occupational Disorders of the Lung*. W.H. Saunders, London.
- Levy, B.S. and Wegman, D.H. (1995). *Occupational Health*, 3rd edn. Little, Brown & Company, Boston.
- Marple, V.A. and Liu, B.Y.H. (1983). *Aerosols in the Mining and Industrial Work Environment*. Ann Arbor Sciences, Ann Arbor, MI.

Chapter 7

The effects of some physical agents

Philip Raffaelli

Noise

- The mechanism of hearing
- Auditory effects of excessive noise
- Non-auditory effects of excessive noise

Temperature

- Effects of increased environmental temperatures
- Heat disorders
 - Heat cramps
 - Heat exhaustion
 - Heatstroke
- Effects of decreased environmental temperatures

Non-freezing cold injuries

- Freezing cold injuries
- Hypothermia

Pressure

- Barotrauma
- Working under increased pressure
- Gas toxicity
- Decompression illness
- Working at altitude
- Flying at altitude
- Further reading

Physical agents are sources of energy that may interact with the body, resulting in a transfer of energy. Examples include light, noise, radiation, vibration, and extremes in temperature and pressure. Exposure to excessive amounts of these physical agents can result in injury or disease and may occur in a wide range of occupations. In the UK, a number of the injuries or diseases that may result from such exposures in certain occupations are *prescribed diseases* and are reportable under law.

The assessment of most of these environmental hazards in the workplace is considered in Part 4. This chapter considers the health effects of exposure to excessive noise and to extremes in temperature and pressure.

Noise, temperature and pressure are features of the normal environment, and cause harm by being in excess of that usually experienced. Under normal circumstances, homeostatic mechanisms maintain the body's internal environment constant within a very wide range of external environmental conditions. Failure of these mechanisms leads to decompensation and injury or disease may result. If the agent is a mediator of a physical sense, the sensory organ is often the area affected. If the primary action of an agent is on a cell surface receptor, the target organ mechanism is important.

Noise

Noise is commonly considered to be unwanted sound. Sound is produced by vibrating objects creating pressure changes that travel as waves in the air or other media. The ear is the body's sensory organ and is the main target for noise damage.

The mechanism of hearing

The first part of the hearing mechanism is mechanical. The outer ear collects sound and directs it to the tympanic membrane, causing it to vibrate. This mechanical movement is then transmitted by three small bones (the auditory ossicles) to a smaller membrane leading to the inner ear. This mechanical part of the hearing mechanism can amplify the sound by up to 20 times. It can also afford some degree of protection via the acoustic reflex; when a loud noise occurs, muscles behind the eardrum contract automatically, suppressing the noise to enhance perception of sound and prevent injury. Unfortunately, its efficiency is limited as it does not occur quickly enough to cope with sudden very loud noise and it is subject to fatigue so cannot protect against sustained noise.

The second part of the hearing mechanism takes place in the inner ear and is sensorineural in nature. The inner ear is a membranous fluid-filled structure lying within the bone of the skull, consisting of the organ of hearing, the snail-shaped cochlea, and the organ of balance, the vestibular labyrinth or semicircular canals. The sensory transducer of the cochlea is the organ of Corti, which lies on the basilar membrane of the cochlea. Within the organ of Corti lies the sensory apparatus, consisting of a highly specialized layer of thousands of 'hair cells'. The ionic composition of the fluids in the various membranous channels of the inner ear is such that depolarization occurs when the hair cells are displaced by the sound wave that travels along the basilar membrane. The resulting electrical impulse is transmitted to the brain via the auditory nerve. Disorders affecting the first part of the hearing pathway before the cochlea cause conductive deafness, whereas disorders of the cochlea or auditory nerve result in sensorineural, or perceptive, deafness.

Auditory effects of excessive noise

High-level explosive noise impulses at around 35 kPa can cause direct trauma to the ear. Perforation of the eardrum and disruption of the auditory ossicles may occur, causing conductive deafness. A tear in the basilar membrane is also possible. This results in severe and usually total sensorineural deafness.

Lesser levels of acoustic trauma are more common in most occupational settings and can lead to classical noise-induced hearing loss. Two mechanisms are involved with these lesser degrees of acoustic trauma. Both lead to a reduction in hearing efficiency, which is measured by assessing the hearing threshold in decibels (dB) at different frequencies using an audiometer and recording the results as an audiogram. First, minor damage may cause swelling of the hair cells in what is thought to be a biochemical disturbance. The resulting hearing loss is initially reversible, a temporary threshold shift (TTS). The second is once again direct mechanical trauma of a lesser degree than that which will cause disruption of the basilar membrane. Over a period of time the hair cells

become damaged, losing their sensory structures and characteristic shape, resulting once again in sensorineural hearing loss. These changes are cumulative and lead to a permanent threshold shift (PTS).

The transitional level between these types of damage is not known. Indeed it may not exist because it is thought that damage may persist if repeated often enough, and the TTS may eventually become a PTS. Furthermore, some individuals are more susceptible to the effects of noise than others and noise-induced hearing loss occurs relatively randomly in exposed persons.

The changes in hearing induced by noise, however, are typical in those affected. The range of hearing extends from 20 to 20 000 Hz. Standard audiograms record the different hearing thresholds from 500 Hz to 6 kHz. Noise affects the higher frequencies first, usually at 4 kHz and extending between 2 and 6 kHz, but rarely occurs outside this range. This dip in hearing efficiency at 4 kHz seen on the audiogram is known as the audiometric notch. Although this change is typical of noise-induced hearing loss, it is not possible to distinguish it from sensorineural hearing loss owing to other causes such as exposure to certain drugs or chemicals. In these circumstances, it is important to have a good record of the occupational exposure to noise that the individual has experienced.

The speech frequencies lie within the range of 500 to 2000 Hz, and because they are not at first affected, noise-induced hearing loss may initially pass unnoticed. The fine pitch discrimination necessary for the comprehension of speech is a complex and incompletely understood process that requires interaction between inner and outer hair cells through neural feedback. As hair cells and neurones become depleted, the understanding of speech suffers because of this poor discrimination. Difficulty with conversation, especially in background noise, is the predominant symptom. In some cases, amplification can make the situation worse, as the remaining neurones suddenly become 'saturated' with sound in a phenomenon known as *recruitment*. The resulting distortion may be quite uncomfortable. The differential between achieving sufficient sound input for compre-

hension and the onset of recruitment may not be very great. As a result, it may be very difficult to communicate with someone who has recruitment. Additionally, hearing sensitivity often declines as people become older and this age-related hearing loss, or presbycusis, will add to the noise-induced hearing loss.

Noise-induced hearing loss may also be complicated by the phenomenon known as tinnitus, which is a constant ringing or buzzing in the ear, even in the absence of external noise. Tinnitus can not only interfere with hearing but is often distressing in its own right. There is no medical treatment for noise-induced hearing loss and, once present, the only option may be a hearing aid. However, if there are problems with speech discrimination and recruitment this may not help.

It is obviously important to guard against this insidious onset of disability, and noise control levels are set by most health and safety legislatures. Even so, it is worthwhile bearing in mind that the sensitivity of hearing is subject to individual variability, and that a minority of workers may suffer hearing loss at levels at or below the legal limits for noise exposure.

Non-auditory effects of excessive noise

There are also some non-auditory effects of noise. At lower levels, the effect may be little more than distraction, which, if sustained, may lead to irritation. In some situations, the individual's pulse, blood pressure and sweat rate may increase. In these circumstances, the noise is probably acting as a general stressor, when these physiological changes are mediated via the autonomic nervous system analogously to the 'flight-or-fight' mechanism.

Temperature

The normal thermoregulatory mechanisms are very efficient, and core body temperature is maintained at close to 37°C (98.4°F) with a diurnal variation of 0.5–1°C. This thermal homeostasis depends on the balance between metabolic heat production (M) and evaporation (E), convection (C), conduction (K), radiation (R) and storage (S):

$$M = E \pm C \pm K \pm R \pm S$$

Thus, this balance is affected by the activity of the worker, on one hand, and the ambient temperature and humidity, the presence of any radiant heat source, air movement and any clothing or personal protective equipment the worker may be wearing on the other. As the worker's metabolic rate increases, energy is produced in the form of heat. Blood vessels in the skin dilate to carry the heat from the core to the surface, where it is dissipated through convection, radiation and conduction – a process that is assisted by the evaporation of sweat. Discomfort due to the increased temperature gives a powerful stimulus to slow the rate of work. In cold climates, the opposite occurs. There is a stimulus to increase the metabolic rate, either through increased physical activity or shivering. Skin blood vessels constrict, helping to conserve core temperature, but causing cooling in the peripheries.

The most common method of assessing the thermal environment is the wet bulb globe temperature (WBGT) measured in degrees Celsius. When the WBGT reading is high, the work regime can be modified taking into account the work load. Adjustment can also be made for the clothing worn by the worker.

For the vast majority of workers, the temperature of their workplace has importance only in terms of comfort zones, and thermal comfort limits are intended to ensure productivity and quality of work, not protect health. Typically, the minimum recommended temperature is 16°C with an upper limit of 30°C; if the work is strenuous, these may be modified to 13°C and 27°C.

However, for an important minority of the workforce the temperatures encountered in the working environment may exceed the body's thermoregulatory capacity and have the potential to be detrimental to health or even to be life-threatening. Such conditions vary greatly depending on the job and where it is being done. It may be extremely hot for the foundry worker or very cold for cold-storage operatives. Oil workers in an Alaskan winter (e.g. –40°C) experience markedly different temperatures from their colleagues undertaking otherwise similar work in the Arabian desert in summer (perhaps +40°C).

Effects of increased environmental temperatures

When individuals are routinely exposed to heat, some adaptation will occur over a period of time. This includes physiological changes in the thermoregulatory system, such as the ability to sweat being optimized, but also behavioural changes such as increased fluid intake, reduced activity where allowed/possible and changes to clothing. There is good evidence that physically fit individuals are better able to adapt to excessive heat and to withstand heat stress. Unfit or obese people are less able to adapt well to hot temperatures. Adaptation may also be compromised by illness or by taking certain drugs, including alcohol.

In most circumstances, those most at risk of heat stress are those whose heat environment changes suddenly, for example the working of a new and deep seam in coal mining or the unexpected hot spell that overwhelms the ventilation system in a boiler room. Management should be aware of the possibility of such occurrences, and should ask for appropriate hygiene advice to determine if the working environment is acceptable. If work is performed under unsuitable conditions, particularly conditions of high humidity and low airflow, the core temperature of exposed workers may start to rise. The first signs of the resulting heat stress are discomfort and fatigue, and the natural reaction is to slow down the work rate or stop work, a protective mechanism that will allow cooling and recovery. In some situations, for example in rescue work and armed service, there may be powerful stimuli to exceed the limits of normal work capacity. Such workers should be trained to recognize this risk and taught how to measure the thermal environment and to enforce a work–rest regime if necessary.

All those involved with work in high temperatures should be aware of the physiological changes that take place and the symptoms that occur on continued and uncontrolled exposure to excessive heat when the body's thermoregulatory mechanisms begin to decompensate. This is particularly important; if the premonitory symptoms are ignored, serious illness may occur with alarming rapidity.

Heat disorders

Heat disorders may range from mild to fatal in severity and may come on after only a few hours in more extreme conditions, particularly when the worker is undertaking strenuous activity but is unable to effectively dissipate heat due to blocked evaporation. They may also develop after some days of prolonged exposure to heat when the excessive sweating has not been made up by sufficient fluid intake leading to dehydration, sodium and potassium salt depletion, and hypovolaemia. Although the different heat disorders are of increasing severity, an affected individual may present with any of the conditions without necessarily having developed one of the less serious disorders.

Heat cramps

Sweating leads to the loss of large amounts of fluid and salt, and the failure to intake adequate replacement, especially in the non-acclimatized worker, can result in painful, incapacitating and involuntary cramps affecting the leg and abdominal muscles. The core temperature is usually within the normal range. Heat cramps are not harmful in themselves and can be effectively treated by rest and adequate fluid replacement. There is no benefit in most circumstances in adding salt to the fluid and, indeed, in some circumstances, this may be harmful.

Heat exhaustion

In heat exhaustion, the initial mechanism is similar with the loss of fluid and salt due to sweating. However, in this case the circulatory system fails to keep up with the extra demands placed upon it because the inadequate fluid replacement causes a fall in circulating blood volume. The body may respond by trying to increase heat loss through peripheral vasodilatation, leading to collapse due to hypovolaemic shock. The patient will look and feel ill with a pale, clammy skin and extreme lethargy. Mental confusion may develop followed by unconsciousness. The core temperature ranges from 38.3°C to 40.6°C (101°F to 105°F). Treatment is by rest, controlled cooling and adequate fluid replacement.

Heatstroke

The seminal event in heatstroke is an impending failure of the sweating mechanism, resulting in decompensation of the thermoregulatory mechanisms and hyperpyrexia. This involves a rapid rise in core temperature to 40°C to 41°C (104°F to 106°F); a temperature above 41°C is a poor prognostic sign. The patient looks hot and flushed and the skin is usually dry. There may have been signs of heat cramps or heat exhaustion accompanied by headache, vertigo and fatigue, but this is not obligatory. The pulse rate is rapid but blood pressure is seldom affected. Mental confusion may develop followed by unconsciousness or convulsions and death. Urgent medical assistance is essential and will involve rapid cooling and treatment by intravenous fluids and drugs as appropriate. In an emergency, first aid by immersion in a cold bath and fluid replacement may be life-saving, although cold water immersion has its inherent dangers as it may lead to peripheral vasoconstriction, further reducing the body's ability to lose heat. Even if the patient recovers, there may be residual damage to the brain or the kidneys.

Effects of decreased environmental temperatures

There are two main types of climate in which cold injuries may occur. In a cold dry climate, snow and ice are usually present, and the temperature seldom rises above 0°C (32°F). A cold, wet climate is more typical of winter in 'temperate' zones, when the temperature may vary from 10°C to 12°C.

In cold climates, the body's principal physiological response is to try to maintain temperature by decreasing the peripheral circulation by vasoconstriction to reduce heat loss. Pre-existent circulatory disorders may therefore be exacerbated by the cold. If the vasoconstriction is prolonged, it may cause functional disturbances or damage to small blood vessels, nerves and the skin. Other responses to cold include increasing the metabolic rate through an impulse to maintain physical activity or, when this is not possible, by shivering.

Cold injuries due to occupation are not common. Toes, fingers, ears and the nose are the most com-

mon sites for cold injury as they lose heat more rapidly due to their higher surface area-volume ratio and the peripheral vasoconstriction. They are also more likely to be in contact with colder surfaces than other parts of the body. Most workers exposed to extreme cold are well aware of the risks and wear suitable protective clothing and are careful about handling metal tools or similar objects. The risk of cold injury is increased in the presence of certain diseases and if there is co-existing dehydration or hypoxia. The risk is also increased by taking certain drugs, including alcohol.

Non-freezing cold injuries

Non-freezing cold injuries occur after prolonged exposure to cold and often damp conditions. Chilblains are a form of mild cold injury following repeated exposures to low temperatures [0–16°C (32–60°F)]. They are characterized by redness and swelling of the skin in the affected area and may be associated with tingling and pain.

Trench foot and immersion foot occur in individuals whose feet have been wet and cold for days or weeks, particularly when associated with the pressure of standing in a relatively immobile position. There is no sharp demarcation between the two conditions. Trench foot tends to occur with shorter exposures to colder conditions, whereas immersion foot is associated with longer exposures to slightly higher temperatures. Despite the epithet, the hands may be affected with a similar condition if cold, wet gloves are worn for prolonged periods. The area involved may initially be pale due to shutdown of the peripheral circulation but then turns red progressing to a purple-blue. The skin becomes macerated and swollen and there may be blisters. The circulatory changes may lead to tissue loss and nerve damage, which may cause permanently altered sensation in the affected part. Symptoms include tingling, numbness and pain.

Freezing cold injuries

Freezing cold injuries are most likely in a cold dry climate in which the extremities are exposed to freezing air. Short exposures may lead to rapid surface freezing that produces a white spot on the

skin known as frostnip. At this stage, the underlying tissues are still viable and early recognition and rewarming limits damage.

Frostbite occurs when the exposure to freezing air is prolonged or the cold is extreme. It can also result from contact with frozen metals or from short exposures in the workplace to cooled or compressed gases. Freezing occurs in the deeper tissues with ice crystal formation in tissue fluid, resulting in structural damage to cells. The skin is pale and solid, looking and literally feeling frozen. The damage is made worse by similar changes in blood vessels. On rewarming, the damaged vessels leak, producing areas of inflammation and swelling that further compromise the circulation and add to the tissue destruction. Blisters may form. Depending on the extent of the damage, the injury may be painful but, in severe cases, the damage to nerves may be such that the injury is pain free. Some degree of tissue loss is inevitable and may be more extensive than the initial injury suggests owing to subsequent infection and gangrene. Treatment is aimed at minimizing this subsequent damage.

Hypothermia

The body's thermoregulatory mechanisms can maintain the core temperature to within 1–2°C of normal in all but the most severely cold environments, as long as dry, warm clothing is worn and physical activity performed. Problems arise if the clothing is inadequate or wet or if physical activity cannot be maintained due to exhaustion. Immersion in water is the most common cause of hypothermia, but cold, wet and windy conditions are also particularly treacherous. Once the body's thermoregulatory mechanisms are overwhelmed, the body's core temperature starts to fall. As the body temperature falls to below 34°C, the initial discomfort associated with the sensation of cold reduces and the shivering reflex diminishes. There is increasing fatigue, apathy and disorientation. Individuals at this stage may be unaware of these changes in themselves so it is important that those who work outdoors in winter should be looking out for these symptoms. As the body temperature falls below 33°C consciousness diminishes and cardiac arrhythmias may occur as the temperature

falls to below 30°C. Coma and cardiac arrest occurs as the temperature reaches 25°C and below, followed by death. Treatment involves rewarming, and, if the exposure has been of short duration and the hypothermia is mild, then this may be all that is required. However, if the hypothermia has been of longer duration or is profound then treatment is more difficult as the body is lacking in oxygen and there may be extensive electrolyte changes during the rewarming phase, with the result that cardiac arrest may occur. Slow controlled rewarming under medical supervision is essential.

Pressure

Normal atmospheric pressure at sea level is 10^5 Pa (in other units of pressure, this equates to 1 atmosphere (atm), 1 bar, 760 mmHg or 14.7 lb in^{-2}). Humans are adapted to live at atmospheric pressure and can acclimatize over a week or two to living at altitude. Acclimatization involves a number of physiological changes to improve the body's ability to utilize oxygen, including increasing the amount of red blood cells and haemoglobin. However, the body has difficulty in adapting to the increased pressures in diving or caisson work, or the decreased pressures in aviation, and may also develop problems due to the rapid changes of pressure that may occur in these environments. Most of the adverse health effects encountered in pressure work are due to decompression.

Barometric pressure on land results from the weight of the air above. At depth, the weight of the water above must be added to this. At 10 m (33 ft) deep in seawater, the diver is exposed to a pressure of 1 atm higher than the barometric pressure at the surface. The total pressure is thus 2 atm absolute. Every additional 10 m of descent adds 1 atm. The pressure in a caisson or tunnel (in which compressed air is used to exclude water from the work site) reflects the pressure of the water outside.

Barotrauma

The volume of a gas is inversely proportional to the absolute pressure (Boyle's law). The resulting

volume changes during compression and decompression are not a problem as long as the air-containing spaces within the body can freely equalize with the ambient pressure. If the air is trapped in the body, then a relative vacuum occurs on descent and excess pressure on ascent. The effect on the body depends on the site affected and the size and rate of the pressure change. The upper respiratory tract and lungs are commonly affected. Trapping of gas in the sinuses causes pain and mucosal trauma, and if the ears cannot be cleared because of a blocked Eustachian tube then there is painful mechanical stress on the tympanic membrane, which may bruise or perforate. If the lungs are affected, rupture of the alveoli and collapse of the lung (pneumothorax) may occur, leading to chest pain and respiratory difficulty. If the air from the ruptured lung enters the bloodstream, this may then lodge in the cerebral blood vessels, preventing blood flow to the affected area. This is known as a *cerebral arterial gas embolism*, which usually causes rapid loss of consciousness and is often fatal. Even if treated early by recompression, there may be residual central nervous system (CNS) effects.

Working under increased pressure

During work under pressure, the gases used in the breathing mixture come to equilibrium in the tissues. The amount of gas dissolved depends on its partial pressure and solubility, and depends on the length of time worked (Henry's law). This can lead to adverse effects either because the gases have toxic properties at pressure, or because of problems when the dissolved gas escapes from the tissues through the bloodstream and lungs during decompression.

Gas toxicity

The partial pressure of a gas is related to its concentration and the absolute pressure (Dalton's law). The essential constituent of air is oxygen. However, extended exposure to normal concentrations of oxygen at depths of below 15 m can result in pulmonary oxygen toxicity, which can range from a cough to respiratory distress. CNS oxygen toxicity can occur below 50 m and may cause

convulsions and unconsciousness. Oxygen toxicity is usually avoided by controlling the length of the diving operation and by diluting the oxygen with an inert gas. As nitrogen at depth has a mild anaesthetic effect similar to alcohol (N_2 narcosis), helium is usually used for deep diving.

Decompression illness

If the rate of decompression is too rapid, the gases dissolved in the tissues become supersaturated and bubbles are formed in tissues or blood vessels. This usually occurs because of repeated exposures to pressure without sufficient time between dives or if there is too rapid a return to the surface. The syndrome that results is complex and depends on the site of the bubble formation. Effects may range from itchy skin and joint pains to breathing difficulties and chest pain, and may include involvement of the nervous system, including paralysis, convulsions and death. Decompression illness (DCI) usually appears a few hours after surfacing, but may be sooner and is unpredictable in its course. Seemingly minor symptoms may be followed by florid neurological signs. There may also be long-term effects such as bone death (dysbaric osteonecrosis). Treatment of acute DCI is recompression but the degree of recovery is dependent on the severity of the injury and the delay before treatment.

Working at altitude

The partial pressure of oxygen at 5500 m is about one-half of that at sea level. Most people can acclimatize up to 3000 m but, before then, a worker will be easily fatigued and prone to breathlessness. About 20% of persons ascending above 2500 m will develop altitude sickness. This percentage increases with the rate of ascent and the altitude. Symptoms include fatigue, headache and visual disturbance. There may be fluid swelling of tissues (oedema), and if this involves the brain, there may be CNS effects. Treatment is by oxygen and descent. Drug treatment is controversial. Wherever possible, an otherwise fit worker should be allowed to acclimatize naturally before being expected to perform at full capacity.

Flying at altitude

Flying at altitude can result in a relatively rapid reduction in the partial pressure of oxygen. Above 3000 m this can result in mental changes including reduced reaction time and decreased ability to concentrate. A combination of hypoxia and anxiety may also lead to hyperventilation, which is ineffective in improving oxygen levels but does reduce the carbon dioxide level in the blood, leading to a respiratory alkalosis which may cause light-headedness, flushing, tingling in the limbs and tunnel vision.

At altitudes above 8000 m, decompression symptoms may occur, but this is avoided in commercial aircraft by maintaining the cabin at a pressure equivalent to around 2500 m. If cabin depressurization occurs in an emergency, barotrauma and DCI may result, along with the effects

of hypoxia and extreme cold (-55°C at 10 000 m).

Mild barotrauma may occur during descents from altitude if there is difficulty in clearing the ears, but there are few other problems due to pressure during a normal descent.

Further reading

- Edmunds, C., Lowry, C., Pennefather, J. and Walker, R. (2002). *Diving and Subaquatic Medicine*, 4th edn. Arnold, London.
- Ernsting, J., Nicholson, A.N. and Rainford, D. J. (1999). *Aviation Medicine*, 2nd edn. Butterworth-Heinemann, Oxford.
- Harrington, J.M., Gill, F.S., Aw, T.C. and Gardiner, K. (1998). *Pocket Consultant: Occupational Health*, 4th edn. Blackwell Science, Oxford.
- Waldron, H.A. (ed.) (1997) *Occupational Health Practice*, 4th edn. Butterworth-Heinemann, Oxford.

Chapter 8

Toxicology

Julian Delic, Steven Fairhurst and Maureen Meldrum

What is toxicology?

How do substances exert toxicity?

Acute toxicity

Irritation and sensitization

Repeated-dose toxicity

Genotoxicity

Carcinogenicity

Reproductive toxicity

Where do we get toxicological information from?

Physicochemical properties

In vitro systems

In vivo systems in experimental animals

Human experience

What key pieces of toxicological information emerge?

What do we use the information for?

Hazard identification

Risk assessment

Standard-setting

Hazard banding

Regulatory framework for industrial chemicals

Some themes for the future?

References

What is toxicology?

Toxicology is the study of the potential of any substance to produce adverse health effects as a result of its physical or chemical properties (the *hazards* of the substance), and the likelihood that such adverse properties might be expressed under specified exposure conditions (the *risk* of toxicity inherent in a particular set of circumstances).

The target species of interest is the human, in the occupational context ‘humans at work’. For those of us concerned with trying to secure protection of workers from any adverse health consequences of their work, toxicology should be able to provide us with two things that are useful prospectively:

1 knowledge of the threats posed by a chemical; and

2 judgement about the level of control of exposure necessary to avoid such threats.

Put another way, for those concerned with workplace control, toxicology should provide answers to the questions: ‘Control of what, to what extent, and why?’

Toxicology can also provide interpretational answers to questions posed by observations already made. For example, when an individual or

population is identified as suffering from ill health, toxicology can answer questions about causation. If a group of workers, known to be exposed to a substance, shows an excessive prevalence of a health problem, toxicological science should be used to examine whether or not it is reasonable to attribute the health effect to the substance.

How do substances exert toxicity?

The different ways in which toxicity can be expressed are all covered by the term *toxicodynamics* (‘what the substance can do to the body’). However, before a substance can exert any toxic properties, it must come into contact with and, in many cases, enter the body. The term *toxicokinetics* covers the way in which the body handles a substance (‘what the body does with the substance’). Toxicokinetics includes the ways in which substances are absorbed into, distributed around, metabolized by (often an important process in relation to the toxicodynamic properties of a substance) and excreted (removed) from the body. Understanding the toxicokinetics of a substance can give important predictive clues or reflective

explanations about the toxicodynamics of a substance. More detailed information on toxicokinetics can be found in the general toxicology references cited at the end of the chapter.

The way a substance exerts its toxicodynamic properties will be dependent upon a number of factors including its physicochemical properties, how the body is exposed to it and its toxicokinetics once inside the body. At one end of the scale, very simple physicochemical properties can lead to the expression of toxicity. For example, extremes of pH damage tissues, and lipophilic solvents can extract fats from the skin leading to irritation and, if left untreated, more severe skin damage. Even seemingly stable unreactive material, such as particulates (dusts, fibres) can cause problems when, for example, they get into the lungs. Because the lining of the lungs is exposed directly to the atmosphere we breathe, a sophisticated defence system has evolved in order to protect this vital organ from attack (for example by microorganisms). When persistent particulates land on the surface of the airways and lungs, if they are resistant to being broken down then the normal clearance mechanisms designed to attack, destroy and clear up unwanted material can overreact, inadvertently leading to damage to the lungs.

Many chemicals are valuable to industry because they are reactive. This same reactivity, although of value in an industrial process, means that such chemicals may also have the potential to react with other molecules, including those found in the body. This reactivity may be expressed towards particular chemical groupings in biological molecules or be of a more non-specific type, but can ultimately result in damage to and dysfunction of key components of biological systems.

For some substances the body 'sees' them (or responds to them) in a manner similar or related to the body systems' interactions with physiological substances. For example, some substances act on the body in a manner similar to that of natural hormones – they can stimulate, or block, natural hormonal-type activities, whereas other substances may block enzyme activity for similar reasons.

One of the ways that the toxicokinetics of a substance can influence its toxicodynamics is through the way in which it can be altered chem-

ically by the body through metabolism. Humans have evolved in environments that continually present a challenge to survival. The food that we eat, the water we drink and the air we breathe have always carried potential problems in the form of poisonous materials. We have evolved to cope with this threat by developing enzyme systems (mainly in the liver but present in many other organs) that can metabolize substances with the aim of making them more water soluble, thus aiding their excretion in the urine. However, paradoxically, the body sometimes creates its own problems in its drive to convert the substance into such a form, with the metabolizing systems transforming the substance from a relatively innocuous material into something much more reactive and potentially dangerous.

Overall, given that biological systems represent a balanced and very complex organization of a myriad of (bio)chemical structures and chemical reactions, every substance thrown into such a system will have the capability of exerting some form of disturbance (toxicity). Hence every substance has some 'toxicity'. However, as outlined above, the toxicity of different substances is hugely variable – a feature of the physicochemistry of the substance, the amount encountered (dose), the frequency of exposure (once only, occasionally, daily, continuously), and the route of exposure (on to the skin, into the lungs).

Clearly, with such a range of potential ways in which a chemical can act on the body, there are different forms of expression of toxicity. Conventionally, these different manifestations are termed 'toxicological end-points' and are grouped as follows:

- acute toxicity;
- irritation;
- sensitization;
- repeated-dose toxicity;
- genotoxicity;
- carcinogenicity;
- reproductive toxicity.

Acute toxicity

Acute toxicity is that which is induced following a single dose or exposure of a substance. People

often associate this with severe effects such as death (see discussion later in the chapter on the use of LD₅₀) but in fact more subtle effects such as those induced by solvents on the central nervous system (CNS depression) are also covered. In many ways, these effects can be more important in the occupational context as they generally are likely to occur at much lower levels of exposure than those that induce death. It is often the case that information is gathered on acute toxicity following oral dosing but, for the workplace, the inhalation and dermal routes are of greater importance.

Irritation and sensitization

Irritation and sensitization are terms used to cover the local effects that generally are associated with the skin and eye, although the respiratory tract is also an important site. As indicated earlier, skin, eye and respiratory tract irritation is often induced by the simple physicochemical properties of a substance, whereas sensitization is a more complex process resulting from the action of a person's immune system on a chemical. Sensitization can also involve the respiratory tract of the individual leading to asthmatic responses (although there may be other mechanisms for the induction of asthma by chemicals that do not necessarily involve the immune system).

Repeated-dose toxicity

Although it is clearly important to know about these short-term effects, workers are often likely to be exposed to chemicals over longer time periods on a daily basis. Thus, it is important to obtain knowledge about the effects of repeated exposure to chemicals, preferably via the route by which exposure is most likely to occur (most often inhalation and dermally in the workplace). Repeated exposure studies can occur over a range of overall timescales from short (14–28 days, so-called 'subacute' studies) to long-term or 'chronic' (2 years or more, depending on the species studied) durations. The overall length of the study can be important as in some cases it may take considerable time for the expression of some forms of

toxicity (for example, because it is the gradual build-up of damage rather than a sudden event than may be important).

Genotoxicity

Genotoxicity is the general term used to describe damage to the genetic material (genome) of a cell or cells by a substance. This is important as it can lead to a mutation that is a permanent change to genetic material. If occurring in a critical part of the genome of a cell (for example in the part which controls cell division), such mutations can have serious consequences such as ultimately the induction of cancer or, if in the germ cells, to a heritable mutation that can impair development or inhibit formation of offspring. As for other forms of toxicity, genotoxicity can be induced either directly (e.g. by reactive chemicals) or via metabolism to a reactive form. In the occupational context, reactive chemicals are of particular concern as they can act at a site of contact (respiratory tract, skin, eye) with the body, a situation that can often arise in the workplace. Thus, study of this aspect of the toxicology of a substance is often critical because of the serious consequences.

Carcinogenicity

Carcinogenicity is the ability of a substance to induce cancer. As indicated, this may be due ultimately to direct or indirect action of a chemical on the genome – so-called *genotoxic carcinogens*. However, in some cases a chemical may have no direct or indirect action on genetic material but may induce through various means long-term division of cells. This is thought to increase the chances of normal background mutations, which arise naturally (e.g. through exposure to background ionizing radiation) of contributing to the process and resulting ultimately in the formation of cancers – so-called *non-genotoxic carcinogens*. The term threshold is used to denote whether or not it is possible to identify a level of exposure below which this effect is unlikely to occur (this issue is discussed further below).

Reproductive toxicity

As suggested by the term, reproductive toxicity covers adverse effects on the ability to reproduce. The term encompasses a very broad range of effects and includes everything from effects on fertility (i.e. on the germ cells and/or reproductive tract/organs) through to effects on the developing offspring both before and after birth; the term would even cover effects on behaviour, which may influence the ability to conceive. In terms of effects on the developing offspring, the term teratogenicity is often used. Most toxicologists would consider this to relate specifically to those effects on development that result in actual malformations (e.g. reduced limb development, cleft palate). The term developmental toxicity is broader and covers teratogenicity as well as other aspects such as effects on body weight and neurological development (e.g. effects on IQ).

Although providing a convenient way of considering the toxicodynamics of a substance, this structural arrangement of end-points is important in other ways. It provides a framework that is followed in the development of toxicity tests, most of which are designed to focus mainly (sometimes exclusively) on one of these end-points. In consequence, the regulatory systems for chemicals that exist in the UK, EU, USA and most other parts of the world are designed to receive and respond to toxicity information acquired in accordance with this framework.

Where do we get toxicological information from?

If the role of toxicology in the occupational context is to develop a knowledge of the hazards and risks to workers from exposure to substances used in the workplace in order to ensure that they are controlled appropriately, then logically the best information would be derived from studies in humans, the target population of interest. Although we sometimes have this type of information, observations of health consequences in workforces exposed to chemicals during their work have

been made and recorded in a very patchy manner. Sometimes there are also difficult confounding factors involved in interpreting such data. In terms of new studies in humans, in general it would be unethical to deliberately generate data in ways that might harm human health.

Thus, a number of approaches have been developed for the generation of data to assess the toxic properties of substances. These range from predictions based on the physicochemical properties of a substance to *in vitro* and *in vivo* test systems through to information based on human experience. The following provides a brief overview of these approaches and their relative contribution to the formation of a toxicology picture of a substance.

Physicochemical properties

A basic consideration of the physicochemical properties of a substance can help to inform on its potential toxicokinetics and toxicodynamics. For example, substances that are relatively small (molecular weight of between 100 and 500), water soluble (between 1 and 10 000 mg l⁻¹) and partition reasonably well into fat (log *P*-values between 1 and 4) are likely to be taken up well across the skin. Similarly, dusts with aerodynamic diameters below 10 µm have the potential to be inhaled and retained in the lungs.

A simple measurement of pH may provide insight into the potential for the induction of local effects; if the pH is excessively acid or alkali then it would be anticipated that the substance could cause damage at a site of contact (e.g. the skin or eyes). These are relatively simple applications of very basic and generally easily obtained information, but they can provide a useful starting point in the consideration of the likely toxic properties of a substance and influence the way further data are gathered.

More sophisticated approaches have been explored, particularly since the advent of readily available computer systems (the term *in silico* has been used in the context of the application of computer-based technologies to toxicological assessment). These approaches depend upon knowledge of the molecular structure of a substance and relationships between this and potential toxic

properties – so-called structure–activity relationships (SAR). There are a number of SAR approaches that have been developed (Barratt, 2000). Some essentially recognize parts of a molecule (a structural alert) that have previously been shown to exert a toxic effect. If these are found on a molecule presented as new to the system then it may ‘predict’ that the substance could be capable of exerting the same effect. A simple example might be the presence of an isocyanate ($-NCO$) functional group being a structural alert for the potential to cause asthma.

Other systems employ more complex procedures such as statistical analyses relating for example the stereochemistry, hydrophobicity and/or electronic properties of substances with their toxicological properties. These approaches are more quantitative in nature and thus are called quantitative SAR (QSAR). However, the degree of understanding of such correlations required in order to use these QSAR approaches with confidence is such that currently they can only be applied legitimately in relatively few, tightly defined circumstances. Overall, although QSAR systems can be useful for screening chemicals for potential toxicity (at least in a qualitative manner), they are not yet at a stage where they enjoy complete confidence in their predictive abilities and do not provide reliable quantitative outputs. However, their role is likely to become of increasing importance with the advent of regulatory changes for industrial chemicals in the EU in the near future (see below).

A further application of computer-based techniques is in the use of physiologically based pharmacokinetic modelling (PBPK). The principle of this technique is to mathematically model the toxicokinetics of a substance by using information on its physicochemical parameters (e.g. partition coefficients) and knowledge of the anatomy and physiology of the human and experimental animal (Andersen and Krishnan, 1994). It is important to understand that the technique is not a simple mathematical exercise of ‘curve-fitting’ but attempts to model the biology of the system of interest and the way a substance would pass through it. An advantage of PBPK modelling is that it allows the exploration of how a substance might be han-

dled over a whole series of exposure conditions for which data do not exist. For example, information from oral dosing studies may be used to model how an inhalation uptake might be handled. Similarly, data from an experimental animal species could be used to predict the kinetics of a substance in humans (Delic *et al.*, 2000). A difficulty with PBPK modelling is that the models require suitable validation in order to be confident of the predicted outputs; the acquisition of such human validity data can be difficult to achieve. Nevertheless, this approach promises much for the future and efforts are already under way to model toxicodynamics aspects in order to provide complete physiologically based biological response models.

***In vitro* systems**

The next step on from a simple consideration of chemical structure/properties and the application of computer technology is the generation of data in living systems that lie outside the body, so-called *in vitro* approaches (Eisenbrand *et al.*, 2002). In these methods, living cells or tissues are maintained in the laboratory under conditions that ideally mimic their natural environment within the body and the effects of the application of substances to them are studied. In some cases, microorganisms, such as bacteria and fungi, are used as simple screening systems. There has been much work in this area, particularly with a view to reducing the use of experimental animals. The most notable success has been in the development of *in vitro* methods to explore the potential for substances to damage genetic material (genotoxicity). There is now an internationally recognized battery of *in vitro* tests that assess genotoxic potential in a stepwise manner in bacterial cells (Ames test) through to mammalian cells that are maintained in laboratory culture (Committee on Mutagenicity, 2000). For the occupational setting, this battery of tests, depending upon the outcome and use pattern of the substance of interest, may provide all the information that is required to judge the potential for a substance to damage genetic material, although, as explained later, further testing *in vivo* may also be required in some cases.

In vitro approaches to address satisfactorily other toxicological end-points have proven to be more difficult to develop, particularly where complex tissue and organ systems exist within the body, though recently useful *in vitro* tests for dermal uptake and corrosive potential have been developed. The development of ‘-nomic’ (see below) technologies is likely to provide a major boost to the use of *in vitro* techniques.

***In vivo* systems in experimental animals**

Most of the data available to the toxicologist is generated in studies using experimental animals. There is much debate about the ethics of using animals in such studies but ultimately despite the continuing development of alternative approaches as outlined above, the ability to reproduce the complex conditions that exist within the tissues and organs within a living organism and the interactions (and their perturbation by substances that enter the body) between these are currently not possible to reproduce *in vitro*. Thus, if we want to understand if and how a substance may induce toxic effects then it remains the case that for many substances this needs to be done within the complex systems of the living organism. This requirement is set against the ethical considerations of using living animals and thus the situation is reached whereby the experimental studies are designed to provide the maximum amount of useful information possible while using the minimum number of animals and keeping their suffering to a minimum. To this end, the three ‘Rs’ of animal welfare have been developed – replacement (where possible do not use animals, use an alternative), reduction (reduce the numbers if replacement is not possible) and refinement (refine techniques to reduce suffering) (Russell and Burch, 1959).

As it is important to limit the number of animals used in testing and in order to ensure that robust standards are met around the world (so that data generated in one place may be used with confidence elsewhere), guidelines have been developed to set out test methodologies to follow. These have been developed at the international level by the Organization for Economic Co-operation and

Development (OECD) and published as a set of agreed test guidelines (OECD, 1993). These guidelines can then be used by anyone as a benchmark against which to conduct studies to investigate the toxicity of a substance. (It should be noted that the OECD guidelines also include methods for *in vitro* techniques and are not exclusive to animal tests.) For regulatory purposes (e.g. when submitting a package to a regulatory authority in the EU) it is necessary that the laboratory performing the experiments can demonstrate that they have been conducted in an appropriate scientific manner, using the methods laid out in their protocols. In order to ensure this, laboratories need to demonstrate that they are operating to the principles of good laboratory practice (GLP), including the use of quality assurance procedures. GLP, in particular, is important to ensure integrity of the procedures used and the results reported.

The range of OECD test methods available to the investigator covers the full range of toxic properties as outlined earlier in this chapter from toxicokinetics through to the exploration for carcinogenic and reproductive toxicity potential. However, as well as these standardized test methods, many studies are carried out which are designed as scientific investigations using techniques that lie outside these standard methodologies but which can contribute further to our understanding of the hazards of occupationally relevant chemicals. As in any other area of science, these studies are published in the relevant peer-reviewed scientific journals and often provide further information over and above that from the standard test methods, particularly in relation to the way in which a substance at the tissue or molecular level expresses its toxicity.

Obviously, all experimental animal work generates data that reflect occurrences in a species other than the one of interest, i.e. the human. There are many similarities between the biology of commonly used experimental animals and humans, such that much of this experimental data is informative and useful. However, there are also important differences that appear between species; these can be apparent as general features or may be specific to the way in which a particular substance is handled by the organism. The allowance

for or accommodation of (potential) differences between experimental animals and humans in the way they respond to a chemical exposure sometimes becomes an area of great controversy and conflict in regulatory toxicology.

Human experience

As stated above, in general it would be unethical to deliberately expose humans in an experimental situation to a substance in order to assess potential toxicity. Nevertheless, in some instances data are available from direct experimentation in volunteers (for example, the effects of substances on enzyme activity such as cholinesterase or the determination of sensory irritation responses on exposure to airborne substances), when the expected responses (in their nature and/or scale) are not anticipated to lead to ill health.

Data in humans may also be available from reports of cases of individuals exposed to substances often in accidental situations, so-called 'case reports'. Generally, these tend to describe people exposed to a substance acutely or over a relatively short space of time, leading to the development of ill health. Although these reports may provide useful indicators of potential toxicity, particularly when grouped together, they rarely contain reliable quantitative information on exposure and may be confounded by unexplored factors such as the original health status of the individual and potential exposure to other substances and/or agents.

For some substances (numerically, very much the minority, but including some individually very important industrial chemicals), the most valuable data derive from epidemiological studies. These can be cross-sectional, cohort or case-control in design. Cross-sectional studies provide a 'snapshot' in time; they inform on the prevalence of a particular health effect in relation to contemporary levels of exposure. Cross-sectional studies can provide important clues concerning the potential health hazards of an industrial chemical but, in general, they cannot reliably inform on exposure-response relationships. This is partly because workers who are more susceptible to the health effect in question may have left the industry prior

to the study, leaving behind a non-representative 'survivor' population. This can bias the results, causing an underestimate of the true health impact of the exposure. Also, if the health effect in question is a slowly developing chronic condition such as silicosis then the precise point in time at which the health effect first developed, and the cumulative exposures that led to the condition, are unlikely to be discernible in a study of cross-sectional design.

Cohort studies are generally more useful than cross-sectional studies. Cohort studies may be either *prospective* or *retrospective* in design, and basically follow up the exposure profiles and health outcomes of a defined cohort of workers over time. In general, to be of value, a cohort study needs to be fairly large scale, particularly when the health outcome of interest is quite rare (e.g. leukaemia). Also, cohort studies can require long periods of follow-up time to allow the health outcome to be expressed. These factors mean that cohort studies are expensive to conduct.

An alternative to the cohort study is the case-control study. Here, the study begins by identifying a group of cases with a defined health condition. The cases can be drawn from the general population, or from within an existing cohort study (nested case-control design). A control group then has to be selected from the same population from which the cases were drawn. Ideally, the cases and control subjects should be *matched* with respect to age/sex and other potentially relevant factors (e.g. smoking status). The investigators then determine the exposure histories of the cases and control subjects, and calculate the odds that cases have been exposed to substance *x* compared with control subjects.

There are a number of potential problems associated with epidemiological studies. Exposure assessment is usually imprecise and often non-existent. A further problem is the potential for mixed-exposure situations, which can confound the ability to characterize the effects of a particular substance. Nevertheless, when performed well under favourable circumstances, epidemiological studies can and do generate invaluable toxicological data.

What key pieces of toxicological information emerge?

A basic aim of toxicology is to identify the nature of any adverse health effects that can be produced by substances, and to provide an understanding of the doses or exposure situations under which these adverse effects will (and will not) occur. At one level, toxicology can provide simple 'yes/no' answers as to whether or not the substance has certain properties, such as (for example) skin irritancy or sensitizing potential, or the ability to cause cancer. But important information can also be provided on whether or not particular organs and tissues are specific targets for toxicity, on the presence or absence of changes and the magnitude of such changes at particular dose levels, and also on the underlying biochemical mechanisms of any toxicity seen.

Probably the most widely known toxicological reference point is the 'LD₅₀' value. Historically, this has been the conventional way of defining acute toxicity. It is the dose that causes death in 50% of experimental animals (normally within 14 days after dosing) after the administration of a single dose. LD₅₀ values have been the traditional means of ranking the acute toxicity of chemicals, and have played a role in various regulatory contexts such as in the transport of dangerous goods. However, in recent years there has been a move away from the so-called 'LD₅₀' tests, and they are being replaced by tests that use fewer animals and are not designed to cause lethality, such as the *fixed dose procedure* (Van den Heuvel *et al.*, 1990).

The exploration of irritancy and sensitization end-points is sometimes restricted to obtaining 'yes/no' answers and responding accordingly in terms of decision-making, warnings and risk management advice. However, particularly for irritancy, the emerging picture can be more sophisticated, enabling the identification of degrees of response (including thresholds) at particular concentration levels.

Repeated-dose studies in experimental animals involve a range of toxicological investigations including pathology, haematology, urinalysis and clinical chemistry tests. When relevant, specific functional investigations such as on pulmonary function or neurobehaviour can also be under-

taken. Key toxicological reference points obtained from repeated-dose studies are the 'no observed adverse effect level' (NOAEL), which is the highest dose in a study that does not cause any observable adverse effects, and the 'lowest observed adverse effect level' (LOAEL), which is the lowest dose in a study at which adverse effects are detected (see Fig. 8.1). NOAELs and LOAELs relate to 'threshold' phenomena such as liver or kidney damage. For such effects, considerations of the biology involved in the toxicological process and the presence of protective and restorative defence mechanisms indicate that there will be a threshold dose level below which the effects will not be induced; the dose threshold will vary from substance to substance, depending on the toxicological mechanism involved and the potency of the substance in question.

In recent years, an alternative reference point to the NOAEL and LOAEL has been developed. This is the 'benchmark dose' (BMD) (Barnes *et al.*, 1995). The BMD refers to a dose causing an adverse effect in a specified percentage of the test group (e.g. 10% of animals showing liver necrosis). One advantage of the BMD is that it is mathematically derived from all of the dose-response data, whereas NOAELs and LOAELs reflect only single data-points. As yet, the BMD has been used mainly in the USA, and it has not gained wide use in regulatory activities in the UK.

Genotoxic potential is an important area in toxicology and is approached and dealt with somewhat differently. In the first instance, the detection of a genotoxic hazard begins with studies carried out *in vitro*, usually in both bacterial and mammalian cells. Such studies are very sensitive to the detection of genotoxicity, given that the target cells are directly exposed to maximal concentrations of the test substance. These studies are also conducted with and without added enzyme systems to mimic the potential for *in vivo* metabolism. Given the rigour that *in vitro* testing allows, negative results generally provide good reassurance for an absence of genotoxic potential.

Positive results *in vitro* do not necessarily indicate that the substance would be genotoxic under physiological *in vivo* conditions. *In vivo* investigations are required to establish whether or not this

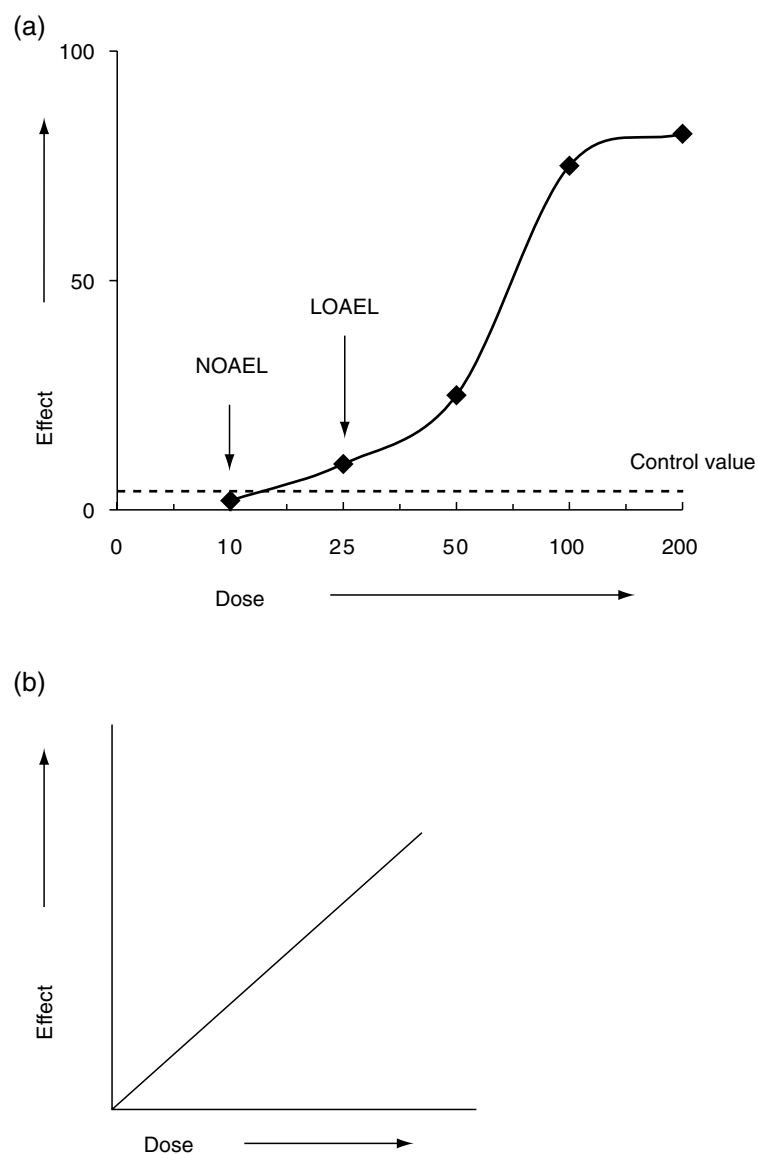


Figure 8.1 Characteristics of dose-response curves for threshold (a) and non-threshold (b) toxicological effects.

would be the case. The standard *in vivo* test methods are designed to detect genotoxic effects in either the rodent bone marrow or liver. However, to obtain meaningful results it is necessary that the test substance is 'bioavailable' to these tissues after dosing.

This can be a problem for highly reactive chemicals that are likely to interact with tissues at the immediate site of contact, usually either the stomach (following oral dosing) or respiratory tract

epithelium (following inhalation exposure). Reliable means of exploring such site-of-contact genotoxic potential have only recently appeared and few highly reactive industrial chemicals have been examined thoroughly in this respect.

Genotoxicity is conventionally treated as a 'non-threshold phenomenon'. This means that any exposure, no matter how small, is regarded as conferring some degree of risk in terms of eliciting a genotoxic change. The idea is that any molecule

of the genotoxic agent could produce one crucial change in the DNA programme, thus catalysing a series of catastrophic changes, e.g. cancer development. In reality, there are physiological defence mechanisms that can 'detoxify' molecules of genotoxicants before they can attack DNA, and there are also DNA repair mechanisms that can restore back to normal sites of DNA damage. These observations suggest that there may be practical dose thresholds for genotoxicity to be expressed. However, limitations in the sensitivity of the test methods available mean that it is not possible to determine reliable dose thresholds for the development of genotoxic effects. Hence ultimately the key issue emerging from genotoxicity testing is usually restricted to an answer to the question 'is it genotoxic *in vivo*?'. The regulatory response to identified *in vivo* genotoxicants is to treat them in a manner commensurate with no reliable threshold dose for the effect being identifiable.

In relation to carcinogenicity, the initial piece of information sought is the answer to the question 'Can it cause cancer?'. An answer of 'no' closes the issue. However, the position becomes more complex if the answer is 'yes' or even 'maybe'. Substances that are seen to have the ability to cause cancer by a DNA-damaging (genotoxic) mechanism are considered in a similar manner to that of *in vivo* genotoxicants (above), i.e. a non-threshold phenomenon is assumed. Indeed, current regulatory thinking is that once a substance has been identified as being an *in vivo* genotoxicant, it seems logical to regard the substance as if it could cause cancer, and that no further testing should be required to confirm this position.

However, as explained earlier, not all carcinogens are genotoxic; there are a number of non-genotoxic mechanisms of cancer development, including, for example, hormonal imbalance (e.g. diethylstilboestrol), altered gene expression (e.g. dioxins), chronic cytotoxicity (e.g. chloroform) and peroxisome proliferation (e.g. trichloroacetic acid). At present there are no *in vitro* predictive tests that can reliably identify potential 'non-genotoxic' carcinogens, and the only experimental means of detecting such carcinogens involves lifetime studies in rodents. Typically, such studies are conducted in two species (rats and mice), and the

design of these studies is meant to be sufficiently rigorous to reduce the chances of obtaining a false-negative result.

The interpretation of non-genotoxic carcinogenicity data requires considerable toxicological skill. Some non-genotoxic mechanisms identified in rodents can be set aside as being of little or no relevance to human health. For example, rodents are prone to the development of thyroid cancer when exposed to substances that cause fluctuations in thyroid hormone levels. It is widely agreed that such cancers are of no relevance to human health owing to species differences in the ability to maintain stable thyroid hormone levels (Grasso *et al.*, 1991; Thomas and Williams, 1991).

In some rodent carcinogenicity studies, a positive result may only occur at a very high dose level. Such tumours may be a consequence of toxicological mechanisms (saturation of metabolic detoxification pathways) that would not occur at more 'realistic' (occupationally relevant) lower doses. To help decide on the human health relevance of tumours occurring under such dosing conditions, biochemical studies may be used to explore the underlying mechanisms, and to inform on possible species differences, supplemented by PBPK modelling, to identify the doses that would reach the target tissues in humans.

For identified 'non-genotoxic' carcinogens, when the mechanism involved is believed to be of relevance to human health, dose-response and NOAEL/LOAEL information is sought and used, in a similar manner to repeated-dose toxicity (see above).

Such is also the case for reproductive toxicity end-points, i.e. toxicological effects on fertility and development are believed to operate via threshold mechanisms. However, the complexity of the biology involved (and in some cases the complexity of study designs used) renders this area one in which much toxicological skill needs to be exercised to interpret the emerging key data.

What do we use the information for?

Knowledge of the toxicological properties of a substance has an intrinsic value and can be put to

a variety of uses. Perhaps the primary one is in relation to regulation. Chemical regulation is about protecting human health and the environment from the harmful effects of chemicals. However, chemicals are of great socio-economic importance; they can generate wealth and employment and confer many advantages on modern society. The 'over-regulation' of chemicals could therefore be just as undesirable as the 'under-regulation'. Furthermore, imbalances at national and international levels in the way chemicals are regulated can lead to barriers to trade. One of the roles of chemical regulation is to provide a 'level playing field' via a harmonization of standards and regulatory requirements.

The regulatory framework surrounding industrial chemicals and the occupational environment characterizes both the generation of toxicological data and the way in which such data are used.

Hazard identification

One of the primary uses of toxicological information is for health hazard identification, and in a formal regulatory sense this can be linked to the way in which a substance is classified (i.e. as a carcinogen, mutagen or reproductive toxicant). In the EU, the legislative basis for classification and labelling derives from the Dangerous Substances Directive (67/548/EEC). The system involves the application of defined R-phrases (e.g. *very toxic by inhalation*) and S-phrases (e.g. *only use in well-ventilated areas*), following classification of substances by comparing the available toxicity information with established agreed criteria.

A key aim of classification and labelling is to warn users of the hazardous properties of the substance (or preparation containing the substance), and to indicate appropriate health and safety precautions. In this manner, classification and labelling may be regarded as a tool for risk management. However, the downstream consequences of classification and labelling can also be of major economic importance, as they determine the ways in which chemicals may be marketed and used. Hazard classification is the trigger for various regulations affecting all sectors involved in the

manufacture or use of chemicals, including transport, industry, consumer and agricultural sectors.

Different hazard identification and classification schemes have evolved in different parts of the world, leading to possible difficulties in the international market for the import and export of chemicals. To enhance consistency/compatibility of hazard communication schemes, there has been an international effort coordinated by the OECD to develop a globally harmonized system (GHS) of classification and labelling. GHS looks likely to become a reality in the next few years.

Further toxicological information about a substance, and appropriate advice regarding the prevention or the response to the potential consequences of such properties becoming expressed, should be contained within and conveyed to chemical recipients by the (Material) Safety Data Sheet [M]SDS].

Risk assessment

Risk assessment processes involve making a judgement about the likelihood that a human health hazard will occur under defined circumstances of exposure. Approaches to risk assessment depend on whether the health effect involved is mediated by a 'threshold' or 'non-threshold' mechanism.

For health hazards mediated via 'threshold' mechanisms, such as chronic liver damage caused by inhalation of a solvent vapour, limitations in the available data mean that often the precise threshold dose level necessary to produce this effect in humans is not known. In such circumstances, risk is assessed using a surrogate reference value such as a NOAEL, LOAEL or BMD obtained from an experimental animal study. Numerical uncertainty factors (or safety factors) are usually applied to these toxicological reference points to take account of possible animal-to-human differences (where relevant), and also to take account of possible inter-human variability within the population. A larger uncertainty factor will typically be applied to a LOAEL or BMD than to a NOAEL. Further uncertainty factors may also be applied to take account of the quality and reliability of the toxicological data. Some authorities may apply a larger uncertainty factor when the health effect

under investigation is, for example, testicular damage, than for respiratory tract irritancy. The application of uncertainty factors can reflect socio-political as well as scientific considerations (Fairhurst, 1995). Ultimately, a level of exposure is estimated at which a human health risk is considered unlikely to occur (the so-called *acceptable intake approach*). This 'safe' level of exposure may be compared to actual exposure estimates for the human population of interest. The ratio of these two values forms the basis for a judgement about human health risk. Alternatively, the size of the margin between the surrogate reference point (e.g. a NOAEL) and actual exposure is used directly, alongside the various uncertainties inherent in the case under examination, to gauge the potential risk involved in a particular situation (the so-called *margin of safety approach*).

For 'non-threshold' effects, the key example being genotoxic carcinogenicity, some organizations conduct quantitative risk assessments based on quantitative extrapolation of a dose–response curve, using mathematical modelling, well beyond the region of the dose axis in which actual observations have been made. A variety of models are available, with probably the most common being the linearized multistage model (LMS). The result can be the estimation, for example, of the unit dose causing a 10^{-6} lifetime risk of cancer. This may be used as a reference point for standard-setting purposes.

In contrast, in UK regulatory activities, such cancer risk assessment approaches are generally not advocated (Committee on Carcinogenicity, 1991). This reflects a lack of confidence in the calculated numerical risk estimates, given that none of the mathematical models is validated, and that different models yield a disturbingly wide range of estimates, possibly three orders of magnitude apart. Furthermore, most cancer risk assessments are based on incomplete data associated with a considerable degree of uncertainty. Hence the position reached in UK risk assessments for 'non-threshold' end-points is often that there might be a risk at the exposure level(s) under consideration; what to do about this situation then becomes a socio-political issue.

Standard-setting

One of the ways in which regulatory authorities and systems facilitate risk assessment and risk management decision-making is through the setting of exposure standards – in the occupational context, occupational exposure limits (OELs). OELs refer to airborne concentrations of a substance measured over a specified time-weighted average (TWA) period, usually either an 8-h TWA or a 15-min short-term exposure limit (STEL). Different countries and organizations follow different principles and approaches to the setting of OELs. A key area of difference relates to the extent to which the costs or practicalities of control to the desired OEL are taken into account. Most systems also include the provision of additional notations to indicate other potential problems, such as the threat of toxicity via skin uptake.

OEL-setting is a resource intensive process that few authorities across the world can support. The process varies from one OEL-setting body to another but, at its most complex level, it involves detailed toxicological assessments, including an evaluation of toxicokinetic information, *in vitro* toxicity data, studies in animals and also (when available) human volunteer and epidemiological studies. Toxicological data are used in the manner described above, including the approaches taken for assessing 'threshold' and 'non-threshold' effects. This is accompanied by occupational hygiene assessments and a review of air monitoring, biological monitoring and analytical methods. There may also be an assessment of the economic costs and perceived health benefits of controlling to various proposed OELs. There is usually at least one multidisciplinary expert committee involved in deriving the OEL and, in some cases, different subcommittees (scientific and socio-economic) may be involved. Only a limited number of OELs can be generated even on a world-wide basis, and existing lists of OELs will eventually become outdated as new data become available. Given the many thousands of industrial chemicals in existence, as well as the novel chemicals coming onto the market, it is clear that substance-specific OELs can only apply to a fraction of substances in use.

Hazard banding

As a more all-encompassing approach to the way in which toxicological data can be used to facilitate risk management, a hazard-banding system has been developed in the UK by toxicologists and hygienists, aimed at providing control advice for the many workplace chemicals that do not have a specified OEL (Brooke, 1998). (The scheme is also designed to provide practical control advice for chemicals that *do* have an OEL.) The scheme is known as COSHH Essentials (COSHH refers to the Control of Substances Hazardous to Health Regulations, which provides the legal framework for OELs in the UK). One of the drivers for the development of COSHH Essentials was the realization that the correct workplace control of chemicals, including the understanding and application of OELs, requires professional expertise, and this is often beyond the scope of many small firms.

The key features of the scheme are as follows. The scheme contains five hazard bands, differentiated according to either the OEL value or the classification of the substance according to its toxicological hazards. The user allocates the substance to the appropriate hazard band, and then works down a decision tree that takes account of the nature of the substance (solid/liquid), other relevant parameters such as vapour pressure, operating temperature, amount used and nature of the task (e.g. laminating or drum-filling). This leads the user to a control solution. The control solution can either be a specific detailed control advice sheet or one of three general control approaches (general ventilation, engineering control or containment). The control solution is designed to control the airborne concentrations of the substance to within a specified target exposure range. Guidance on the use of personal protective equipment is also a feature of this hazard-banding scheme. The aim is that for substances with an OEL, by following the scheme, compliance with the OEL should be achieved. It is a precautionary scheme (i.e. errs on the side of 'over-control' rather than 'under-control', and has been validated to demonstrate that this is the case). COSHH Essentials is not a comprehensive scheme, it has not been designed to work for gases, nor does it work for process-

generated dusts and vapours. However, it offers a simple means of controlling chemicals, based on their toxicological hazards.

Regulatory framework for industrial chemicals

The regulation of industrial chemicals can be thought of as being aimed at either 'supply-side' or 'user-side' controls (Fig. 8.2). These affect all stages of the life cycle, from manufacture through to distribution and use and, finally, dispersal to the environment. 'Supply-side' regulation places the onus onto manufacturers and suppliers of chemicals to accurately communicate the hazards of the substances supplied, and to provide advice on appropriate protective measures. This can be achieved via classification and labelling information on containers/packages and on safety data sheets. In addition, the chemical industry itself has become more proactive in recent years in terms of 'self-regulation', and initiatives such as 'product stewardship' programmes aim at providing hazard and control advice to downstream users. 'User-side' controls place responsibility on those at the end of the supply chain to take heed of the health and safety information on the substances supplied, and for managing the risks. In the occupational context, 'user-side' controls include OELs and biological monitoring standards. To varying extents, chemical regulation may be backed up by regulatory inspection and enforcement activities.

When it comes to deciding how best to regulate a chemical, one of the problems encountered by regulatory authorities is that many of the chemicals on the market have very few toxicity data. The hazards and risks posed to man and the environment are inadequately characterized for large numbers of chemicals. This situation prompted the development of a number of national and international programmes aimed at stimulating the chemical industry to generate basic hazard information. Of key importance has been the High Production Volume (HPV) programme coordinated by the OECD. More recently, an additional voluntary programme has been developed by industry, the ICCA (International Council of Chemicals Associations). This has essentially the same

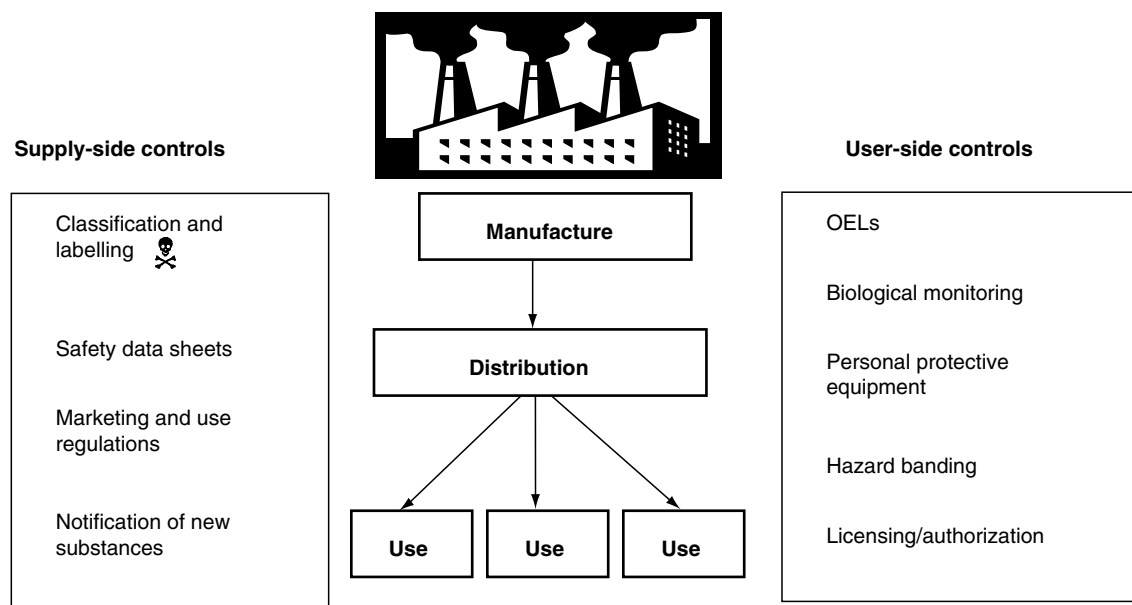


Figure 8.2 Regulatory framework for the control of industrial chemicals.

aims as the OECD HPV programme and is feeding into it.

Earlier, a programme was developed in the EU under the framework of the Dangerous Substances Directive for the regulation of ‘new’ chemical substances. In the UK this is administered via the Notification of New Substances (NONS) regulations. The purpose is to ensure that there should be an assessment of a new chemical’s toxicological, environmental and physicochemical hazards before it is placed on the market. A ‘new’ substance is arbitrarily defined as one that was not registered on EINECS (European Inventory of New and Existing Chemical Substances) by 1981. In the mid-1990s, a further EU programme was developed, which focused on ‘existing’ chemical substances. This is administered via the Existing Substances Regulation (ESR). The requirements of NONS and ESR are fundamentally similar. In each programme, the level of information submitted by the notifier depends on the tonnage levels manufactured or imported into the EU. The information generated by both programmes is used for hazard identification (classification and labelling), for decisions on whether further testing is needed, and for risk assessment with the possibility of leading to risk

management outcomes. These may include restrictions on marketing and use, or when occupational health concerns have been identified, the development of an EU-wide package of occupational risk management measures, including an OEL.

Some themes for the future?

At the time of writing this chapter in 2004, there is a general feeling that the science, and the occupational/regulatory application of toxicology are about to undergo substantial changes over the next decade. It will be interesting to see just how much of a real shift occurs as a result of the developments now under way.

In terms of the application of toxicology to regulate and control industrial chemicals, there are changes afoot wherever one looks. An EU ‘Chemicals White Paper’ produced in 2001 heralded the intention to introduce a new regime to cover the ‘supply-side’ of industrial chemicals regulation in the European Union. The acronym ‘REACH’ (Registration, Evaluation and Authorization of Chemicals) has been introduced to denote the planned new system.

The basic philosophy of REACH is to secure, collate and make available better information on the potential adverse properties of all significant industrial chemicals; to investigate further, in an intelligent and substance-specific manner, those substances deemed worthy of this; and, for those substances with properties regarded as posing the most worrying threats to human health or the environment, to require industry to seek clearance (authorization) from the regulatory system to use them for any particular purpose – a request for permission that might be denied.

It is envisaged that this legislation will replace NONS and ESR. It is also envisaged that under REACH, substances and preparations will be classified not in accordance with the existing EU Classification and Labelling system, but according to the criteria of the newly developed GHS, which is aimed at commonality of classification around the world.

The desire for more toxicological information on industrial chemicals, when put alongside the huge numbers of substances involved (approximately 30 000 substances marketed in quantities of 1 tonne or more per year), has triggered much concern about the implications for the scale of animal usage (and the costs), if further information is pursued in the conventional manner of animal experimentation. Sensitivities have been heightened on what is already a sensitive, controversial ethical issue. Hence there is at present a strong drive to further develop, promote and bring into real meaningful use alternative means of investigating or deducing the toxicity of industrial substances. It is anticipated that the new REACH system will see much greater use (compared with current programmes) of techniques involving predictions of toxicity that are based on chemical structure (QSAR) and/or physicochemical properties; and of *in vitro* methods using cultured biological systems (isolated biological molecules, cells, tissue sections, whole organs) to identify toxic properties.

It will be interesting to see how much can be done with such approaches. Another significant and challenging development in the science of toxicology is the advent of ‘-omics’: genomics, proteomics, etc. In essence, new biochemical techniques

of the types developed and used in the recently undertaken human genome project have given toxicologists the ability to readily and simultaneously monitor for changes vast numbers of different pieces of genetic material and the proteins derived from their genetic codes. There is a hope (perhaps an expectation?) that such approaches will deliver an ability to be much more specific, selective and incisive in our understanding of toxicological processes, and that this will in turn bring about further new approaches and capabilities, in both a predictive (what adverse effect might happen in exposed humans?) and a diagnostic (what adverse effect has happened/is happening in exposed humans?) sense. Whether or not such hopes/expectations are realized and, if so, over what timescale the significant changes in what toxicology can achieve are delivered are big questions with no definitive answers at present.

References

- Andersen, M.E. and Krishnan, K. (1994). Physiologically based pharmacokinetics and cancer risk assessment. *Environmental Health Perspectives*, **102**(Suppl), 103–8.
- Barnes, D.G., Daston, G.P. and Evans J.E. (1995). Benchmark dose workshop: criteria for use of a benchmark dose to estimate a reference dose. *Regulatory Pharmacology and Toxicology*, **21**, 296–306.
- Barratt, M. (2000). Prediction of toxicity from chemical structure. *Cell Biology and Toxicology*, **16**, 1–13.
- Brooke, I.M. (1998). A UK scheme to help small firms control health risks from chemicals; toxicological considerations. *Annals of Occupational Hygiene*, **42**, 337–90.
- Committee on Carcinogenicity of Chemicals in Food, Consumer Products and the Environment (1991). *Guidelines for the Evaluation of Chemicals for Carcinogenicity*. Department of Health. Report on Health and Social Subjects 42. HMSO, London.
- Committee on Mutagenicity (2000). Guidance on a strategy for testing of chemicals for mutagenicity. Department of Health. HMSO, London.
- Delic, J.I., Lilly, P.D., MacDonald, A.J. and Loizou, G.D. (2000). The utility of PBPK in the safety assessment of chloroform and carbon tetrachloride. *Regulatory Pharmacology and Toxicology*, **32**, 144–55.
- Eisenbrand, G., Pool-Zobel, B., Baker, V., Balls, M., Blaauboer, B.J., Boobis, A., Carere, A., Kevekordes, S., Lhuguenot, J-C., Pieters, R. and Kleiner, J. (2002). Methods in *in vitro* toxicology. *Food and Chemical Toxicology*, **40**, 193–236.

- Fairhurst, S. (1995). The uncertainty factor in the setting of occupational exposure standards. *Annals of Occupational Hygiene*, **39**, 375–85.
- Grasso, P., Sharratt, M. and Cohen, A.J. (1991). Role of persistent, non-genotoxic tissue damage in rodent cancer and relevance to humans. *Annual Reviews of Pharmacology and Toxicology*, **31**, 253–87.
- Organization for Economic Co-operation and Development (OECD) (1993). *OECD Guidelines for the Testing of Chemicals*, vols 1 and 2. OECD, Geneva.
- Russell, W.M.S. and Burch, R.L. (1959). *The Principles of Humane Experimental Technique*. Methuen, London.
- Thomas, G.A., and Williams, E.D. (1991). Evidence for and possible mechanisms of non-genotoxic carcinogenesis in the rodent thyroid. *Mutation Research*, **248**, 357–70.
- Van den Heuvel, M.J., Clark, D.G., Fielder, R.J., Koundakjian, P.P., Oliver, G.J.A., Pelling, D., Tomlinson, N.J. and Walker, A.P. (1990). The international validation of a fixed dose procedure as an alternative to the classical LD₅₀ test. *Food and Chemical Toxicology*, **28**, 469–82.

Part 3

Principles of occupational hygiene

Chapter 9

The nature and properties of workplace airborne contaminants

Lisa M. Brosseau and Claudiu T. Lungu

Introduction	The motion of airborne particles
Physical properties of matter	Drag force on a particle
Basic properties of gases and vapours	Motion under the influence of gravity
Vapour pressure	Impaction and interception
Density	Elutriation
Humidity	Aspiration
The ideal gas laws	Diffusion
Partial pressure	Interactions of airborne pollutants with electromagnetic radiation
Transport properties of gases and vapours – diffusion	Aerosols
Adsorption	Gases
Basic properties of aerosols	Summary
Aerosol generation in workplaces	Further reading
The evolution of aerosols	
Particle shape	
Particle size	
Elementary particle size statistics	

Introduction

The science and practice of occupational hygiene is concerned with the interaction between humans and the working environment. An important facet of the working environment is the surrounding air, in which numerous hazardous materials may be present. Broadly speaking, these materials are classed as air pollutants, existing as matter (gases or aerosols) or energy (heat, sound, light and ionizing or non-ionizing radiation). This chapter is concerned with airborne pollutant matter in the form of gases, vapours and aerosols.

The physical properties of matter are important in helping to understand how pollutants (gaseous or aerosol) are generated and dispersed in the workplace air, transported to the part of the worker–environment interface where they are likely to be troublesome, and monitored and controlled. This chapter sets out to provide a basic framework of relevant physical ideas. It starts with a brief résumé of the general physical

properties of matter, describing how the gaseous, liquid and solid phases are related to one another and how phase changes can take place. We then describe some important properties of the air which determine the behaviour of airborne pollutants. There follows a description of the nature and behaviour of gases, vapours and aerosols, with an emphasis on those properties important to their generation and measurement in workplace environments. We then briefly describe interactions between electromagnetic radiation and airborne pollutant matter (gases, vapours and aerosols) with emphasis on applications in monitoring methods. Examples are given throughout to illustrate how knowledge of these scientific subjects is important to occupational hygienists.

In view of the wide range of topics encompassed, the treatment throughout is necessarily of an introductory nature, and the interested reader is recommended to consult more specialized texts for in-depth coverage of specific areas. Some of these

are listed in the 'Further reading' section at the end of the chapter.

Physical properties of matter

Matter is usually acknowledged to exist in three phases: solid, liquid or gas. It consists of small particles called atoms, which in turn are made up of combinations of so-called fundamental particles of matter, namely protons, neutrons and electrons. Each particular combination of these defines an element. Under certain conditions, atoms may combine together to form larger entities known as molecules. Whether or not atoms or molecules come together to form solids, liquids or gases depends on combinations of pressure, volume and temperature. The most familiar example is water, which, over the ranges of familiar terrestrial conditions, can exist as solid ice, liquid water or gaseous water vapour.

In the solid state, atoms are located in fixed positions, which, for many stable materials, are arranged in regular and periodic patterns constituting the familiar stable crystallographic lattice structure. So-called amorphous (non-ordered) materials (e.g. glasses) are not strictly stable, and in time – sometimes a very long time – become crystalline. The atoms of a solid material are held together in this ordered way by inter-atomic forces, electrostatic in nature, which may be likened to a system of invisible springs by which each is connected to its neighbours. When we speak of atoms occupying fixed positions in the crystal lattice of a solid material, it should be understood that we are referring to their mean positions. In fact, at any temperature above absolute zero, 0 kelvin (K), the atoms are in oscillatory motion about their mean locations and, as in any spring–mass system in motion, energy is continually being exchanged between the kinetic form (associated with velocity) and the potential form (associated with displacement). Averaged overall, energy is shared equally between the two energy forms.

If extra internal energy is given to a lump of solid matter in the form of heat, then the atoms perform greater excursions about their mean locations. If enough energy is supplied, the solid melts and

enters the liquid phase. At that point, bonds may be broken and remade, and individual atoms can move around in the lattice, changing places with one another. The state of the material has now become 'fluid'. It is of particular interest to occupational hygienists to consider what happens near the surface of a liquid. Atoms there are connected by their invisible 'springs' only with other atoms in the general direction of the body of the liquid; so, unlike atoms in the body of the liquid, they experience a net inwards-seeking force. This accounts for the well-known phenomenon of surface tension. However, there is a statistical probability that a given atom located instantaneously near the surface of the liquid may escape from the surface as a free entity and enter the gaseous vapour phase. Thus, we have the phenomenon of evaporation. Conversely, atoms of molecules in the vapour phase may enter the liquid through the surface, and so contribute to condensation. The magnitude and direction of the net flux of molecules across the surface are controlled by complex thermodynamic considerations.

If enough energy is supplied to a liquid, then a temperature is eventually reached at which all the inter-atomic bonds can be broken permanently. All the atoms or molecules now become free to move at random, and the liquid becomes a gas in which all of the internal energy exists as kinetic energy.

The preceding scenario for the transition from the liquid to the gaseous or vapour phase applies in principle to all substances. For example, under extreme thermodynamic conditions (e.g. very low temperature), even a gas such as helium can become a liquid. In relation to occupational hygiene, however, it is the convention to refer to gases as substances, which, under workplace conditions, are always found in the free molecular phase (e.g. air). On the other hand, vapours are regarded as the free molecular phase of some other substances (e.g. organic solvents), which can, in the workplace, also be found in the liquid state.

Basic properties of gases and vapours

Occupational hygiene is concerned with the transport of pollutants of various kinds in the vicinity of

human subjects, both through and by the workplace atmospheric air. In general, vapours are produced in the workplace as a consequence of volatile liquid evaporation. Most solvents, cleaning liquids and oil-based products are highly volatile and they have an increased potential to become airborne vapour at room temperature. Gases that are non-condensable at room temperature can be released in the environment as a result of various industrial processes or from chemical reactions of solid or liquid chemicals.

Air is a mixture of gases, the main constituents being nitrogen (about 78% by volume) and oxygen (21%), with a variety of other trace gases (amounting to about 1% in total), including argon, carbon dioxide and water vapour. It is a colourless, odourless gas with a density of 1.29 kg m^{-3} at standard temperature of 293 K and pressure of $1.01 \times 10^5 \text{ Pa}$ sea at level (STP).

In the widest sense, 'air pollution' defines the presence in the atmospheric air of entities of matter or energy, naturally occurring or synthetic, which have the potential to cause harm. In the context of occupational hygiene, this relates to the health and well-being of employees.

The universal unit of the concentration of any pollutant is its mass per unit volume of the atmosphere itself (e.g. micrograms of pollutant per cubic metre of air, or mg m^{-3}). However, for gases and vapours it is also common to talk in terms of the partial volume occupied [e.g. parts per million (ppm) or parts per billion (ppb)]. For gases and vapours, the relationship between forms of expression is given (for STP conditions) by:

$$(\text{mg m}^{-3}) = \frac{(\text{ppm}) \times \text{MW}}{24.5 (\text{L mol}^{-1})} \quad (9.1)$$

where MW is the molecular weight (mass of gas or vapour in g mol^{-1}) of the material in question. Take, for example, the common gaseous air pollutant, sulphur dioxide. At a mass concentration of 0.3 mg m^{-3} , an exposed person would soon become aware of its presence. From Equation 9.1, this is equivalent to a partial volume of about 10^{-7} (or 0.1 ppm).

For aerosols, concentrations are usually expressed in terms of the mass per unit volume of air (e.g. mg m^{-3}). However, depending on the measurement method used, aerosols may also be expressed in terms of the surface area of particulate per unit volume air (e.g. as might be obtained using a light-scattering instrument) or the number of particles per unit volume of air (e.g. as might be obtained for asbestos fibres using an optical microscope).

Some of the above principles can be applied to materials, which, although normally existing in the liquid phase, can also appear as vapours in air. This is a situation commonly encountered by occupational hygienists, as not all such materials are harmless.

Vapour pressure

Vapour pressure (VP) represents the pressure that would be exerted by vapour molecules in equilibrium with the same material in liquid form inside a closed container. For a material starting out as 100% liquid in such a closed system, some of the molecules will evaporate into the vapour phase. For some materials, the attractive molecular forces between liquid molecules are relatively weak, so that the pressure exerted by that liquid in the closed container would be relatively high, as a high proportion of the material will be present in the vapour phase. Conversely, for materials with stronger intermolecular forces, relatively fewer molecules will be present in the vapour phase – so the vapour pressure will be correspondingly lower. Thus, it follows that materials with high vapour pressures are more likely to evaporate into the air than those with relatively lower vapour pressure. For example, hydrazine (N_2H_4 , a colourless liquid) has a vapour pressure at STP (VP_{STP}) of 10 mmHg, whereas hexane ($\text{CH}_3(\text{CH}_2)_4\text{CH}_3$, another colourless liquid) has a vapour pressure of 124 mmHg. Thus, the magnitude of vapour exposure is likely to be greater for hexane than for hydrazine.

This discussion leads to a concept useful to occupational hygienists – the vapour-hazard ratio (VHR). For a given material, this is defined as:

$$\text{VHR} = \frac{\text{SC}}{\text{OEL}} \quad (9.2)$$

where OEL is the relevant occupational exposure limit for the material in question, established on the basis of the material's toxic properties for humans (in parts per million by volume, ppm) and where SC is the saturation concentration (also in ppm) given by:

$$\text{SC} = \frac{\text{VP}_{\text{STP}} \times 10^6}{\text{BP}} \quad (9.3)$$

in which barometric pressure (BP) is 760 mmHg.

A liquid's ambient saturation concentration reflects the magnitude of its vapour pressure compared with the vapour pressure of the air above it. When a pool of liquid is evaporating within an enclosed space, the amount of evaporated vapour within that space eventually will stabilize at an equilibrium level called *ambient saturation concentration*. If a chemical has a high ambient saturation concentration, it has a strong ability to displace air, and the concentration of the chemical's vapour in the air will be high. This property changes with temperature such that a liquid at higher temperature will have a higher ambient saturation concentration.

Applying the above to the examples of hydrazine and hexane, we obtain the following

- hydrazine: SC = 13 158 ppm and OEL = 1 ppm → VHR = 13 158
- hexane: SC = 163 158 ppm and OEL = 500 ppm → VHR = 326

from which we see that hydrazine is potentially much more hazardous to health, despite its lower vapour pressure and, hence, lower magnitude of exposure.

In some cases it is also important to consider the extent to which a material, when it is airborne, can exist as a vapour or an aerosol. To quantify this, SC – as defined above in Equation 9.3 – is first converted into a mass concentration (mg m^{-3}). This is then compared with the OEL (also expressed in mg m^{-3}). Thus, we have the following possible scenarios:

1 if $\text{SC/OEL} < 1$, the airborne material will appear mostly as aerosol;

2 if $1 < \text{SC/OEL} < 100$, the airborne material will contain some aerosol;

3 if $\text{SC/OEL} > 100$, the airborne material will appear as vapour.

For example, mercury has an OEL listed as 0.05 mg m^{-3} , under the assumption that the material is present as vapour and that there is no aerosol exposure. Mercury has a vapour pressure of $1.8 \times 10^{-3} \text{ mmHg}$, leading to $\text{SC} = 19.6 \text{ mg m}^{-3}$. In turn, this leads to $\text{SC/OEL} \rightarrow 19.6/0.05 = 393$. Therefore, this confirms that the setting of an OEL for mercury, based on the assumption of a vapour, is correct.

Density

Another physical property of pollutant gases and vapours in air is that associated with their density. Significant differences in density in relation to that of the air itself can lead to stratification. For example, we note that the density of carbon dioxide is 1.98 kg m^{-3} (compared with 1.29 kg m^{-3} for air), and it is well known that, in still atmospheres, it tends to accumulate near the floor. Although carbon dioxide is not toxic in itself, the fact that it displaces oxygen during this stratification can present a hazard to the unwary in certain confined spaces. However, this is not a problem in most industrial settings when there is usually sufficient mixing to prevent stratification.

Humidity

Water vapour is a normal constituent of air and is innocuous. So it is not a pollutant. However, it does not form a constant atmospheric constituent as the changes between phases for water (between solid ice, liquid water and gaseous water vapour) can all occur within the range of expected atmospheric conditions, even in workplaces. Atmospheres with high humidity can affect the properties and distribution of vapour and aerosol airborne pollutants. Some vapours can, under certain conditions, chemically interact with water, resulting in more harmful compounds. For example, sulphur dioxide can interact with water to produce sulphuric acid. Aerosol size distribution

and settling velocity can also be affected by the presence of water vapour.

The physical picture presented earlier to describe how molecules of a liquid can enter the gaseous vapour phase may be enlarged to enable discussion of the important environmental question of humidity. This relates to the presence in the air of free water molecules. The mass of water vapour per unit volume of air is referred to as the absolute humidity. Its partial pressure cannot exceed the vapour pressure of water for a given temperature and atmospheric pressure. It reaches a pressure of 1 atmosphere (atm) (1.01×10^5 Pa) at the temperature at which water boils (393 K).

Air is considered to be saturated with water vapour when its partial pressure becomes equal to the vapour pressure. At lower pressures it is unsaturated, and relative humidity (RH, expressed as a percentage) is given by:

$$\text{RH} = \frac{\text{partial pressure of water}}{\text{vapour/vapour pressure of water at the same temperature}} \quad (9.4)$$

For a given mass concentration of water vapour in the air, RH can be raised by lowering the temperature. Conversely, raising the temperature lowers RH. The temperature at which water vapour becomes saturated is known as the *dew point*. Below this, nucleation and condensation may take place, hence the appearance in the air of water droplets visible as mist or fog.

The ideal gas laws

A gas or vapour can be completely characterized by the volume it occupies, its pressure and temperature. If N molecules are trapped in a box of volume V , the collision with the box walls will create a net force exerting a pressure p on the walls.

The temperature, T , of a gas is the result of intermolecular collision and depends on the velocity of colliding molecules. The behaviour of both vapours and gases in the workplace can be described in most cases by the ideal gas law expressed by:

$$pV = nRT \quad (9.5)$$

At sufficiently low pressure (as in the case of workplace atmosphere) the product of pressure p and

volume V is proportional to the amount of gas (described as the number of kilomoles, n), the absolute temperature of the gas T , and a constant R . Experiment shows that at low enough density and pressure R has the same value for all gases, namely:

$$R = 8.314 \text{ J/kmol K} \quad (9.6)$$

R is called the *universal gas constant*. Equation 9.5 represents the equation of state for a gas or vapour – it is impossible to force a gas into a state of pressure, volume, temperature and amount that does not satisfy this expression.

Of particular interest for the occupational hygienist are the expressions derived from this law concerning the response to pressure and temperature. At constant temperature, the volume of a gas is proportional to the inverse of its pressure ($V \sim 1/p$). On the other hand, both the volume and the pressure are proportional to the temperature ($V \sim T$ with p held constant; $p \sim T$ with V held constant). These relationships are important for determining the properties of gases and vapours under various environmental or process conditions.

Because the properties of gases and vapours are listed in most cases at standard conditions of temperature and pressure, STP, it is often necessary to apply the equation of state to determine the properties of the pollutant under the real working conditions. For example, if an analytical method requires the sampling of a volume of 24 l of air at STP, to calculate the volume of air needed to be sampled at 30°C (303 K) and 1.0×10^5 Pa the following conversion needs to be applied.

$$V_{\text{new}} = V_{\text{STP}} \frac{p_{\text{STP}}}{p} \times \frac{T}{T_{\text{STP}}} \quad (9.7)$$

$$V_{\text{new}} = 24 \frac{1.01 \times 10^5}{1.0 \times 10^5} \times \frac{303}{293}$$

Under these conditions, 25 l should be sampled.

Partial pressure

The working environment is rarely composed of a single gas or vapour. In many processes, a pollutant gas or vapour is introduced into a container or room containing an initial gas or vapour.

To determine the total pressure exerted on the container or room, Dalton's law of *partial pressure* is applied. This law states that the pressure exerted by a mixture of gases behaving ideally is the sum of the pressure exerted by the individual gases occupying the same volume alone. To determine the total pressure exerted on a 10-l (10^{-2} m^3) container from two gases (e.g. 1 mol of nitrogen and 3 mol of hydrogen), the partial pressure of each component can be determined from the ideal gas law: $p_i = n_i (RT/V)$. The calculations give a partial pressure of $2.47 \times 10^5 \text{ Pa}$ (2.44 atm) for nitrogen and $7.42 \times 10^5 \text{ Pa}$ (7.32 atm) for hydrogen. According to Dalton's law, the total pressure will be the sum of the two partial pressures: $p_{\text{tot}} = (2.47 + 7.42) \times 10^5 \text{ Pa} = 9.89 \times 10^5 \text{ Pa}$ (9.76 atm).

Another way of expressing Dalton's partial pressure law is by using the *mole fraction*. If we have a mixture of gases A and B in the amount of n_A and n_B , then the fraction $x_A = n_A/n$ represents the mole fraction of component A present in the mixture. The sum of the total mole fractions in a mixture is unity, and the partial pressure law for the two gases can be expressed as:

$$p = p_A + p_B = \frac{(n_A + n_B)RT}{V} = \frac{(x_A + x_B)nRT}{V} \quad (9.8)$$

Transport properties of gases and vapours – diffusion

Transport processes occur in gases, liquids and solids and are not confined to mass transfer. Electrical charge, energy (heat) and momentum can be transported from one region of a system to another. In the case of gases and vapours, the transport of mass will allow the molecules to 'flow' from one region to another until equilibrium is reached. It is important to mention here that the main mass transport process for gases and vapours (diffusion) can be viewed as a passive process. The flow of gas or vapour molecules inside a container or process room, for example, will occur without any outside intervention until equilibrium is reached. Through ventilation the gas and vapour molecules can also

be transferred from one region to another of the system, but in this case the pressure drop needed for the flow is mechanically created.

The mass transfer of a gas or vapour as a result of the random motion of its molecules when a concentration gradient is present is called diffusion. Diffusion is a general term referring either to a single gas or vapour that seeks to attain concentration equilibrium or to mixtures for which the equilibrium consists of uniform composition. For example, if a gas is confined to a container that is open to a low-pressure region through a small hole, the gas will flow through the hole until the pressures are equal on both sides. This process is called *effusion*, and in this case the concentration gradient is created by the differences in pressures on either side of the hole. In stricter terms, *diffusion* is the penetration of the molecules of one gas or vapour through the molecules of another until the composition is uniform throughout. In the practice of occupational hygiene, when a volatile liquid evaporates in air, the concentration in a room tends to become uniform because of the diffusion. Obviously, this process will allow for the contaminant to be diluted, creating a lower concentration throughout the room. On the other hand, as in ventilation, the contaminant will be transported to regions of the room far away from the contaminant source. In this way, people can suffer the effect of the contaminant without being aware of its immediate presence.

Graham's law states that the rate of diffusion of a gas is inversely proportional to the square root of its molecular weight. In the case of a gas A that diffuses through a gas B, Graham's law can be expressed as:

$$\frac{R_A}{R_B} = \frac{\sqrt{m_B}}{\sqrt{m_A}} \quad (9.9)$$

where R_A and R_B are the rates of diffusion of gases A and B and m_A and m_B are the molecular weights.

An important application of diffusion in occupational hygiene is the principle of passive, or diffusive, sampling of gases and vapours. The passive sampling is based on Fick's first law of diffusion:

$$J = -D \frac{dC}{dx} \quad (9.10)$$

This law states that the amount of gas or vapour passing per unit time through the unit area perpendicular to the direction of diffusion (the entity defined in this way is called flux) is proportional to the concentration gradient, dC/dx , and the diffusion coefficient, $D(\text{cm}^2 \text{s}^{-1})$. The negative sign in this expression is due to the fact that a positive mass flow is determined by a decrease in concentration. If the geometrical parameters (surface area, length of diffusion path) of a sampler are known then the mass of gas or vapour transported through diffusion and adsorbed by a sorbent material inside the sampler is proportional to the ambient concentration of the pollutant.

Adsorption

Adsorption has been defined as the enrichment of one or more components in an interfacial layer. Although the adsorption of liquid molecules on to solid surfaces is possible, the most relevant process for occupational hygiene is the adsorption of gases and vapours onto solid, porous materials. There are two basic types of adsorption. In *chemisorption* or chemical adsorption the molecules bind to the surface as a result of the formation of a chemical – usually covalent – bond. The energy of attachment is strong and the molecules are transformed and lose their identity.

Physical adsorption occurs as the result of van der Waals interactions between the surface and the adsorbed molecules. As a result of this long-range, weak interaction, molecules are attached to the solid surface. However, the energy of this interaction is insufficient to lead to bond breaking, and thus in physical adsorption the adsorbed molecules retain their identity, although they might be stretched or bent at the proximity of the surface. Both of these processes are exothermic, but the amount of heat released in chemisorption is usually much higher than the heat released in physical adsorption, owing to the strength of the interaction.

Adsorption of gases and vapours on porous surfaces has two main applications in occupational hygiene:

1 Respirators use adsorbing materials inside chemical cartridges to prevent various organic vapours and gases entering the respiratory tract.

2 One method of sampling for gases and vapours is by drawing contaminated air through a sampling tube filled with gas/vapour-adsorbing material, and afterwards analysing the content of the adsorbent.

Both of these applications use a variety of adsorbents, but the material used most widely is activated carbon. Zeolites (an artificial adsorbent), silica, alumina or titanium oxides, as well as catalyst-impregnated activated carbon are also used, depending on the properties of the adsorbed gas or vapour.

‘Activated carbon or charcoal’ usually means a porous form of carbon produced by the carbonization of some naturally occurring material such as wood, peat or nut shells. The activation process takes place in two stages: carbonization followed by the removal of hydrocarbon tarry products from the interstices formed during carbonization. The final goal is to obtain a material composed of carbon crystallites with a complex pore structure increasing the surface area of the adsorbent. The internal surface of the product obtained through this process ranges from 400 to 1600 $\text{m}^2 \text{g}^{-1}$.

The adsorption isotherm represents the central concept for characterizing adsorbents with respect to different gases and vapours. The adsorbed volume of a gas or vapour per unit mass of solid adsorbent, at a constant temperature, is a function of ambient concentration. This relationship is termed the *adsorption isotherm*, and is a measure of how an adsorbent reaches its capacity. A number of theoretical models have been developed to predict the adsorption isotherm and thus the adsorption capacity of certain adsorbents when challenged by various concentrations of gases and vapours. However, no single model can predict the adsorption behaviour of all adsorbent vapour/gas pairs.

The concept of adsorption capacity is of great importance for occupational hygiene practice. When the adsorption capacity is exceeded, the contaminant is no longer retained inside the adsorbent material and breakthrough occurs. In the case of chemical cartridges after breakthrough, toxic gases and vapours can penetrate into the respiratory system of the wearer, with possible health consequences. When breakthrough of

sampling tubes is reached, sampled gases or vapour will be lost and the obtained concentration will differ from the ambient concentration.

When a respiratory protection programme is implemented, a cartridge exchange schedule is set in place, based on the chemical usage, the type of cartridge (capacity of the activated carbon), activity and environmental conditions. Intense physical activity increases the respiration rate and therefore the service life of the cartridge will be diminished. The performance of the cartridge will also be affected by the presence of water vapours or another vapour or gas in significant amounts. Because adsorption is a surface process, if the adsorption sites are occupied by water vapours or other vapours or gases the capacity of the cartridge for a given contaminant will be significantly diminished.

Basic properties of aerosols

‘Aerosol’ is a scientific term that applies to any disperse system of liquid or solid particles sus-

pending in a gas – usually air. It applies to a very wide range of particulate systems encountered terrestrially. Aerosols occur widely in workplace environments, arising from industrial processes and workplace activity, and so are of considerable interest to occupational hygienists. They take many different forms. A summary classification of a range of typical aerosols is given in Fig. 9.1. It contains not only examples of the workplace aerosols with which this book is primarily concerned, but also, for the sake of comparison, some naturally occurring and synthetic aerosols found in the outdoor atmospheric environment. Aerosols of interest to occupational hygienists include:

- *Dust*. An aerosol consisting of solid particles made airborne by the mechanical disintegration of bulk solid material (e.g. during cutting, crushing, grinding, abrasion, transportation), with sizes ranging from as low as 1 to over 100 μm .
- *Spray*. An aerosol of relatively large liquid droplets produced by mechanical disruption of bulk liquid material, with sizes ranging from 10 to 100 μm or more.

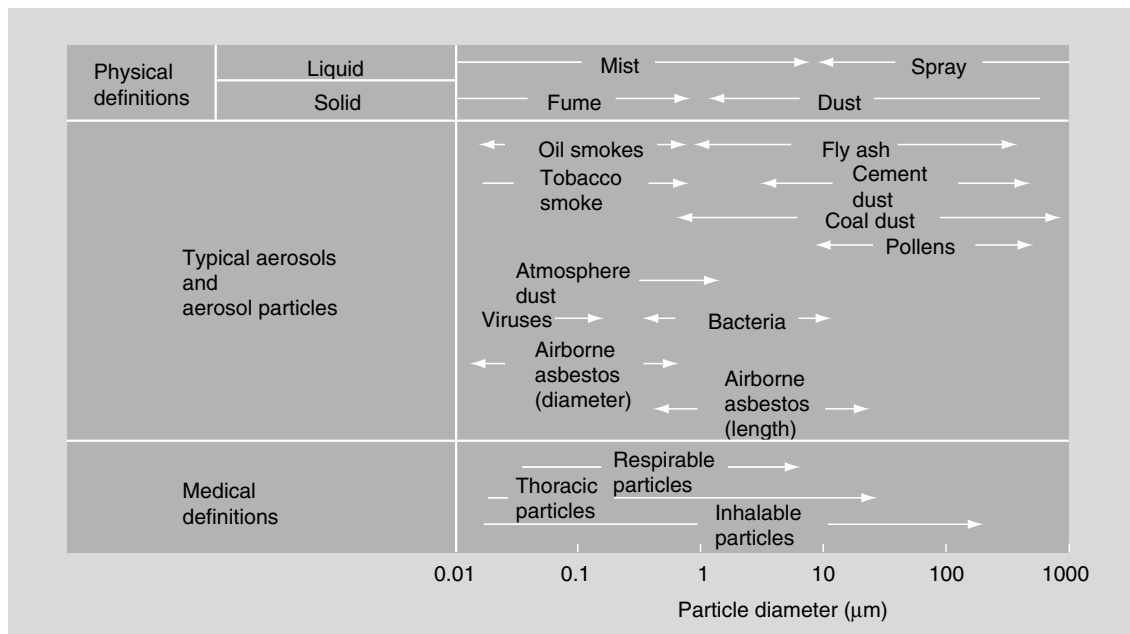


Figure 9.1 Classification of typical aerosols ('medical' definitions refer to particle size fractions where exposure to the various parts of the human respiratory tract is possible).

- *Mist*. An aerosol of finer liquid droplets produced during condensation or atomization, with sizes ranging from 0.01 to 10 μm .
- *Fume*. An aerosol consisting of small solid particles produced by the condensation of vapours or gaseous combustion products. Usually, such particles are aggregates of very small primary particles, with the individual units having dimensions of a few nanometers. Aggregates range from 0.1 to 1 μm .
- *Smoke*. An aerosol of solid or liquid particles resulting from incomplete combustion, again usually in the form of aggregates of very small (nm-sized) primary particles. The aggregates themselves have extremely complex shapes, frequently in the forms of networks or chains, having overall dimensions ranging from 0.01 to 1 μm .
- *Bioaerosol*. An aerosol of solid or liquid particles consisting of, or containing, biologically viable organisms. Viruses are generally very small, ranging from 0.01 to 0.5 μm . Bacteria are larger than viruses, ranging from 0.5 to 30 μm . Pollens are generally larger than bacteria, ranging from 10 to greater than 100 μm .

Aerosol generation in workplaces

The majority of industrial processes generate aerosols in one form or another, usually as a side-effect of the process itself and by a wide variety of physical and chemical means. These may include:

- mechanical generation of dry aerosols (e.g. during mineral extraction, smelting and refining of metals, textiles manufacture, bulk chemical production and handling, woodworking);
- mechanical generation of liquid droplet aerosols (e.g. during paint spraying, crop spraying);
- formation by molecular processes (e.g. during combustion, chemical reactions, condensation).

The evolution of aerosols

It cannot be assumed that an aerosol, once it has been dispersed, will necessarily remain in equilibrium and so retain the properties with which it began. Depending on the material in question, the initial generation process and the concentration of the aerosol, and other conditions in the

surrounding air, a number of possibilities exist for evolutionary changes. These include:

- growth by coagulation, agglomeration and coalescence (by the contact of particles with and attachment to one another), in which the number concentration of particles decreases but the mass concentration stays the same;
- disintegration (when a system of particles combined together to form a single particle is subjected to external forces such that the adhesive and cohesive bonds that hold its individual elements together are broken), in which the number concentration increases but the mass concentration stays the same;
- condensation (when particles are formed and grow by the condensation of molecules out of the vapour phase), in which the mass concentration increases;
- evaporation (where particles are decreased in size – or even disappear – by the transfer of molecules from the liquid to the vapour), in which the mass concentration decreases.

These last two phenomena extend the earlier discussion about atmospheric water and other vapours to aerosols. From detailed consideration of the physics of phase transitions from liquid to vapour – and vice versa – it may be shown that, in a system of droplets of a wide range of sizes, larger droplets can grow at the expense of smaller ones.

Particle shape

Particle shape can have a significant bearing on effects relevant to occupational hygiene, for example on the way in which particles behave in the air, and how they behave after they have been deposited in the respiratory tract. Particle shape falls into a number of categories, some of which are shown schematically in Fig. 9.2. These include:

- spherical particles (e.g. liquid mists, fogs and sprays and some dry aerosols such as glassy spheres condensing out of some high-temperature processes);
- regular or isometric, non-spherical, angular particles which have no preferred dimension or whose aspect ratio cannot be said to be substantially different from unity (e.g. most dusts, including coal dust);

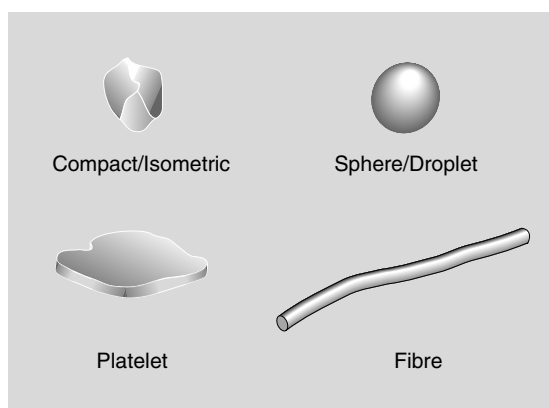


Figure 9.2 Examples of some particle shapes found in occupational hygiene.

- platelet particles (e.g. some dusts, such as mica);
- fibrous or acicular particles which are long, thin, needle-shaped particles (e.g. asbestos and man-made mineral fibre dusts);
- fractal particles, complex aggregates of much finer primary particles (e.g. fumes and smokes).

Particle size

Particle size is a property that is extremely important in virtually all aspects of aerosol behaviour. But it is a property whose definition is not always as simple as might at first appear, and can be somewhat elusive. Indices of particle size include:

- true geometric diameter (d) for a particle that is perfectly spherical;
- ‘effective’ geometric diameter (d') for a non-spherical particle, based on representative widths; for example, dividing a two-dimensional image of the particle into equal areas (Martin’s diameter) or contained within a pair of parallel tangents to the particle perimeter (Feret’s diameter);
- equivalent projected area diameter (d_p) is the diameter of a sphere that, in two dimensions, projects the same area as the particle in question;
- equivalent surface area diameter (d_A) is the diameter of a sphere that has the same surface area;
- equivalent volume diameter (d_v) is the diameter of a sphere that has the same volume;
- aerodynamic diameter (d_{ae}) is the diameter of a sphere of water (density 10^3 kg m^{-3}) that has the

same falling speed in air as the particle in question (see below).

Of these, perhaps the most important in the occupational hygiene context is the last one – particle aerodynamic diameter – as this governs the airborne behaviour of most particles under most conditions, and so is relevant to the inhalation of particles by humans, deposition in the respiratory tract, sampling and air cleaning.

For some particles, none of the above definitions of particle size is truly appropriate, and further considerations need to be invoked. This is the case for fibres for which both diameter and length need to be defined. Complex aggregates such as those formed during combustion (e.g. smokes) also pose special problems. As already mentioned, these are made up of large numbers of very small primary particles and the degree of complexity is such as to render difficult the definition of size in relation to any of the measurable geometrical properties like those described above. So, although aerodynamic diameter can be usefully applied to describe aerodynamic behaviour, and a geometric diameter can be applied to describe aspects of visual appearance of individual particles or aerosols as a whole, these do not always properly convey the full nature of the particles. For many complex aggregated particles, therefore, the concepts of fractal geometry can provide further information, leading to the concept of a fractal dimension.

Elementary particle size statistics

Only rarely in practical situations – usually under controlled laboratory conditions – do aerosols consist of particles of all one size. Such aerosols are referred to as ‘monodisperse’. More generally, however, in workplaces and elsewhere, aerosols consist of populations of particles having wide ranges of sizes, and so are termed ‘polydisperse’. For these, particle size within an aerosol needs to be thought of in statistical terms.

Consider an ensemble of particles whose sizes can be represented in terms of a single dimension (say, d). The fraction of the total mass of particles with dimension falling within the range d to $d + dd$ may be expressed as:

$$dm = m(d)dd \tag{9.11}$$

where

$$\int_0^{\infty} m(d)dd = 1 \tag{9.12}$$

in which $m(d)$ is the mass frequency distribution function. Alternatively, we have directly analogous expressions for the number frequency distribution function, say $n(d)$.

In particle size statistics it is often helpful to plot distributions in the alternative cumulative form. For example, for the distribution of particle mass this is given in terms of the mass with dimension less than d , thus:

$$C_m(d) = \int_0^d m(d)dd \tag{9.13}$$

where C_m is the cumulative mass distribution. The fraction of mass with dimension less than d is given by:

$$\frac{\int_0^d m(d)dd}{\int_0^{\infty} m(d)dd} \tag{9.14}$$

A typical mass distribution for a workplace aerosol is shown in Fig. 9.3, both in the frequency and cumulative forms. Note here that the cumulative distribution describes the mass (e.g. in units, mg) contained in particles below the stated dimension (where now we have replaced L with d , where d is the particle diameter). As the cumulative distribution is obtained by integrating the frequency distribution, it follows conversely that the frequency distribution derives from differentiating the cumulative distribution. Thus, it is seen that the frequency distribution represents the mass fractions of particles contained within narrow size bands [and so may be expressed, for example, in units of $(\text{mg } \mu\text{m}^{-1})$].

Figure 9.3 contains a number of important features. First, the mass median particle diameter

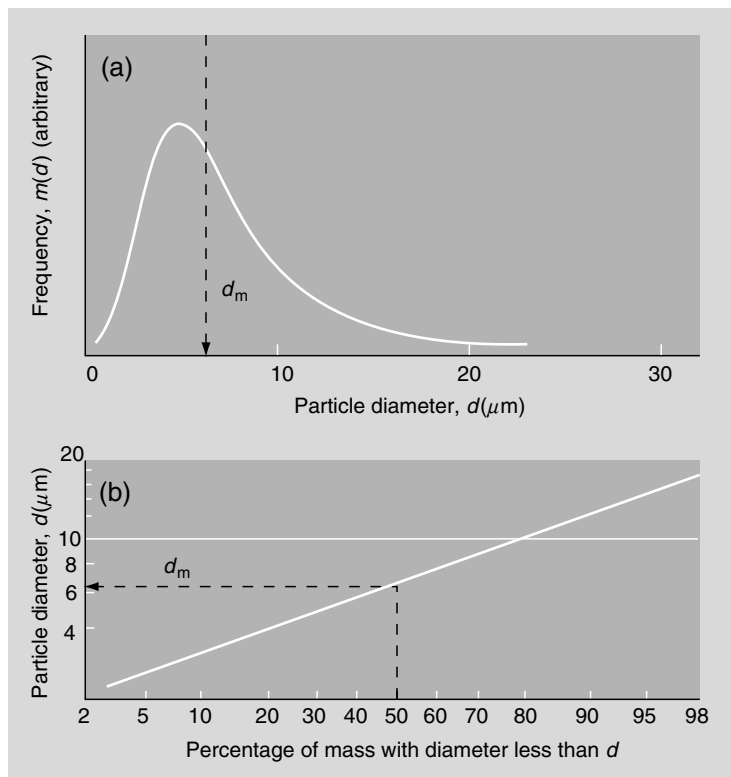


Figure 9.3 Examples of aerosol size distributions. (a) Frequency distribution; (b) cumulative distribution on log-probability axes.

(d_m), at which 50% of the mass is contained within smaller particles and 50% is contained within larger ones, can be read off directly from the cumulative plot. Second, the frequency distribution shown exhibits a strong degree of asymmetry such that the peak lies at a value of d , which is substantially smaller than d_m , and there is a long tail in the distribution that extends out to relatively large particles. This characteristic is very common in polydisperse aerosols that are found in workplace environments. Very often, the overall distribution can be represented to a fair first approximation by the log-normal mathematical function:

$$m(d) = \frac{1}{d\sqrt{2\pi} \ln \sigma_g} \exp \left[-\frac{(\ln d - \ln d_m)^2}{2(\ln \sigma_g)^2} \right] \quad (9.15)$$

where σ_g is the geometric standard deviation, reflecting the width of the distribution. This is given by:

$$\sigma_g = \frac{d_{84\%}}{d_m} = \frac{d_m}{d_{16\%}} \quad (9.16)$$

For a perfectly monodisperse aerosol, $\sigma_g = 1$. More typically for aerosols found in the workplace environment, σ_g ranges from about 2 to 3. At this point, it is useful to note that, when the cumulative distribution is plotted on log-probability axes, it appears as a straight line if the distribution is log-normal (see Fig. 9.3b). Such log-normality (or even a reasonable approximation to it) provides some additional useful aspects. In particular, it enables conversions between relationships for distributions based on particle number, mass, surface area and any other aerosol property, using a set of equations (known as the Hatch–Choate equations) that have the form:

$$qMD = NMD \exp(q \ln^2 \sigma_g) \quad (9.17)$$

where NMD is the number median particle diameter and qMD is the median diameter weighted by dq . For a given particle size d , in order to get from particle number to mass we need to multiply by d^3 . Therefore, it becomes obvious that $q = 3$ if we wish to use Equation 9.17 to convert distributions from number to mass.

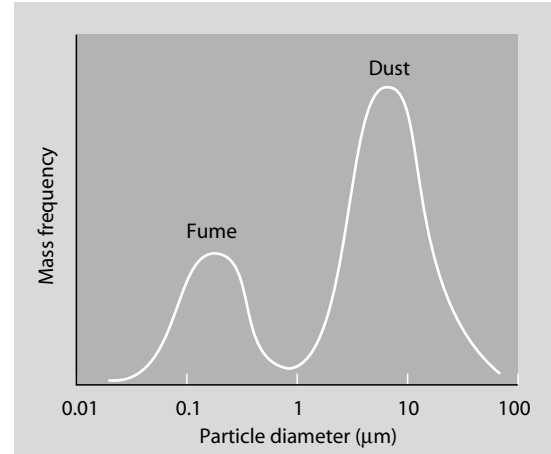


Figure 9.4 Typical bimodal aerosol frequency distribution (from an underground mining situation, showing the dust and diesel fume components).

The appearance of a log-normal particle size distribution is usually associated with a single aerosol generation process. In many workplaces, there may be more than one type of aerosol. In such cases, it is not unusual, for the aerosol as a whole, to find two or more particle size distributions superimposed. These are referred to as multimodal. A typical example is given in Fig. 9.4 for an aerosol in an underground mining situation where there is both relatively coarse dust (generated by the extraction process itself) and relatively fine diesel particulate (generated by underground transportation).

The motion of airborne particles

The physical processes governing the motion of airborne particles are highly relevant to the transport and deposition of particles in ventilation ducts, deposition onto workplace surfaces, inhalation into and deposition inside the human respiratory tract, sampling and filtration, and so on. So, an elementary appreciation of the physics of particle motion is important to occupational hygienists.

Drag force on a particle

When a particle moves relative to the air, it experiences forces associated with the resistance (by the

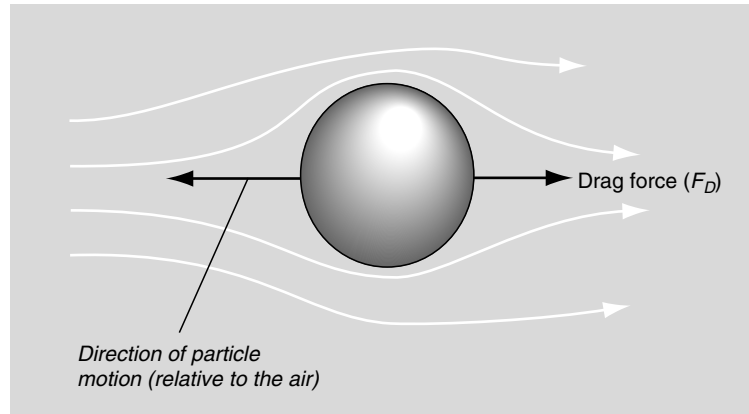


Figure 9.5 Schematic to show the drag force on an aerosol particle.

air) to its relative motion (as shown schematically in Fig. 9.5). For very slow, ‘creeping’ flow (at low Reynolds number) over the particle at velocity v , the drag force (F_D) is given by the well-known Stokes’ law:

$$F_D = -3\pi d\eta v \quad (9.18)$$

where η is the viscosity of air. The Reynolds number for the particle, defined as:

$$Re_p = \frac{dv\rho}{\eta} \quad (9.19)$$

is very small ($Re_p < 1$) and the minus sign indicates that the drag force is acting in the direction opposing the particle’s motion (ρ = density of air). Strictly, this expression should be modified by three factors. The first is the Cunningham correction factor, which derives from the fact that, in reality, the air surrounding the particle is not continuous but is made up of individual gas molecules that are in random thermal motion (so particle motion takes the form of ‘slip’ between collisions with individual air molecules). The second factor concerns deviations from Stokes’ law at Re_p values exceeding 1. The third relates to cases (the majority in practice) when particles are non-spherical. These corrections are described in detail in the aerosol science literature. They should never be ignored. But in many occupational hygiene situations they may be quite small, so that – to a first approximation – Stokes’ law may be a reasonable working assumption.

The starting point for all considerations of particle transport is the general equation of particle motion, again based on Newton’s second law (‘mass \times acceleration = net force acting’). For the forces, the drag force describing the resistance of the fluid to the particle’s motion has already been described. In addition, there may be an external force (e.g. gravity, electrical or some combination of forces), the effect of which is to generate and sustain particle motion. As long as the particle is in motion relative to the fluid, the drag force will remain finite. The proper relationship for describing the particle motion is a *vector equation*, embodying the motion of the air and the particle and the forces acting, each in all three available dimensions. It is not difficult, therefore, to envisage that the resultant set of equations that needs to be solved for particle motion in specific cases can become quite complicated. However, the important principles involved can be illustrated by reference to one simple – but nonetheless extremely important – example.

Motion under the influence of gravity

The case of a particle falling under the influence of gravity in still air is shown schematically in Fig. 9.6. The equation of motion for a spherical stokesian particle (i.e. a particle obeying Stokes’ law) moving in the vertical (y) direction is given by:

$$m(dv_y/dt) = mg - 3\pi\eta dv_y \quad (9.20)$$

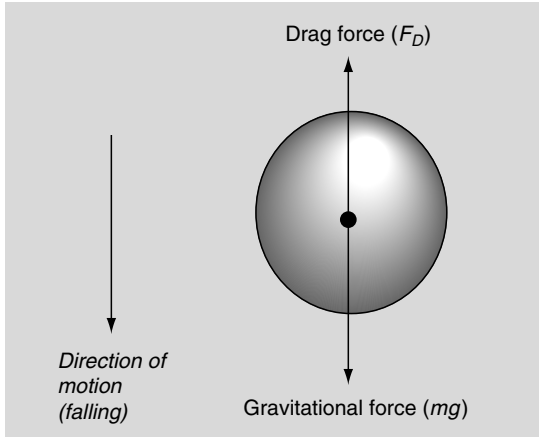


Figure 9.6 Schematic to show the forces acting on a particle of mass m moving under the influence of gravity.

where v_y is the particle's velocity in the y -direction, m is its mass and g the acceleration due to gravity. For a spherical particle, this expression may be reorganized to give:

$$dv_y/dt = (v_y/\tau) - g = 0 \quad (9.21)$$

where

$$\tau = d^2\gamma/18\eta \quad (9.22)$$

in which γ is particle density. In Equation 9.22, closer inspection reveals that τ has dimensions of time, the significance of which we shall see shortly. Equation 9.21 is a simple first-order linear differential equation of the type familiar in many areas of science and engineering. In terms of the particle velocity at time t , it has the well-known form:

$$v_y = g\tau[1 - \exp(-t/\tau)] \quad (9.23)$$

for the case when the particle starts from rest ($v_y = 0$) at time $t = 0$. This shows that particle velocity under the influence of gravity tends exponentially towards a terminal value, the sedimentation or falling speed, given by:

$$v_s = g\tau \quad (9.24)$$

Regardless of particle size (but under the broad simplifying assumption that Stokes' law applies), particle velocity reaches $1/e$ of its final terminal value at $\tau = t$. The quantity t is therefore referred to as the particle relaxation time. Based on the

above equations, we can estimate that for a 'fine' particle with the same density as water (i.e. 103 kg m^{-3}) with $d = 1 \text{ }\mu\text{m}$, we get $v_s \sim 30 \text{ mm s}^{-1}$; for $d = 5 \text{ }\mu\text{m}$ we get $v_s \sim 0.8 \text{ mm s}^{-1}$; and for a 'coarse' particle with $d = 20 \text{ }\mu\text{m}$ we get $v_s \sim 12 \text{ mm s}^{-1}$ and so on. If we wished, at that stage we could estimate the appropriate value of Re for each particle size and so may inspect the extent to which the assumption of stokesian conditions is valid.

From the above, we could perform a simple 'back-of-the-envelope' calculation of the time it would take for particles of given type and size to sediment out completely in a room of given dimensions. For example, consider a cloud of monodisperse water droplets with a diameter of $20 \text{ }\mu\text{m}$ uniformly dispersed into a room of a height of 3 m . Under the simplest assumptions (no air movement or other deposition mechanisms), we may estimate that all particles will have settled to the floor of the room in a time $3 \text{ m} \times 12 \text{ mm s}^{-1}$, i.e. in about 4 h .

Although the mechanism of gravitational settling is perhaps the most important in occupational hygiene, other relevant examples are those involving particle motion in electric fields or in thermal gradients. For these, the general physical approach is directly analogous to that for gravitational settling.

For two spherical particles having different diameters (d_1 and d_2) and different densities (γ_1 and γ_2), their falling speeds in air will be the same provided, from Equations 9.21 and 9.23, that:

$$d_1^2\gamma_1 = d_2^2\gamma_2 \quad (9.25)$$

where for simplicity, slip, Reynolds number and particle shape corrections have been neglected. Equation 9.25 leads directly to a new definition of particle size based on falling speed, namely the particle aerodynamic diameter (d_{ac}) referred to earlier. Thus, for a given near-spherical particle we have:

$$d_{ac} = d(\gamma/\gamma^*)^{1/2} \quad (9.26)$$

where d is the geometric diameter of the particle and γ^* is the density of water (103 kg m^{-3}). Note that this does not apply to particles of extreme aspect ratio, notably long and thin fibres, for which separate equations have been developed and are described in the literature.

Impaction and interception

Consider what happens in a distorted aerosol flow, as for example around a bend in a duct or about a bluff flow obstacle (Fig. 9.7). The air itself diverges to pass around the outside of the body. The flow of airborne ‘inertia-less’ particles would do the same. However, as described above, real particles exhibit the features of inertial behaviour, in particular the tendency to continue to travel in the direction of their original motion upstream of the body. This tendency is greater the more massive the particle, the greater its approach velocity and the more sharply the flow diverges. In the aerosol flow shown in Fig. 9.7, the result is that some particles will ‘impact’ on to the surface of the body. The effect is greatest for the heaviest particles approaching the body at the highest velocity. The efficiency of impaction is:

$$E = \frac{\text{Number of particles arriving by impaction}}{\text{Number of particles geometrically incident on the body}} \quad (9.27)$$

and is a strong function of the Stokes’ number:

$$St = \frac{d^2 \gamma U}{18 \eta D} \quad (9.28)$$

where D is the body dimension and U is the velocity of the approaching airflow. If all the particles that impact onto the body in the manner indicated actually stick and so are removed from the flow,

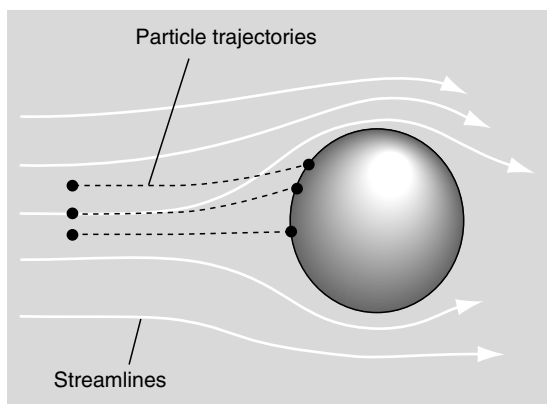


Figure 9.7 Schematic to illustrate the phenomenon of impaction.

then E is also equivalent to the collection efficiency. Therefore, it is seen that impaction is important in aerosol collection in many situations, including during filtration and aerosol sampling.

This discussion can be extended to a particle whose trajectory, as traced by the motion of the particle’s centre of gravity, passes by outside the body. If this trajectory passes close enough to the surface of the body and if the particle is geometrically large enough, it may be collected by interception, as illustrated in Fig. 9.8. Although for $d \ll D$ this effect on E is negligible, it becomes a significant influence if d becomes of the order of D , as for example it might in a filtration device made up of thin fibrous collecting elements.

Elutriation

The general term ‘elutriation’ is used to refer to another mode of particle deposition relevant to industrial hygiene – from a moving air stream under the influence of an externally applied force. Traditionally, the term has been used to describe the gravitational separation of particles carried along by smooth laminar flow through a narrow horizontal channel in which particles are deposited on to the floor of the channel. An extension of this idea is the gravitational elutriation that occurs during aerosol flow vertically upwards (e.g. through a vertical tube or into an inverted sampling device). The general principle also applies if some other force (e.g. electrostatic) is the main agency of deposition. The process is relevant to aerosol behaviour not only in sampling devices but also in the airways of the lung after inhalation.

Aspiration

Aspiration concerns the process by which particles are withdrawn from ambient air through an opening in an otherwise enclosed body. It is therefore relevant to aerosol sampling systems. It is also relevant to the inhalation of aerosols by humans through the nose and/or mouth during breathing.

In order to identify the nature of the process of aspiration and to enable some generalizations, Fig. 9.9 shows schematically a body of arbitrary shape placed in a moving airstream. It has a single

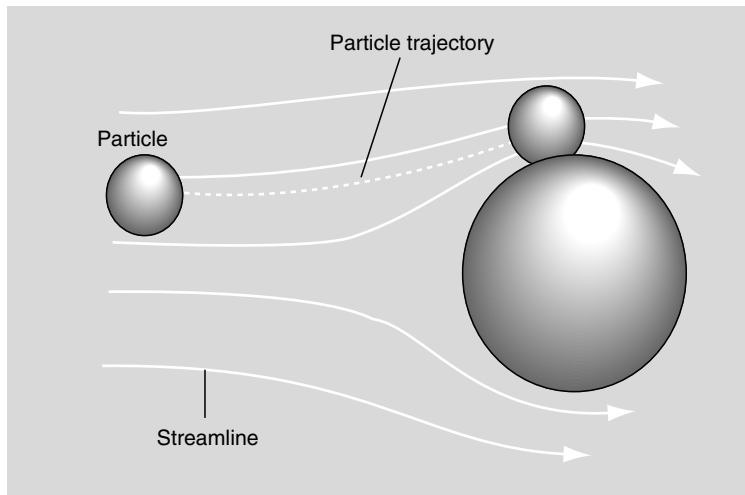


Figure 9.8 Schematic to illustrate the phenomenon of interception.

orifice located at arbitrary orientation with respect to the wind through which air is drawn at a fixed volumetric flow. There are two competing flow influences on particle transport – the external wind, which diverges to pass around the outside of the body, and the convergent flow into the orifice. The interaction between these two gives rise to the complex distorted overall flow pattern that is shown. It may be thought of as having two parts: the external divergent part and the internal convergent part.

Particles moving in this flow system respond to the changes in velocity and direction in the ways described earlier. Generally, in moving air the wind brings particles into the region of influence of the

aspirating body and inertial forces provide the dominant influence on aerosol transport in that region. In fact, the system shown may be regarded as just a more complicated version of the impaction of particles on to a bluff body. This time, however, particles may be thought of as having to undergo two successive impaction processes. The first involves particles impacting on to the surface of the body and is governed by the external part of the flow. The second involves the impaction of particles in the plane of the orifice and is governed by the internal part of the flow. Having established this picture, we may begin to construct a quantitative physical model for the efficiency with which particles are aspirated from the ambient air and into the body through the orifice.

Aspiration efficiency (A) may be defined for given particle aerodynamic diameter (d_{ac}), body and orifice geometry and dimensions (D and d respectively), orientation with respect to the wind direction (q), external wind speed (u) and mean aspiration velocity (u_s) as:

$$A = \frac{\text{Concentration of particles in the air actually entering the orifice}}{\text{Concentration of particles in the undisturbed upstream air}} \quad (9.29)$$

provided that the airflow and aerosol upstream of the sampler are uniformly distributed in space. Aspiration efficiency defined in this way is the

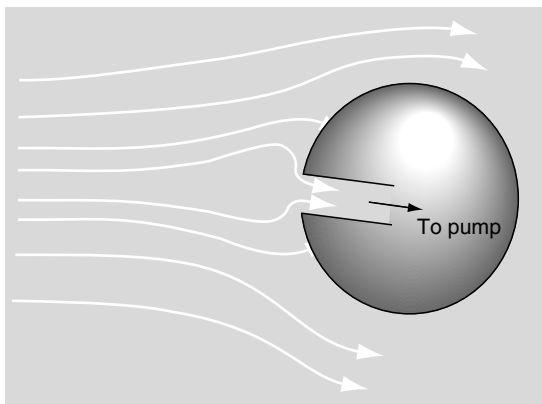


Figure 9.9 Schematic to illustrate the concept of aspiration.

most basic description of performance for an aerosol aspirating system (such as an aerosol sampler). Starting with Equation 9.27, and from considerations of particle impaction from one region of the flow to another, a system of equations may be developed which can, in principle, provide estimates for A . For the present purposes, it is sufficient to express some of the generalizations that arise. In the first instance:

$$A = f(\text{St}, U/U_s, \delta/D, \theta, B) \quad (9.30)$$

where $\text{St} (= d_{ac} 2^* U / 18 \eta D)$ is a characteristic Stokes' number for the aspiration system and B is an aerodynamic shape ('bluffness' or 'bluntness') factor. Second,

$$A \rightarrow (U/U_s) \cos \theta \text{ as } \text{St} \rightarrow \infty \quad (9.31)$$

indicating that A levels off for large particles approaching the body at high wind speed. For very large particles and/or in environments with very little air movement, gravity may also play a role, and so an additional term – to reflect the effect of gravitational settling – may be required in Equation 9.31.

This forms the basis for understanding the performance characteristics of the simplest – and most widely researched – sampling system, the thin-walled tube. For many years, this has formed the basis of aerosol sampling in stacks and ducts under what have come to be known as *isokinetic sampling conditions*. Here, with the thin-walled sampling tube aligned axially with the flow and the sampling flow rate adjusted so that the velocity of the air entering the tube matches that in the duct (in the absence of the sampler), there is no distortion of the airstream and so particles of all sizes are aspirated with 100% efficiency (Fig. 9.10).

Diffusion

Particle motion has so far been assumed to be well ordered and – in theory at least – deterministic. In reality, however, even in apparently smooth airflow, aerosol particles exhibit random movement associated with their collisions with gas molecules, which themselves are in thermal motion (as described by the classical kinetic theory of gases). Such movement is independent of any convection associated

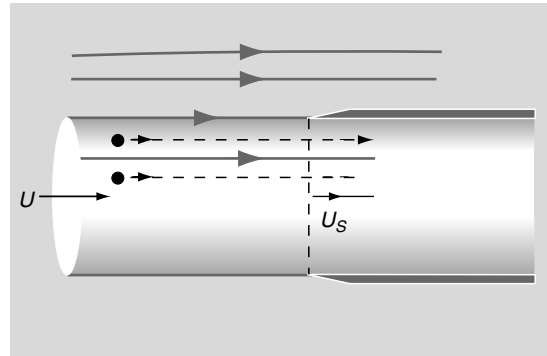


Figure 9.10 Schematic to illustrate isokinetic sampling with a thin-walled sampling tube, in which sampling velocity (U_s) is matched to the wind speed (U).

with the air itself, and is known as molecular (or Brownian) diffusion. As a result of this phenomenon, there is a net migration of particles from regions of high concentration to regions of low concentration. That is, although individual particles may diffuse in either direction, a greater number end up travelling down the concentration gradient. The resultant local net flux of particles by this process is described by the well-known Fick's law of classical diffusion, which is described for the simple one-dimensional case (for the x -direction) by:

$$\text{Local net flux} = -D_B \frac{dc}{dx} \quad (9.32)$$

where c is the local concentration and D_B is the coefficient of Brownian diffusion. From classical kinetic theory for a small particle in the Stokes' regime, the latter is given by:

$$D_B = \frac{kT}{3\pi\eta d_v} \quad (9.33)$$

where T is the air temperature (in K) and k is the Boltzmann constant ($= 1.38 \times 10^{-23} \text{ J K}^{-1}$). Here, the numerator represents the thermal energy of the gas molecules that is being transferred to the particles and the denominator represents the loss of particle energy due to viscous effects. Therefore, D_B embodies the continual interchange of thermal energy between the gas molecules and particles, and vice versa. Typically, for a particle of a diameter of $1 \mu\text{m}$ in air, D_B is very small – only of the order of $10^{-11} \text{ m}^2 \text{ s}^{-1}$.

Equation 9.32 leads directly to the general diffusion equation describing the local rate of change of concentration:

$$\frac{dc}{dt} = D_B \frac{d^2c}{dx^2} \quad (9.34)$$

whose solution for the simple one-dimensional case of N_0 particles released initially at $x = 0$ at time $t = 0$ gives the Gaussian form:

$$c(x, t) = \frac{N_0}{(2\pi D_B t)^{1/2}} \exp\left(\frac{-x^2}{4D_B t}\right) \quad (9.35)$$

for the concentration distribution along the x -direction at time t . The root mean square displacement of particles (in the one-dimensional case chosen) from their origin at time t is:

$$x' = (2D_B t)^{1/2} \quad (9.36)$$

Aerosol diffusion in a flowing gas system is referred to as 'convective diffusion', and this is perhaps the aspect that is most relevant to occupational hygiene, especially with respect to deposition. In simple terms, this may be envisaged by superimposing the possible excursion due to diffusion on the trajectories that would otherwise result in the absence of diffusion. The scaling parameter for this situation, analogous to the Stokes' number already described for inertial behaviour, is the Peclet number (Pe), given by:

$$Pe = \frac{UD}{D_B} \quad (9.37)$$

where, as before, D and U are dimensional and velocity scales respectively. The smaller Pe , the more pronounced the contribution due to diffusion.

The phenomenon of diffusion is important not only in how particles move from one point in an aerosol system to another but also how they move in relation to one another.

Interactions of airborne pollutants with electromagnetic radiation

Aerosols

Whereas most of the properties of aerosols outlined above can be directly linked – in one way or

another – with health effects or environmental control, optical properties may appear to be somewhat peripheral. However, there are two aspects that are particularly relevant to occupational hygiene. The first concerns the visual appearance of a workplace aerosol. The fact that it is visible at all is usually an indication that worker exposure is high enough to demand attention. Furthermore, its visible intensity is a direct indication of the level of exposure. In addition, other qualitative features of the aerosol's appearance (e.g. colour) can provide some information about its physical nature. From such considerations, therefore, an experienced and enlightened occupational hygienist can learn a great deal from the visual appearance of a workplace aerosol. At the more quantitative level, however, the optical properties of aerosols can form the basis of sophisticated aerosol instrumentation for measuring not only aerosol concentration but also particle size characteristics.

The basic physical problem involved in the optical properties of aerosols concerns the interaction of electromagnetic radiation with individual suspended particles and with ensembles of such particles. If a particle has different dielectric properties to those of the surrounding medium, as reflected in their refractive indices, then it represents a dielectric non-homogeneity. As a result, interactions with incident light can be detected from outside. In general, the whole problem can be treated in terms of a plane electromagnetic wave incident on a particle whose geometric surface defines the boundaries of the non-homogeneity, and whose dielectric properties are described by the refractive index for the particle medium. Mathematically, it is based on Maxwell's theory of electromagnetic radiation, the solutions of which explain the well-known phenomena of reflection, diffraction, refraction and absorption. The first three of these constitute the phenomenon of light scattering; the last concerns that part of the incident energy that goes into increasing the vibrational energy of the molecules in the ordered array inside the solid particle. Such absorbed energy appears in the form of heat, raising the temperature of the particle.

There is one further process that deserves mention, namely the physical mechanism by which

radiation incident at one wavelength can be scattered at another. This occurs by virtue of so-called ‘inelastic’ interactions involving the absorption and re-emission of radiation energy by the individual molecules of the particle. However, such interactions do not have much direct relevance to workplace aerosols. So attention here will be focused on the simpler cases where the wavelengths of the incident and scattered radiation are the same. Such interactions are referred to as ‘elastic’.

The first theory of light scattering was by Lord Rayleigh in the late 1800s, and applies to very small particles and molecules much less than the wavelength of the radiation. In effect, for visible light, this means particles with diameter of less than about $0.05\ \mu\text{m}$. Under these conditions, the particles may be treated as ‘point scatterers’, and the resultant mathematical treatment is relatively simple. But the most significant advance, in terms of its relevance to aerosols, came in the early 1900s when Mie extended Rayleigh’s theory to larger particles.

For a beam of light energy incident on a system of many suspended particles (e.g. an aerosol), the fraction of energy that interacts in the manner indicated is either scattered or absorbed. This energy is effectively removed so that the beam itself may be regarded as having been attenuated or undergone extinction. The energy that remains in the beam is transmitted. From this picture, the interaction of light with an aerosol may be considered in one of two ways: either in terms of the extinction of the beam (or, conversely, its transmittance) or in terms of the scattered component.

The phenomenon of extinction is described by an important relation, the well-known Lambert–Beer law, which appears widely in science for describing the effects of the interactions between energy (of all types) and matter. For the passage of light through an aerosol, it is written in the form:

$$\frac{I}{I_0} = \exp(-\alpha ct) \quad (9.38)$$

where I_0 and I are the light intensities before and after passing through the aerosol respectively, c is

the aerosol concentration and t is the path length through the aerosol. The important quantity is an extinction coefficient that embodies the physics of the interactions between the light and each individual particle.

Both light extinction and light scattering are relevant to occupational hygiene, in relation both to the visual appearance of aerosols and to aerosol monitoring instrumentation. Possible scenarios are summarized in Fig. 9.11.

Gases

When electromagnetic radiation passes through a gaseous medium, energy may be removed from the beam if the wavelength of the radiation is such that energy may be absorbed by the molecules of the gas. So the phenomenon of extinction again applies, and the Lambert–Beer law reappears, this time in the form:

$$\frac{I}{I_0} = \exp(-\alpha_\lambda ct) \quad (9.39)$$

where c is the concentration of the molecules with which the radiation is interacting, and I , I_0 and t are as defined above. In Equation 9.39 there is an extinction coefficient embodying the physics of the interaction between the radiation and the gas through which it passes. It refers to the absorption spectrum of the gas and is strongly dependent on the wavelength (λ). In the ultraviolet region from 0.25 to $0.40\ \mu\text{m}$, absorption takes place by electronic transitions in the gas molecules (e.g. excitation or ionization). In the visible region from 0.40 to $0.70\ \mu\text{m}$, absorption is by vibrational–rotational excitation, although this is very weak in most gases (hence their invisibility to the human eye). The only common gas for which there is significant absorption in the visible region is nitrous oxide, which occurs as a visible brown gas. In the infrared region above $0.70\ \mu\text{m}$, there are strong vibrational–rotational modes of excitation for most gases and vapours. This region is therefore particularly useful for application in detection systems for pollutant gases and, indeed, is employed widely for instruments used in the occupational hygiene setting.

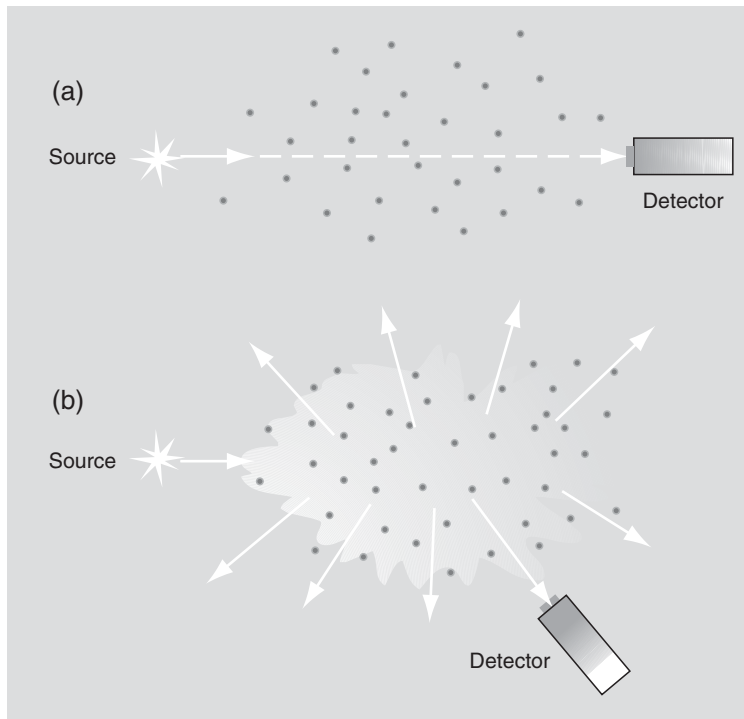


Figure 9.11 Examples of practical scenarios involving the interaction of light with particles. (a) Detection of transmitted light; (b) detection of scattered light.

Summary

This chapter has given short descriptions of various physical aspects of gases and aerosols relevant to the science and practice of occupational hygiene. These relate to the properties of airborne contaminant materials as they influence worker exposure, in particular their recognition, evaluation and control. This review underlines the point that occupational hygiene is first and foremost a scientific discipline. More in-depth background in any particular aspect described may be obtained by reference to the reading list given below.

Further reading

- Cohen, B. and Hering, S. (eds) (2001). *Air Sampling Instruments for Evaluation of Atmospheric Contaminants*, 8th edn. ACGIH, Cincinnati, OH.
- Hinds, W.C. (1999). *Aerosol Technology*. John Wiley & Sons, New York.
- Ness, S.A. (1991). *Air Monitoring for Toxic Exposures, An Integrated Approach*. Van Nostrand, New York.
- Vincent, J.H. (1995). *Aerosol Science for Industrial Hygienists*. Elsevier Science, New York.

Chapter 10

Principles of risk assessment

Steven S. Sadhra

Introduction	Exposure assessment
Hazards, risks and risk assessment	Selection of sampling and analysis methods
Risk assessment and legislation	Exposure variation
Risk rating	Exposure modelling
Individual and societal risk	Risk characterization
Risk assessment and occupational exposure limits	Acceptability and tolerability of risks
Cost–benefit analysis	Comparison of exposure data with exposure limits
Communicating risks	Recording the risk assessment
A model for the assessment and management of occupational health risks	Conclusions
Risk assessment in practice	References
Hazard identification techniques	
Assessment of dose–response	

Introduction

Risk assessment is a structured and systematic procedure, which is dependent upon the correct identification of the hazards and an appropriate estimation of the risks arising from them, with a view to making inter-risk comparisons for purposes of their control or avoidance (Health and Safety Executive, HSE, 1995). Risk assessment aims to improve the quality of the decision-making process and attempts to reduce the uncertainty as much as possible. However, it must be realized that uncertainty does not disappear just because the risks are being assessed. Evaluation of risk is complex, especially when deciding whether the risks to the individuals or the public are acceptable.

In occupational health terms, the purpose of risk assessment is to enable a valid decision to be made about measures necessary to control exposure to hazards arising in any workplace. It enables the employers to demonstrate that all factors pertinent to the work have been considered, and that a valid judgement has been reached about the risks. An important part of the assessment process is to

define clearly the steps that need to be taken to achieve and maintain adequate control. In defining adequate control, one must decide on the acceptability of risks, which will depend on factors such as legal requirements, costs, availability of controls, toxicity of substances and the number of individuals exposed. The effort, detail and expertise required in conducting an assessment will depend largely on the nature and degree of risk as well as the complexity and variability of the process. For a risk assessment to be meaningful its outcomes must trigger actions to manage the defined risks; hence risk assessment and risk management are two inter-related processes. It is commonly believed that risk assessment is based on scientific principles alone, whereas risk management also takes into consideration issues such as technological feasibility, cost–benefit, public perception and government policy (Sadhra and Rampal, 1999).

Although there are inter-country and inter-agency differences in the methodology used in conducting risk assessments, attempts continue to be made to standardize the approaches in risk assessment

methodology, e.g. in the UK the Interdepartmental Liaison Group on Risk Assessment reviewed the use of risk assessment within government departments (HSE, 1996). A general guide on risk assessment for the European Union member states has also been published providing practical guidance on implementing the requirements of the Council framework Directive 89/391/EC (European Commission, 1996).

Hazards, risks and risk assessment

A *hazard* is a substance, agent or physical situation with a potential for harm in terms of injury or ill health, damage to property, damage to the environment or a combination of these. Hazards can be physical, chemical, biological, ergonomic (including mechanical) and psychosocial. *Hazard identification*, the first step in the risk assessment, is purely qualitative and is defined as the process of recognizing that a hazard exists and defining its characteristics.

Risk is the likelihood of the harm or undesired event occurring and the consequences of its occurrence. It is the probability that the substance or agent will cause adverse effects under the conditions of use and/or exposure and the possible extent of harm. Hence it is a function of both exposure to the hazard and the likelihood of harm from the hazard. The *extent of risk* covers the population that might be affected by a risk, i.e. the numbers of people who may be exposed and the consequences from them. *Risk assessment* is the overall process of estimating the magnitude of risk and deciding whether or not the risk is tolerable or acceptable, taking into account any measures already in place.

Risk assessment and legislation

Legislation has been the main driving force behind formal risk assessments. Earlier legislation tended to be prescriptive, laying down specific sets of rules to follow, and has often been reactive following major incidents. For example, the Offshore Instal-

lation (Safety Case) Regulations, 1992, was introduced after Lord Cullen's Report on the Public Inquiry into the *Piper Alpha* tragedy, which killed 167 people.

In the UK, an important move towards proactive health and safety management was made with the introduction of the Health and Safety at Work, etc. Act, 1974 (HSWA). The HSWA contains an implied duty to carry out risk assessment by virtue of the phrase 'so far as is reasonably practicable'. The requirement for assessments of risk and proposal for ensuring safety (safety case reports) in some high-risk situations, such as major hazards sites and offshore oil and gas operations, are covered by the Control of Major Accident Hazard (COMAH) Regulations 1999. However, the UK legislation which probably made the greatest impact in terms of coverage and specific duties to undertake risk assessment was the Control of Substances Hazardous to Health Regulations 1989 (COSHH) (now replaced by the 1999 COSHH Regulations). Examples of other UK legislation that has included definite requirements for risk assessments to be carried out include: Control of Asbestos at Work Regulations, 1985; Ionising Radiation Regulations, 1985; Noise at Work Regulations, 1989.

More recently, the concept of risk assessment has been introduced into a set of European health and safety directives, the requirements of which were implemented in the UK in the so-called 'six pack' of regulations. The specific requirements of these regulations have been brought together under the umbrella regulation 'implemented in the UK as the Management of Health at Work Regulations 1992 which require every employer to carry out "a suitable and sufficient assessment of risk" to the health and safety of employees and to anyone else at work who may be affected by the work activity'.

Paustenbach (1995) highlights the historical evolution of risk assessment in the USA and how improvements made in this process have been the basis of both the environmental and occupational health regulations. The Occupational Health and Safety Administration (OSHA) Act in the USA does not mention risk assessment *per se*, but focuses on individual risk to employees exposed

to agents at the permissible exposure limits (PELs) for a working lifetime. The workplace standards in the case of carcinogens were set as low as was deemed to be technically feasible and at reasonable cost. The ‘benzene decision’ by the Supreme Court in 1980 ruled that before OSHA issues a standard it must first demonstrate that the chemical poses a ‘significant risk’. OSHA’s attempt to set standards for 426 chemicals, based on outside standards or setting standards based on general risk assessments instead of case-by-case demonstration of significant risks was also struck down by the courts in 1992.

In Europe the general principles of risk assessment of new and existing substances are laid down in Directive 93/67/EEC and Regulation 1499/94. Technical guidance documents have been written, which provide a detailed framework for conducting human health risk assessment. The risk assessment process entails assessment of effects and exposure, which are then integrated to characterize human health risk. The risk assessments are carried out by competent authorities designated by the responsible Member States to act as rapporteurs. The proposed risk assessment scheme is intended to integrate occupational, indirect air and consumer exposure to a single substance rather than being industry based. Hence additive or synergistic effects, which may be caused by a combined action of several substances, are not considered. The risk assessment procedure covers the whole life cycle of the substance in all environmental compartments. The Council Regulation (EEC) No. 9793/93 on the evaluation and control of existing substances requires under article 10 the actual or potential risk to man of priority substances to be assessed using principles which have been laid down in the Commission Regulation (EC) No. 1488/94 on risk assessment for existing substances. The risk assessment process for human health entails the following steps:

1 Assessment of effects, comprising:

- (a) hazard identification: identification of the adverse effects that a substance has an inherent capacity to cause; and
- (b) dose (concentration–response (effects) assessment): estimation of the relationship be-

tween dose, or level of exposure to a substance, and the incidence and severity of an effect, where appropriate.

2 Exposure assessment. Estimation of the concentrations/doses to which human populations (i.e. workers, consumers and man exposed indirectly via the environment) are, or may be, exposed.

3 Risk characterization. Estimation of the incidence and the severity of the adverse effects likely to occur in a human population due to actual or predicted exposure to a substance, and may include ‘risk estimation’, i.e. quantification of that likelihood.

This Commission Directive requires that the risk assessment should address a number of defined potential health effects and human populations considering each population’s exposure by the inhalation, dermal and oral route.

Health effects

- Acute toxicity.
- Irritation.
- Corrosiveness.
- Sensitization.
- Repeat-dose toxicity.
- Mutagenicity.
- Carcinogenicity.
- Toxicity for reproduction.

Human populations

- Workers.
- Consumers.
- Humans exposed indirectly via the environment.

In essence, the risk assessment procedure involves comparing the exposure level(s) to which the population is exposed, or likely to be exposed, with the exposure levels at which no toxic effects are expected to occur. The risk assessment is conducted by comparing the exposure level, the outcome of the exposure assessment, with the no observed effect level (NOAEL), the outcome of the dose–response assessment. Depending on the exposure level–NOAEL ratio, the decision is taken as to whether a substance presents a risk to human

health. In cases when it is not possible to establish a NOAEL but a lowest observed adverse effect level (LOAEL) can be derived, the latter is compared with the exposure level. If it is not possible to identify N(L)OAEL, a qualitative evaluation of the likelihood that an adverse effect may occur is carried out. In the risk characterization steps, exposures from different compartments and routes are combined for each potential health effect using the following categories:

- 1 There is need for further information and/or testing.
- 2 There is at present no need for further information and/or testing and no need for risk reduction measures beyond those that are being applied already.
- 3 There is a need for limiting the risks; risk reduction measures that are already being applied shall be taken in to account.

Risk rating

Risks may be defined in qualitative, semiquantitative or quantitative terms:

- *qualitative*: no figures; judgement is used to estimate the risk level;
- *semiquantitative*: risks may be ranked on a comparative scale or by using a risk matrix;
- *quantitative*: risks may be described as a frequency of death.

When risks in the workplace or those faced by society need to be prioritized, attempts are made

to rate or rank them. Although questions continue to be raised about the methodology involved in rating risks, it has continued to dominate risk assessments carried out by government agencies and safety and health professionals in industry. Although there are numerous methods to estimate risk, a common method that continues to be used frequently uses a risk rating derived from a matrix based on rating of the hazards in the workplace and rating of the likelihood of exposure (Fig. 10.1).

Risk rating is dependent on both the hazard rating and the likelihood rating with the equation used to determine the risk rating being: $\text{risk rating} = \text{hazard rating} \times \text{likelihood rating}$. The hazard rating is based on the severity of harm and damage that can occur. The severity categories can range from near miss/minor consequence to catastrophic consequences.

Acceptability of the consequences can also be based on the frequency or likelihood of the event. Consequences that are major or catastrophic are only acceptable when the likelihood or frequency of the event occurring is very small. The likelihood of the untoward events can be assessed using historical information or special techniques like *fault tree analysis* and *event tree analysis*.

The American Institute of Chemical Engineers has developed semiquantitative guidelines for the ranking of risks where class 1 risks are those considered to be of sufficient significance to warrant immediate shutdown until the hazard is mitigated. Under these guidelines, class II risks are those requiring immediate action to mitigate the risk,

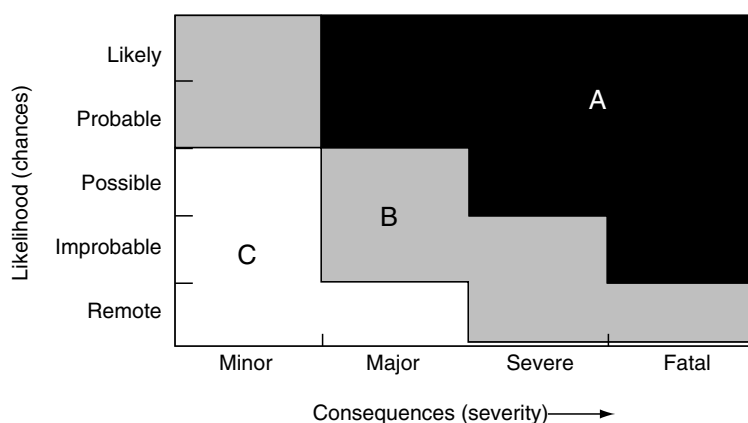


Figure 10.1 A risk matrix (Sadhra and Rampal, 1999).

and a programme to provide a permanent solution should be initiated immediately; class III risks are those of a less serious nature and where the situation is to be corrected as soon as possible. These risks are either related to the specific processes or the management systems. Class IV risks are other areas where risk reduction or improvement in risk management is advised (Wells, 1996).

Individual and societal risk

The most common question that is posed to risk assessors by the public and decision-makers is what is the risk to individuals and what is the risk to the society at large.

Individual risk is the risk of the agent harming a hypothetical person assumed to have representative characteristics of the population. It is the risk to a particular individual and is the frequency at which that individual is expected to sustain a given level of harm from the hazard. An estimate of individual risk will depend on the frequency of harm or undesired event, proportion of time an individual is exposed to the hazard and the vulnerability of the individual. The estimate of individual risk of death from a particular cause is the chance in a million that a person will die from that cause in any one year, averaged over a whole lifetime. Risk estimates are made of those exposed to the risk.

Mathematically, individual risk can be calculated using the following formula: individual risk = $F \times P_1 \times P_2$, where F = frequency of the undesired event, P_1 = probability of the person being killed and P_2 = probability of the person being exposed to the hazard.

Societal risk is the likelihood of the untoward events occurring in a group of people. The societal risk is calculated as for individual risk including the number of individuals in the group. The larger the group exposed the greater the societal risk.

There are various kinds of societal risk, e.g. national, local or risks linking several communities. Societal risks are normally assessed when assessing risk from major industrial hazards, which have the potential of causing large-scale injury and loss of life from a single event. In the context of these major industrial hazards, the *FN* curve is

commonly used. The *FN* curve displays the experienced or predicted frequency (F) of an event killing more than a certain number of people (N). Quantitative risk assessments carried out to estimate risk from these major hazards are only able to make an estimate of real risk within an order of magnitude.

Risk assessment and occupational exposure limits

One of the most common methods that have been used to manage risks at a societal level by regulatory agencies is by establishing occupational exposure limits (OELs). Industry is then expected to comply with these standards. Although industry is expected to self-regulate, enforcement is crucial. There are inter-country differences in the terminology used for the exposure levels established (permissible exposure levels, maximum allowable concentrations, etc.) and also what they mean. The setting of occupational exposure limits usually involves consultation with the interested parties. The degree of consultation varies from being an integral part, as in the rule-making procedure in the USA and in Europe, to one where it is policy maker-driven with little consultation with the stakeholders. In the developed countries, an opportunity is provided for consideration of scientific, technical and socio-economic factors in setting OELs. However, with more than 100 000 existing chemicals in the European Inventory of Existing Commercial Chemical Substances (EINECS) it was obvious to the Commission of the European Community that rapid adoption of Community Occupational Exposure Levels would not be easy. The steps for establishing standards under the Commission of the European Community on Health and Safety (1993) include preparing a scientific dossier, scientific review, recommendation to the Commission on a scientifically based occupational exposure level, evaluation of technical and policy aspects and proposal by the Commission on an occupational exposure level. Then there is consultation with government authorities, the Advisory Committee and other interest groups on the Commission's proposal and finally adoption of the directive.

When it comes to standard setting, three broad approaches have emerged to derive numerical values. One favours the use of mathematical modelling based on quantitative linear extrapolation and makes many assumptions; the second tries to use a NOAEL approach but then uses safety factors of perhaps 100 or 1000; the third approach uses 'lowest technically achievable approach, but takes into account the presumed risk levels' (Vainio and Tomatis, 1985). Probably the most abundant and widely known are the airborne standards, called OELs; however, it should be borne in mind that for a number of chemicals that can be absorbed through skin, relying on an airborne standard may underestimate the total uptake. In this case, biological monitoring and biological effect monitoring can give a truer representation of the risk. For such substances, biological exposure indices (BEIs) [American Conference of Governmental Industrial Hygienists (ACGIH, 2003)] have evolved in the USA and biological tolerance values (BATs) (DFG, 1994) in Germany. Although these two countries are most advanced in setting such biological limits, other countries are now beginning to set their own, such as the UK, where biological monitoring guidance values (BMGVs) (HSE, 2002) have been set but with no regulatory status. In the case of physical hazards, a number of countries have developed standards to protect health, e.g. radiation (lasers, microwaves and ionizing radiation), noise, vibration, heat or for comfort, e.g. workplace illumination, ventilation rates in offices, theatres and kitchens. Most standards are based on protecting people from both short- and long-term exposure and the working schedule.

The application of a standard in the risk assessment process relies on good professional judgement. All standards will have limitations in their settings, either through incomplete data or socio-economic compromise, resulting in some people being at risk. For this reason, the limitations of the standard must be appreciated before comparing limits with exposure levels to estimate risks. For example, most OELs are set for single substances; however, in the workplace individuals are commonly exposed to a cocktail of mixtures, often of varying proportions. Several different

types of interactions may occur between the chemical constituents present in a mixture. Chemicals may act independently of each other, or may interact in an additive manner or synergistically, i.e. the overall effect on health is greater than the sum of the individual effects. However, the use of OELs can assist in justifying the selection and use of control measures for minimizing exposure to hazards in the workplace.

Cost-benefit analysis

Regulators conduct cost-benefit analysis before making new regulations. In the UK, this comes from the principle of 'reasonable practicability' under the Health and Safety at Work Act etc., 1974. The Health and Safety Commission has required cost-benefit assessments to be carried out for all proposals of health and safety regulations and Approved Codes of Practice since 1982. Cost-benefit assessment for regulators assists in deciding which is the next chemical or agent to target, and in determining more effective ways to manage risks (Lopez, 1996). In the UK, cost-benefit assessments (CBA) have been conducted for each individual maximum exposure limit (MEL) proposal. In a CBA, cost and benefits of the proposed level are measured in monetary terms. The aim of the CBA is to ensure that the proposals are worthwhile and consistent across the various industry sectors. The key issues involved are determining the magnitude of the problem, quantifying the costs and benefits and putting a value to the benefits. At the plant level, CBA will involve determining the benefits obtained by ensuring risks are controlled. These gains could be health benefits, reduction in injury and ill health, increased productivity and decreased wastage. The quantum of risk is weighed against the sacrifice involved in averting risk. This analysis requires a financial value to be placed on preventing death, injury, pain and suffering. There is always uncertainty in this exercise but it is required, however, to err on the side of safety when calculating cost and benefits. Benefits from activities entailing risks not accrued by the individual or a particular community are given less weight.

Communicating risks

Communicating information on risks to the stakeholders is imperative. The estimate of the seriousness of the risk is often influenced by personal benefit or loss, information available to them on the risk and familiarity with and understanding of the risks. Their perception may not be based on assessments by experts and in some cases they may even query the opinions of the experts on the real risks. Usually, it is also based on how the harm can affect what they value. The real risks and the perceived risk may be markedly different.

Communication on risk should aim at bridging the gap between perceived risks and real risks. This should lead to action being taken, demonstrating transparency in the decision-making process and also management commitment in maintaining a safe and healthy workplace. Issues that need to be considered when communicating information to stakeholders include a clear objective of risk com-

munication, language and literacy factors and quality and quantity of the information available. It is important to inform those who are most affected and to make sure they understand what is being told to them. When communicating with stakeholders it is useful to show concern, involve them in the process and to acknowledge uncertainty when it exists. ‘Take the jargon out of risk assessment and demystify risk assessment and make it more accessible to the public and decision makers’ was a call made by David Eves of the HSE in 1992.

A model for the assessment and management of occupational health risks

A number of models exist for assessing human health risks (Covello and Merkhofer, 1993). In the UK, Sadhra and Rampal (1999) proposed a model for assessing and managing health risk in the workplace (Fig. 10.2). The model integrates

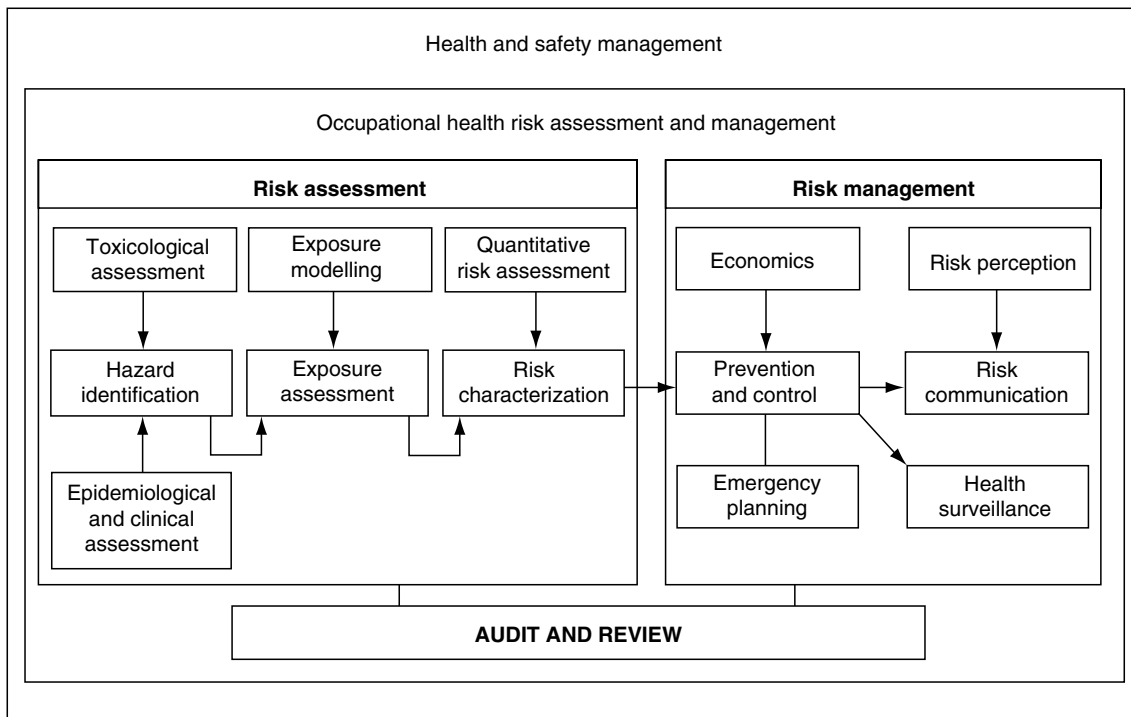


Figure 10.2 Model for risk assessment and management (Sadhra and Rampal, 1999).

various scientific, economic and psychosocial tools that play an important part in defining the extent of risk and decision-making with regard to their communication and management. The inputs to hazard identification include data from toxicological, epidemiological and clinical assessments. Hazard identification leads to the assessment of exposure, which can be measured directly or, where no actual exposure data are available, it may be necessary to rely upon theoretical predictions to model exposure. An important element of data collection is to ensure that suitable sampling strategies are developed and implemented so as to maximize the validity of the data generated, thereby strengthening the inference drawn. In the risk characterization steps, appropriate methods are used to rank and prioritize risks so that suitable actions can be justified. In the management of risks prevention and control measures take into consideration the costs of the action and the benefits to be derived. How the measures taken to manage health risks improve the overall performance of the company including financial performance need to be quantified with well-defined performance indicators. The risks identified need to be communicated to employees and stakeholders as well the steps to be taken to manage these risks. In order to communicate risks effectively we need to understand how individuals perceive risks and the basis for these perceptions. Health surveillance measures are instituted when measures to prevent and control risks are associated with residual risks that are not acceptable. Health surveillance is also the litmus test of the effectiveness of control measures. The collective measures taken to manage risk need to be continuously reviewed through a system of inspections and audits.

Risk assessment in practice

In practice, risk assessment effectively involves four key steps: hazard identification (determining the presence and quantity of contaminants that affect human health); dose–response assessment (the relationship between concentration of the contaminants and incidence of adverse health outcome); exposure assessment (determining condi-

tions of exposure and doses received by those exposed); and risk characterization (estimating the likelihood of adverse health outcome in exposed individuals and the uncertainties associated with the estimate) (WHO, 1999).

Hazard identification

A hazard is defined as a source of potential harm; this can include substances or machines, method of work and other aspects of work organization. Harm includes death, injury, physical or mental ill health, damage to property, loss of production or any combination of these. In the context of occupational health, ill health includes acute and chronic effects caused by physical, chemical or biological agents as well as adverse effects on mental health. Thus, the hazard identification process must evaluate all activities to determine whether a substance, situation or activity has the potential to cause harm.

The purpose of hazard identification is to evaluate the weight of the evidence for adverse effects in humans, based on assessment of all available toxicological and epidemiological data available for the agent of concern. The main purpose of hazard identification is to address two important questions:

- whether the agent poses a health hazard to humans, and
- under what circumstances the identified hazard may be expressed.

From the above definition and questions, it is clear that the process of hazard identification has two distinct steps, one largely involves a scientific judgement as to whether an agent can cause an adverse effect and its toxicological classification (labelling) and the other step focuses on both the circumstances of exposure and the inherent characteristics of the hazard. Classically, the latter is seen as the role of the occupational hygienist and the former of the occupational toxicologists, and both are discussed below.

Toxicological basis of hazard identification

In the EC, the approach to hazard identification differs according to whether a substance is an

existing substance or a new substance. Existing substances are substances that are listed in the European Inventory of Existing Commercial Chemical Substances (EINECS). Any substance not listed in EINECS is called a *new substance*. The assessment of the toxicity of new substances takes place before the substance is placed on the market, and involves animal experiments and *in vitro* assays. The statutory requirements for testing new substances for toxicological, physicochemical and ecotoxicological properties are contained in the Dangerous Substances Directive (Council Directive 67/548/EEC, 1997a). In the UK, this requirement is implemented by the Notification of New Substances (NONS) Regulations 1993. Guidelines for the conduct of tests investigating these properties have been drawn up and agreed by the international competent organizations, such as the Organization for Economic Co-operation and Development (OECD). Guidelines for testing new substances are contained in

Annex V of the Dangerous Substances Directive (Council Directive 67/548/EEC, 1997b).

Suppliers of substances and preparations that are considered to be dangerous also have a responsibility by law to produce safety data sheets. Safety data sheets contain, in addition to toxicological data, information on the identity of the substance, recommendations for handling, storage and disposal, physicochemical properties, first aid and fire fighting measures and transport information. Safety data sheets are intended to provide users of a substance with sufficient information about the hazardous properties of the substance, which, together with labelling information and risk and safety phrases, enable the user to establish the potential adverse health effects of the substance. For instance, on the basis of the results from acute toxicity tests, a substance may be classified as harmful, toxic or very toxic. Table 10.1 summarizes how results from acute toxicity tests are used to classify and label a substance.

Table 10.1 Classification and labelling of substances using acute toxicity data (Sadhra and Rampal, 1999).

Test result	Classification	Label	
		Symbol (indication of danger)	Risk phrase
LD ₅₀ (oral) ≤ 25 mg kg ⁻¹	Very toxic	T + (very toxic)	R28 – very toxic if swallowed
LD ₅₀ (dermal) ≤ 50 mg kg ⁻¹	Very toxic	T + (very toxic)	R27 – very toxic in contact with skin
LD ₅₀ (inhalation) ≤ 0.25 mg l ⁻¹ per 4 h (aerosols and particulates); ≤ 0.5 mg l ⁻¹ per 4 h (gases or vapours)	Very toxic	T + (very toxic)	R26 – very toxic by inhalation
LD ₅₀ (oral) > 25, ≤ 200 mg kg ⁻¹	Toxic	T (toxic)	R25 – toxic if swallowed
LD ₅₀ (dermal) > 50, ≤ 400 mg kg ⁻¹	Toxic	T (toxic)	R24 – toxic in contact with skin
LD ₅₀ (inhalation) > 0.25, ≤ 1 mg l ⁻¹ per 4 h (aerosols and particulates); > 0.5, ≤ 2 mg l ⁻¹ per 4 h (gases or vapours)	Toxic	T (toxic)	R23 – toxic by inhalation
LD ₅₀ (oral) > 200, ≤ 2000 mg kg ⁻¹	Harmful	Xn (harmful)	R22 – harmful if swallowed
LD ₅₀ (dermal) > 400, ≤ 2000 mg kg ⁻¹	Harmful	Xn (harmful)	R21 – harmful in contact with skin
LD ₅₀ (inhalation) > 1, ≤ 5 mg l ⁻¹ per 4 h (aerosols and particulates); > 2, ≤ 20 mg l ⁻¹ per 4 h (gases or vapours)	Harmful	Xn (harmful)	R20 – harmful by inhalation

When integrating the available data (human data, animal data, structure–activity relationships) for toxicological classification of substances, it is important to assess both the quality of the data and the weight of evidence available on which it is based. For instance when evaluating human and animal studies, one needs to take account of their relevance, reliability, quality and consistency. Each dataset needs to be evaluated carefully and criteria must be established to assess the strengths and weaknesses of studies. For instance a positive association between an agent and an effect may be interpreted as implying causality, if the following criteria are met: (1) there is no identifiable positive bias; (2) the possibility of positive confounding has been considered; (3) the association is unlikely to be due to chance alone; (4) the association is strong; and (5) there is a dose–response relationship (IARC, 1990).

Identifying hazards in the workplace

It could be argued that the identification of hazards is the most important step in any risk assessment. Only the identified hazards can be assessed and risk assessments will rarely reveal unidentified hazards. Ideally, hazard identification techniques should be applied as early as possible in the development of a process, i.e. the concept stage (and especially at the process design stage), when it is generally possible to make changes that are less expensive rather than having to enter into costly modifications once the process is up and running. The hazard identification process must then continue in different forms throughout the life of the process to ensure that procedures developed are correctly followed and process modifications, variations and faults are identified. Regulations in a number of European countries as well as the USA now require the application of hazard identification to existing plants presenting major hazards.

In the workplace, the risk assessor begins with an inventory of all known hazards that exist. This inventory of hazards can be developed from a list of chemicals purchased and used, understanding the process to determine intermediate products

and final products, by conducting a walk-through survey and ‘brainstorming’ by those who work in specific areas.

Both *continuous* and *non-continuous* hazards need to be identified. Continuous hazards are those that are inherent in the work activity or equipment under normal conditions; non-continuous hazards are hazards that arise from system failures (machine breakdown), non-routine operations (handling spillages, emergency procedures) or human errors. Various qualitative approaches to hazard identification are available, and the selection of the appropriate procedure will depend largely on the type of process and hazards involved. Procedures range from the use of a simple checklist carried out by a single person to the more complex and detailed open-ended hazard operability studies (Hazops) carried out by multidisciplinary teams. In every case it is essential to ensure that the reviewers are properly qualified and adequately experienced in the technique to be used and the process being reviewed.

Hazards in the workplace can be identified by a number of methods, ranging from simple checklists to specialized techniques used mainly in process design stage. Hazards identification techniques listed below are described by Sadhra and Rampal (1999).

Methods used to identify workplace hazards

- Accident and ill health statistics.
- Investigation of accidents, ill health effects and complaints.
- Audits.
- Checklists.
- Workplace inspections, including discussions and use of basic occupational hygiene instrumentation.

Specialized techniques used to identify hazards in the planning and design stages

- *Hazard and operability studies* (Hazops) – a qualitative technique used to identify hazards from hardware failures and human errors.

- *Failure mode and effect analysis* (FMEA) – an inductive technique used to identify hardware failures.
- *Task analysis* – an inductive technique used to identify likely sources of human error.

Use of occupational hygiene instrumentation in hazard identification

The majority of the hazards may be identified from knowledge of process and materials safety data sheets, observing and understanding the process, previous experience, records and the literature. However, in some situations the hazard may not be obvious or may require confirmation before proceeding to the risk assessment. Furthermore, as part of the workplace inspection an initial assessment of ventilation systems and an understanding of how the contaminated volume of air moves may also be helpful. In such cases the use of basic occupational hygiene instrumentation (smoke tubes, Tyndall beam dust lamp), photography techniques and numerous direct reading instruments can play an important role. These are described below.

A smoke tube consists of a glass tube containing concentrated sulphuric acid that is absorbed into inert granules. A continuous stream of smoke is produced by coupling the smoke tube to a hand-held positive pressure pump (aspirator). The smoke may be released at different points in the workplace to trace airflow patterns. The technique is particularly useful for the rapid assessment of the suction inlets of local exhaust ventilation (LEV) systems, i.e. a smoke cloud could be released at the source of the contaminant to show whether air is drawn from the source into the LEV system. The 'dust lamp' employs the Tyndall effect to reveal the presence and direction of movement of respiratory particles, which are normally invisible to the naked eye in normal lighting. The lamp produces a horizontal beam of light. When the beam passes through a cloud of dust, forward scattering of light occurs, which is visible to an observer looking along the beam of light in the direction of the lamp. The lamp can also be used to watch the performance of LEV systems associated with dust-emitting processes and the design can be modified if required to improve the capture

efficiency. It is useful to photograph or videotape the results from the use of smoke tubes and the Tyndall beam.

Other qualitative techniques used to visualize the flow of certain pollutants include infrared and Schlieren photography. Infrared photography is based on the fact that the majority of gases and vapours have dipole moments in their molecular structure, which cause them to have strong absorption peaks in the infrared region. Schlieren photography is based on the visualization of small differences in density, which cause changes in the refractive index of the gas, which can be made visible. It is also possible to integrate a pollutant sensor (for instance a monitor for measuring particulate concentrations) with a video recorder so that a concentration marker can be displayed on the recording, enabling a direct correlation between work activity and exposure profile.

Direct-reading instruments also play an important role in the identification or confirmation of suspicious hazards in the workplace. Instruments are available for measuring toxic gases, combustible gases, oxygen analysers, flammable substances, aerosols, noise levels, lighting levels, radiation intensity, vibration, heat stress, etc. The instruments range from the simple, easy-to-use colorimetric detector tubes to multispecific gas detectors (portable gas chromatographs, infrared analysers, photo-ionization detectors).

Assessment of dose–response

The dose–response refers to the relationship between the dose of a chemical and the response that it elicits. As the dose of the substance increases, e.g. the amount of the substance ingested or absorbed through the skin, or the airborne concentration of a gas, the response increases. The response may be expressed in terms of either the severity of a graded response or the percentage of a population affected by an adverse effect, and may range from no measurable response to maximal response. For instance in the case of an irritant gas the response could range in severity from slight irritation of the nasal passages to pulmonary oedema and bronchial constriction. The typical dose–response is presented by a cumulative frequency distribution curve, shown

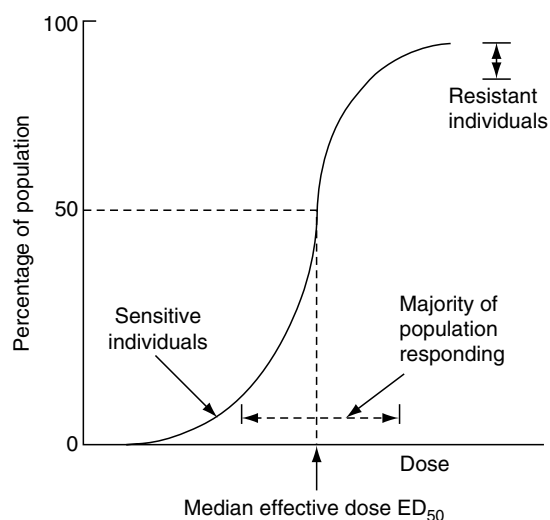


Figure 10.3 Typical dose–response curve.

in Fig. 10.3. In this figure, the response is expressed as the percentage of a population who respond to a low dose (sensitive individuals), the small proportion of the population who only respond to a high dose (resistant individuals) and the majority who respond around the mid-point, the median effective dose (ED_{50}).

In risk assessment, the dose–response relationship provides the basis for estimating the response associated with a particular exposure level. Implicit in the dose–response relationship is that, for most substances, there will be a dose below which no response is detectable, i.e. a threshold. The threshold may also be referred to as the no observable adverse effect level (NOAEL). For certain substances, however, such as those known to cause cancer involving a genotoxic mechanism of action, it is generally recognized that there is no threshold, or that a threshold cannot be identified with any certainty. The NOAEL is an estimate of the highest dose at which the incidence of a toxic effect or change in target organ weight was not significantly different from the untreated group. The NOAEL is thus an observed value, which does not take account of the nature or steepness of the dose–response curve. As discussed above, the identification of the threshold or NOAEL plays an important part in the establishment of occupational exposure limits.

Chemical exposures can lead to a variety of toxic effects that may be described in several ways. For instance, toxic effects may be described according to:

- duration of exposure (acute or local);
- site of tissue damage (local or systemic);
- occurrence of effect in relation to time of exposure (immediate or delayed);
- reversibility of effect (reversible or irreversible);
- target organ (e.g. renal toxicity, neurotoxicity);
- nature of toxic effect (e.g. whether functional, biochemical or morphological);
- specific toxic effects (e.g. carcinogenesis, sensitization, mutagenesis).

Exposure assessment

Humans are exposed to substances in the workplace, from use of consumer products and indirectly via the environment. Exposure is normally understood as external exposure, which can be defined as the substance ingested, the total amount in contact with the skin, or either the amount inhaled or the concentration of the substance in the atmosphere. However, if exposure occurs by more than one route it may be necessary to determine the total body burden. According to WHO (1999), exposure to a chemical is defined as ‘the quantitative or qualitative evaluation of the contact’, which includes consideration of the intensity, frequency and duration of contact, the route of exposure (e.g. dermal, oral or respiratory), rates (chemical intake and uptake rates), the resulting amount that actually crosses the boundary (a dose) and the amount absorbed (internal dose). For risk assessment based on dose–response relationships, the output usually includes an estimate of dose (Fig. 10.4). Doses are often presented as dose rates, or amount of a chemical dose (applied or internal) per unit time (e.g. mg per day), for instance, as dose rates per unit per body weight basis, e.g. mg kg^{-1} per day.

Given uncertainties in the assessment of exposure, the exposure levels should be derived from measured data (if available) and model calculations. In all cases, regardless of the sources of the data, they should be representative of the exposure situation being evaluated. The duration and fre-

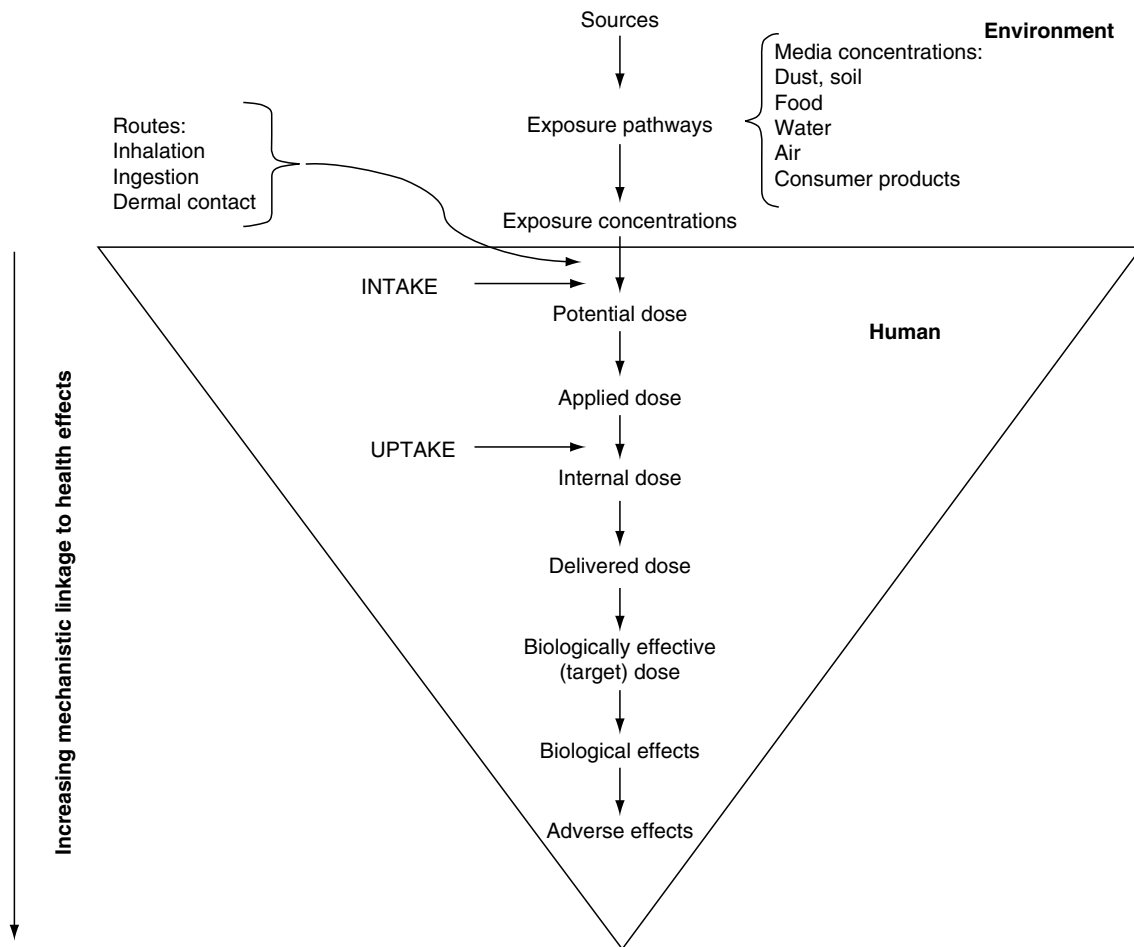


Figure 10.4 Exposure assessment and dose (from WHO, 1999).

quency of exposure, routes of exposure and work practices as well as control measures need to be considered carefully. The prediction of exposure should describe ideally both typical and reasonable worst-case scenarios. The section below discusses the general requirements of measuring workplace exposure.

Workplace exposure assessment

Whenever possible, high quality and relevant measured data should be used. Exposure data may already exist in organizations collected through various monitoring programmes over different time periods and for different reasons. It is

thus important to assess the relevance of the data both in terms of current practices and sampling strategies used to collect the data. Particular attention should be given to the conditions under which the data have been collected, in order to establish how representative they are both of the time periods and processes sampled. The confidence in the measured data is largely determined by the appropriateness of sampling techniques, sampling strategies employed and quality standards applied for sampling and analysis.

Two common techniques used to evaluate individual exposure are atmospheric monitoring and biological monitoring. Atmospheric monitoring is used to evaluate exposure where the main route

of uptake is inhalation, whereas biological monitoring of blood, urine or exhaled air may be the most appropriate to evaluate exposure in certain occupations where skin absorption and ingestion are the most important routes of entry to the body.

Sampling and analysis techniques can be divided into several categories (listed below) based on factors such as time, location and methods of collection and analysis:

- instantaneous or real-time monitoring;
- integrated or continuous sampling;
- personal monitoring;
- static (area) monitoring;
- active or passive (diffusion) flow monitoring;
- bulk sampling.

Exposure by inhalation is defined as the concentration of the substance in the breathing zone (personal sampling) and is usually expressed as an average concentration over a reference period. The reference period may be either 8 h to represent long-term exposure or 15 min to represent short-term exposure. The recorded exposure levels should take account of the use of personal protective equipment including respiratory protective equipment (RPE). The effectiveness of the RPE will depend on the inherent efficiency of the equipment and its correct use by the wearer. It is important that the exposure data are accompanied by sufficient information to place exposures in context with respect to pattern of use, pattern of control and other relevant process parameters.

Samples that are not taken on the individual are generally referred to as static (or area) samples. The two main benefits of static samples are when they are used to collect a large sample volume when the air concentration is low and when the aim is to determine the concentration of a contaminant in a specific location over time, such as before and after implementation of control techniques. Static sampling techniques are not always capable of measuring an individual's daily exposure accurately, especially when exposure sources are close to the breathing zone or when unplanned incidents such as machine breakdown or spillages occur or when the individual is very mobile in the workplace. Thus, personal sampling is preferred over static sampling when assessing exposure

of specific personnel conducting jobs/tasks of interest.

Selection of sampling and analysis methods

Exposure data should be collected following good occupational hygiene practice employing standardized and validated procedures, particularly on measurement and analysis methods. Depending on the survey objectives, very precise and accurate measurements may not always be necessary. However, the accuracy and precision of the method needs to be known. In selecting the most appropriate sampling and analysis methods the following factors should be considered:

Sampling method

- The physical and chemical properties of the contaminant.
- The stability of the sampling medium.
- Compatibility of the sampling medium with the subsequent analytical method.
- The capacity and the collection efficiency of the sampling medium.
- Type of analysis and information required.
- The intrinsic safety of the equipment.
- The portability, reliability and ease of equipment maintenance.

Exposure variation

Any sampling exercise will be limited by a number of factors, which influence the individual's exposures to a specific agent. The fluctuations in concentration of air contaminants are dependent on a number of factors (listed below). Exposure can also vary greatly, both within and between individuals, days, shifts, etc. In order to obtain representative exposure data for risk assessment, these variables need to be understood and considered carefully in the design of appropriate sampling strategies.

The main sources of variation that need to be considered include variation in:

- shift patterns and the average exposure of individuals;
- type and nature of processes;

- contaminant concentration in the breathing zone of operators over the duration of the shift;
- individual exposure levels, even when working in the same place, carrying out the same tasks on the same shift.

Within- and between-shift fluctuations in exposure concentrations can be due to variation in any combination of the following:

- number of emission sources;
- rate of release of the contaminant from a source;
- dispersion of a contaminant, i.e. the effect of air current and turbulence in the workplace;
- ambient conditions such as air temperature and humidity.

By careful consideration of the factors affecting exposure, the validity of the data generated can be maximized and thereby strengthen the inferences drawn. This should also ensure the most cost-effective approach is adopted.

Exposure modelling

As mentioned above, for occupational exposures, measured data are preferred to data derived from modelling. However, some measured data may be incomplete or of poor quality, necessitating the need to use both measured and modelled data. Exposure may also be modelled in the following circumstances:

- prior to the introduction of new processes, equipment or substances;
- prior to process modification;
- in the selection of substitutes for hazardous substances;
- to predict potential exposures from accidental releases of substances;
- to help reconstruct exposures in the aftermath of an accident;
- to help reconstruct exposures in retrospective epidemiology.

Exposure models can be conceptual, qualitative or quantitative, and might involve simple algorithms, complex numerical techniques or even scale models built of Perspex. Exposure models consist essentially of sources, a transmission path and a receiver (the worker), each of which can be modelled separately as described by Gray (1999).

Exposure models can be general in their application or specific evaluating single processes under defined conditions. An example of a general-purpose predictive model is EASE (Estimation and Assessment of Substance Exposure), which is used for general exposure profiling. EASE is a simple model that can be used to estimate exposure by inhalation or the skin. EASE is essentially a series of decision trees. The software asks a number of questions about the physical properties of the substance, the circumstances of its use and measures used to control exposure. The model is designed to predict 8-h time-weighted average (TWA) concentrations for normal use of the substance, but not unplanned events such as breakdown, spillages, etc.

Exposure modelling is often subject to considerable uncertainties. One way of dealing with uncertainties is to assume the 'worst case', e.g. choosing conditions that give the highest possible air concentration. It is also possible to carry out a 'best case' calculation so that a range of possible exposures can be estimated. This approach may appear to be reasonable but may be highly misleading as the best and worst cases could be improbable, extreme values and the range could be unrealistically wide. An alternative approach is to consider the natural ranges and uncertainties in all of the parameters of the model, and to use these to calculate a probability distribution for the exposure, rather than to calculate point estimates with unknown certainties. This is normally carried out using a 'Monte Carlo' simulation.

Risk characterization

Risk characterization aims to provide a synthesis of estimates of exposure levels and health risks: it also summarizes sources of uncertainties in scientific data and provides the primary basis for making risk management decisions.

Definitions and guidance for risk characterization have been published in US EPA (1996) as: 'a summary, integration, evaluation of the major scientific evidence, reasoning and conclusion of a risk assessment. It is a concise description of the estimates of potential risk and the strengths and weaknesses of those estimates'. The EU defines risk characterization as: 'the estimation of the

incidence and severity of the adverse effects likely to occur in a human population due to actual or predicted exposure to a substance' (Hertel, 1996).

It is evident from the above definitions that risk characterization is the final step in the risk assessment. It is designed to provide the critical scientific evidence and rationale required for decision-making. In order to manage risks effectively, risk characterization should provide information on the following:

- the extent of risk;
- individuals (populations) at risk;
- conditions of exposure believed to cause or contribute to the risk;
- the nature and magnitude of adverse consequences;
- the degree of confidence in the quality of data and risk estimates;
- uncertainty in the risk estimates;
- what data gaps exist and their impact on the evaluation of risk.

The risk criteria usually fall into three categories: (1) *comparative or equity based*, where the standard is what is held to be usually acceptable in normal life, or refers to some other premise held to establish an expectation of protection; (2) *cost-benefit analysis based*, where some direct comparison is made between a value placed on the risk of ill health and the cost of risk reduction/prevention measures; and (3) *technology based*, which essentially reflects the idea that a satisfactory level of risk prevention is attained when relevant best or 'state-of-the-art' technology is employed.

Acceptability and tolerability of risks

Hazards continue to exist either because society does not know of the risks associated with them or knows of the risks and accepts the levels of risk at which they exist. Are these risks acceptable or are they being tolerated because of the perceived or real benefit accrued from them?

When attempts are made to measure and rank hazards in order of priority for the purposes of control, questions of acceptability of risks emerge repeatedly. What level of risk is deemed to be acceptable? These questions have led to risk comparisons being made. Risks are less likely to be

acceptable if individuals or the community bearing the risks did not derive any benefit from them. Tolerability does not mean acceptability. It refers to the willingness to live with a risk to secure certain benefits, and in the confidence that it is being properly controlled. Tolerable risk is a range of risks that are not negligible and cannot be ignored, but which need to be reviewed and reduced still further if possible. These risks are undertaken on a regular basis for a benefit. The decision on what is a tolerable level of risk is usually a political one, taking into consideration the opinions given by various parties.

A comprehensive list of factors for judging tolerability of societal risk has been presented by the HSE (1993). These include factors related to the nature of the hazard, consequential risks and benefits, nature, purpose and limitation of the risk assessment. Economic factors, matters affecting the interest of the nation, political aims of government and interest groups, public concern about the activity and public confidence in regulatory authorities, plant operators, experts and emergency services also need consideration.

The HSE (1992), reviewing tolerability of risk from nuclear power stations, proposed levels that regulators in the UK apply. The HSE developed a three-tier system shown in Fig. 10.5, where risks are divided into three groups: the acceptable region, the ALARP (as low as reasonably practicable) or tolerability region, or the unacceptable region. The lower risk level is the level below which it does not warrant regulatory concern and for which no further action is necessary except to ensure that risk is maintained at the same level (1 in 1 million per person per year). The upper risk level is the level above which risks are not acceptable and cannot be justified on any grounds. The ALARP or tolerability region (risk level was 1 in 1000 per person per year for workers and 1 in 10 000 per person per year for the public) is the intermediate region between the broadly acceptable region and the intolerable risk region. The risks in this region are tolerable; however, there is a need to reduce the risks to as low as reasonably practicable. What this infers is that risks are tolerable only if risk reduction is impracticable or if the cost of risk reduction is grossly disproportionate to the improvement gained.

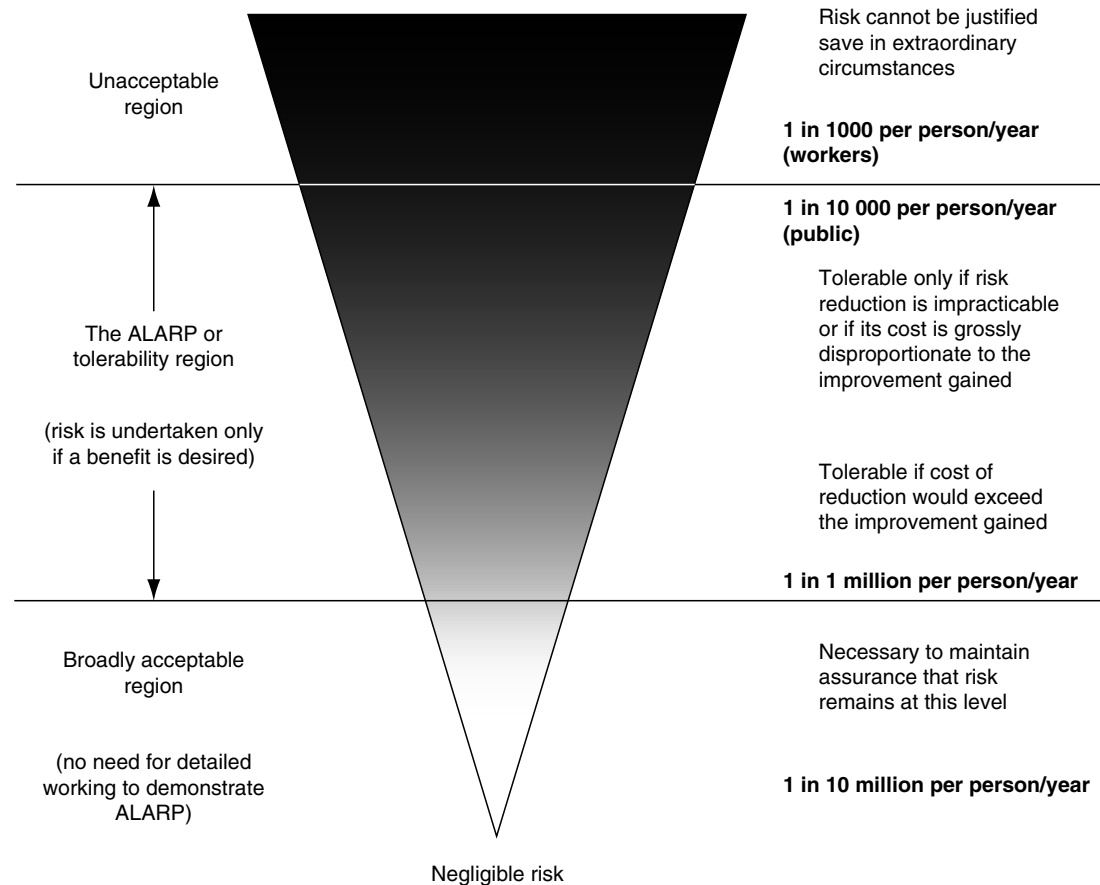


Figure 10.5 UK criteria for the tolerability of risk (HSE, 1992).

Comparison of exposure data with exposure limits

When comparing personal exposure data with exposure limits for compliance purposes, three basic conclusions can be reached:

- Exposures are above the occupational exposure limit, hence the need to identify reasons for the result and steps required to control the workers' exposure.
- Exposures are well below the exposure limit, hence controls need to be maintained.
- There is insufficient information to decide if exposures are either above or below the limit; either more information is necessary or prudent action should be taken to reduce the exposure of the workers.

When assessing for compliance, it is possible to take a pragmatic approach and simply divide the measured concentration by the OEL and make decisions based on this dimensionless index of exposure. For example, if the value is 0.1 and the short-term exposure limit (STEL) conditions are fulfilled then compliance is assumed. However, when there is a high degree of variability in high-risk situations, incorrect conclusions may be drawn.

Another technique to determine compliance is to use the mean and variance of the exposure distribution to calculate the probability of a measurement exceeding the OEL. If the probability is $\leq 0.1\%$ then compliance is assumed; if the probability is $> 0.1\%$ but $\leq 5\%$ then the situation is probably compliant but more measurements are

needed and, finally, if the probability is $> 5\%$ then the situation is not in compliance (Comité Européen de Normalisation, 1992).

Recording the risk assessment

Records should be kept of all risk assessment, especially when significant risks have been identified. The record serves as a point of reference to indicate the information and criteria used in the decision-making process. Regardless of the outcome of the assessment, reliable information should be available to defend judgements. The risk assessment should be conducted by a competent person and should include the following:

- activities/processes assessed;
- hazard types (planned and unplanned) and their characteristics;
- individuals potentially exposed;
- exposure routes;
- potential health effects (acute and chronic);
- frequency and pattern of exposure;
- estimate of exposure levels, including monitoring strategies employed;
- control types used and their effectiveness;
- reported symptoms;
- individuals at risk and extent of risk;
- recommendations and actions to minimize risks.

The results and recommendations of the risk assessment should be discussed with the various stakeholders so that the risks, uncertainties and the need for further measures, including additional resources, are understood and agreed.

Conclusions

The objective of risk assessment is to determine an estimate of the risk that will guide decision-making on the actions to be taken. Risk assessment is a dynamic science that continually evolves as a result of both scientific developments and the concerns of different stakeholders. Increasingly safety and health practitioners are expected to demonstrate how they have managed the risks and how this has contributed to the overall performance of the company. Therefore, the risk assessment process needs to be transparent, with statements describing un-

certainties and assumptions affecting the scope and interpretation of the assessment. Furthermore, risk assessment should only be regarded as acceptable when it is understandable and is compatible with the attitudes and the perceptions of those exposed to the hazards in the workplace.

References

- ACGIH (2003). *Documentation of the Threshold Limit Values and Biological Exposure Indices*. ACGIH, Cincinnati, OH.
- Comité Européen de Normalisation (1992). *Workplace Atmospheres – Guidance for the Assessment of Exposure to Chemical Agents for Comparison with Limit Values and Measurement Strategy*, PrEN 689. CEN, Brussels.
- Council Directive 67/548/EEC (1997a) Annex V. *Methods for the Determination of Physico-chemical Properties, Toxicity and Ecotoxicity*. EC Publication, Luxembourg.
- Council Directive 67/548/EEC (1997b) Annex VI. *General Classification and Labelling Requirements for Dangerous Substances and Preparations*. EC Publication, Luxembourg.
- Covello, V.T. and Merkhofer, M.W. (1993). *Risk Assessment Methods – Approaches for Assessing Health and Environmental Risks*. Plenum, New York.
- DFG (1994). *Biological Exposure Values for Occupational Toxicants and Carcinogens*, Vol. 1. VCH Verlagsgesellschaft, Weinheim.
- European Commission (1996). *Guidance on Risk Assessment at Work*. EC Publication, Luxembourg.
- Gray, C (1999). Exposure modelling. In *Occupational Health-Risk Assessment and Management* (eds S. Sadhra and K.G. Rampal), pp. 161–176. Blackwell Science, Oxford.
- Hertel, R.F. (1996). Outline on risk assessment of existing substances in the European Union. *Environmental Toxicology and Pharmacology*, 2, 93–6.
- HSE (1992). *The Tolerability of Risk from Nuclear Power Stations*. HMSO, London.
- HSE (1993). *Quantified Risk Assessment: Its Input to Decision Making*. HMSO, London.
- HSE (1995). *Generic Terms and Concepts in the Assessment and Regulation of Industrial Risks*. HSE Books, Sudbury, Suffolk.
- HSE (1996). *Use of Risk Assessment Within Government Departments*. HMSO, London.
- HSE (2002). *EH40/2002 Occupational Exposure Limits 1997*. HSE Books, Sudbury, Suffolk.
- IARC (1990). *Cancer: Causes, Occurrence and Control*, IARC Scientific Publications No. 100. International Agency for Research on Cancer (IARC), Lyon.
- Lopez, J. (1996). Taking chemicals to the limits. *Health and Safety at Work*, May, 21–4.

- Paustenbach, D.J. (1995). The practice health risk assessment in the United States (1975–1995): how the US and other countries can benefit from that experience. *Human and Ecological Risk Assessment*, 1, 29–79.
- Sadhra, S. and Rampal. K.G. (1999). *Occupational Health-Risk Assessment and Management*. Blackwell Science, Oxford.
- US EPA (1996). Proposed guidelines for carcinogen risk assessment. *Federal Regulations*, 6, 17960–8011.
- Vainio, H. and Tomatis, L. (1985). Exposure to carcinogens: scientific and regulatory aspects. *Annals of the American Conference of Governmental Industrial Hygienists*, 12, 135–43.
- Wells, G. (1996). *Risk Criteria. Hazard Identification and Risk Assessment*. Institute of Chemical Engineers, Rugby, UK.
- WHO (1999). *Environmental Health Criteria 210. Principles for the Assessment of Risks to Human Health from Exposure to Chemicals*. WHO, Geneva.

Chapter 11

Design of exposure measurement surveys and their statistical analyses

Hans Kromhout, Martie van Tongeren and Igor Burstyn

Overview	Observational and experimental study designs
Setting priorities: why measure exposures?	Documenting determinants of exposure in observational studies
Exposure variability	Analysis of data: multiple linear regression
Non-detectable exposures	Controlling for all sources of variability in exposure
Compliance with exposure limits	Understanding exposure to mixtures
Traditional strategies and their limitations	Exposure surveys and models for risk assessment
Efficient measurement strategies for hazard control	Summary of recommendations
Measurements for epidemiological studies	References
Identifying determinants of exposure	
Selecting determinants of exposure to document and study	

Overview

Any professional practice, such as occupational hygiene, is fraught with uncertainty. This is most easily seen when one considers measurements of exposure. Facing the task of exposure assessment and evaluation, an occupational hygienist has to decide both which exposures to give a priority to and how to best determine at what level (e.g. concentration in air) these exposures occur. Next to measurement devices and analytical techniques, essential tools in accomplishing these tasks are a variety of statistical techniques, which find their application in both design and analysis of exposure measurement surveys. Statistics comes to the aid of the occupational hygienist because it was explicitly designed to cope with uncertainty.

The array of statistical tools that an occupational hygienist has to be aware of (and preferably be a competent user of) has grown significantly in the last 20 years. This chapter will provide an overview of these statistical tools and the underlying concepts within a framework of exposure measurement surveys. This will be accomplished by first providing a motivation for measuring

exposures, as opposed to applying ‘expert systems’ that have been proposed as an alternative to direct measurements. Next, we will discuss some of the properties of exposure variability, and will introduce statistics that can be used to describe exposure levels observed in a survey. Upon these foundations, we will present methods for using exposure measurements to test whether observed exposure levels are in compliance with regulatory and health-based exposure limits. Both traditional and efficient strategies for assessing compliance and overexposure will be presented and contrasted. Readers will be introduced to a new computer application that greatly simplifies application of efficient measurement strategies for hazard control.

Other professional activities that an occupational hygienist might be involved in include exposure assessment for epidemiological studies, investigations of determinants of exposure (e.g. factors that can be used to reduce exposure levels) and risk assessment. In separate sections, we will present some basic guidelines on the measurement survey design and analysis considerations that are pertinent to these activities. The chapter is not meant to

replace an introductory course in applied statistics. However, occupational hygienists with no previous training in statistics will also find the chapter comprehensible and useful. Nonetheless, we would encourage such readers to consult many excellent introductory textbooks on statistics (e.g. Devore, 1982; Snedecor and Cochran, 1989), as

we will not be able to develop theoretical background for all the presented formulae and ideas in detail.

Throughout the chapter the ideas presented will be illustrated through an example of an exposure measurement survey, which was conducted at shipyards of the Royal Dutch Navy.

Example

The survey at the shipyards of the Royal Dutch Navy focused on exposure to welding fumes and solvents among workers of three large departments. An initial walk-through survey identified these groups of workers and exposures to be associated with the highest health risks. The goals of the subsequent measurement survey were threefold: estimation of long-term average exposure concentrations, their evaluation with respect to occupational exposure limits and, if necessary, identification of adequate exposure controls. Finally, a prospective measurement programme had to be developed to control exposures in the future.

Setting priorities: why measure exposures?

Most strategies for controlling workplace exposures start with what is often called an initial survey. The reason for this is obvious: when circumstances are so badly controlled that one can smell or see the risks involved, it would be a waste of resources to measure exposure concentrations before implementing control measures. On the other hand, when qualitative or semiquantitative methods yield results, which indicated that the situation is well controlled, it would be a waste of money to perform measurements. However, in many circumstances careful evaluation of existing conditions via exposure monitoring will be necessary to avoid 'being penny wise but pound foolish' (Kromhout, 2002). Recently, in the UK, a tendency has developed to stay away from exposure monitoring and to promote the application of general risk assessment tools like COSHH Essentials (Topping, 2001), especially in small- and medium-sized enterprises with exposures to chemical agents. Methods like COSHH Essentials as well as expert judgement by occupational hygienists are known for their inaccuracy and lack of appreciation for

the various components of exposure variability (Kromhout *et al.*, 1987; Hawkins and Evans, 1989; Post *et al.*, 1991). This will pose a problem particularly in situations where the working environment is neither very bad nor very good in terms of control of hazardous exposures. Nevertheless, it is essential to have a qualitative or semiquantitative scheme in place in order to prioritize the need for actual monitoring of exposures. Given the costs of monitoring programmes, we will never end up in a situation when every exposure would be measured on every occasion.

Exposure variability

In the occupational environment, exposure concentrations (intensities) vary enormously. For instance the variation in results of 8-h-shift-long measurements has been estimated to be between three- to 4000-fold (Kromhout *et al.*, 1993). For results of measurements with a much shorter averaging time, this variation can be much larger, as has been shown for instantaneous measurements of magnetic field levels (van der Woord *et al.*, 1999). The probability distribution of

concentrations is most often better described by a log-normal distribution than a normal one. This implies that the logarithms of concentrations (or other measures of intensity) are distributed normally (Oldham, 1953). The simple log-normal distribution can be described by only two parameters: the geometric mean (GM) and the geometric standard deviation (GSD). These parameters can be estimated by taking the antilog of the mean (m) and standard deviation (s) of the logarithms of N concentration measurements:

$$m = (N)^{-1} \times \sum_{i=1}^N \ln(X_i) \quad (11.1)$$

$$s = [(N-1)^{-1} \times \sum_{i=1}^N \{\ln(X_i) - m\}^2]^{0.5} \quad (11.2)$$

$$\text{GM} = \exp(m) \quad (11.3)$$

$$\text{GSD} = \exp(s) \quad (11.4)$$

Log-normality can be checked by formal statistical testing or by plotting the data-points on so-called probability plots (see example below).

The estimated mean and variance of a series of measurements are valid only for a stationary situation and should be re-estimated whenever the underlying exposure distribution changes, for instance owing to an increase in production levels for a significant period of time or installation of control measures. In general it is recommended to repeat exposure surveys on an annual basis, or after major process changes, in order to be sure that exposure distribution remains relatively stationary.

Example

In the shipyard, exposure to welding fumes was monitored prospectively for a period of 1 year. The measurement strategy was such that repeated personal measurements were collected within each of the four measurement periods and some repeated measurements were collected from persons employed in different time periods. This allowed estimation of the amount of variability present in average concentrations between workers within each period, but also over the entire period of 1 year. In addition, this measurement strategy enabled estimation of daily variation in shift-long averages. In Fig. 11.1 we see the sampling scheme as it was applied in shipyard 1 for measurement of the exposure to welding fumes. The results of the successful measurements fitted the log-normal distribution. In Fig. 11.2, the frequency distribution of measurement results for shipyard 1 is presented. It shows the characteristic skew to the right of a log-normal distribution.

The log-normal probability plot in Fig. 11.3 shows a straight line for most of the data. At the left tail (the lower end of the distribution), four values are visible, which were below the limit of detection. At the right end (the higher end of the distribution), three values with relatively extreme values are visible.

Formal statistical testing of this dataset revealed a Shapiro–Wilk W -statistic of 0.981 ($P = 0.07$). Indicating that the null hypothesis that these data-points are from a log-normal distribution could not be rejected. The Kolmogorov–Smirnov test statistic was close enough to 0 ($D = 0.074613$; $P = 0.08$) to assume log-normality.

As was indicated before, workers within the same department performing the same tasks and sharing the same working environment can have considerable differences in average exposure levels (between-worker variance) and will definitely experience varying exposure concentrations from day to day (within-worker or day-to-day variance) (Kromhout *et al.*, 1987, 1993; Rappaport *et al.*, 1993). The actual sizes of these variance components can be estimated from log-transformed exposure concentrations using a random-effects ANOVA (analysis of variance) model when the same workers are sampled on more than one occasion.

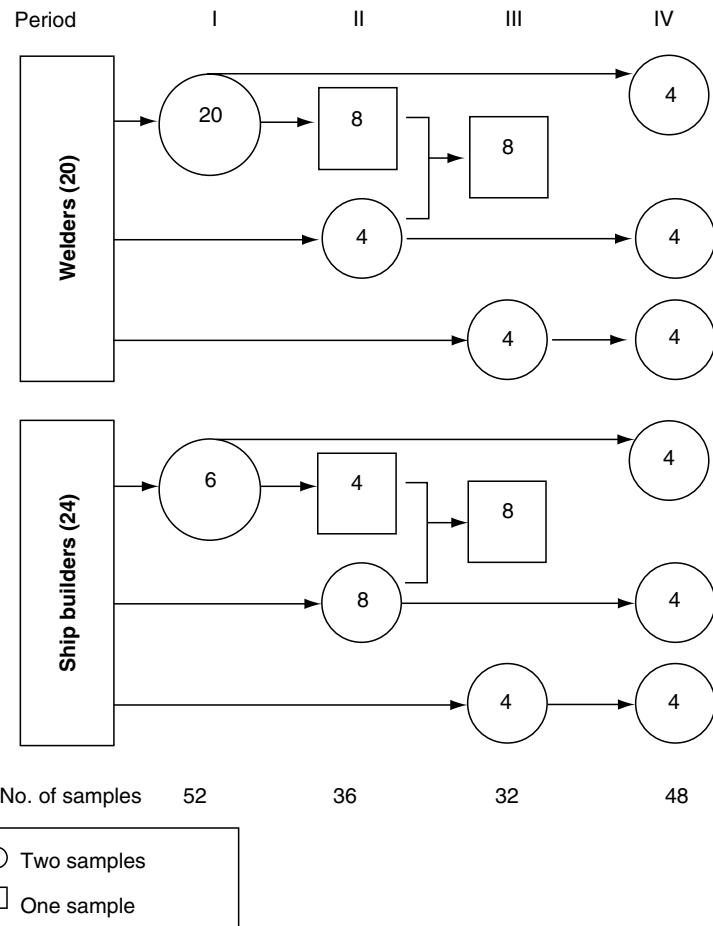


Figure 11.1 Measurement strategy applied to estimate long-term exposure to welding fumes in a shipyard.

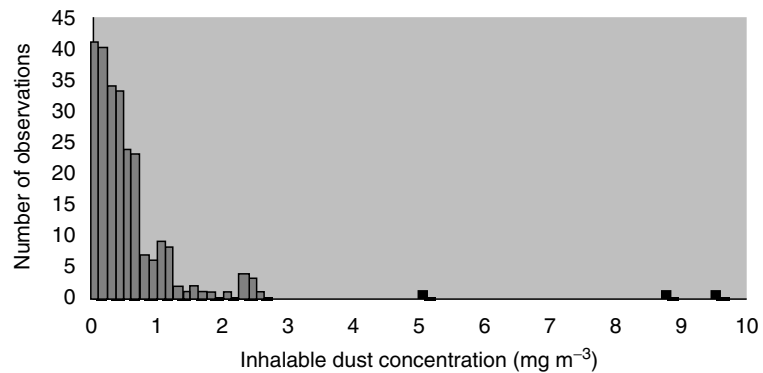


Figure 11.2 Frequency distribution of 129 inhalable dust (welding fumes) measurements from 32 workers during a 1-year period in shipyard 1 (AM = 0.71, GM = 0.40 and GSD = 2.89).

The random-effects ANOVA model is specified as follows:

$$Y_{ij} = \ln(X_{ij}) = \mu_y + b_i + \varepsilon_{ij} \quad (11.5)$$

for $(i = 1, 2, \dots, k)$ and $(j = 1, 2, \dots, n_i)$, where X_{ij} = exposure concentration of the i th worker on the j th day; μ_y = mean of Y_{ij} ; b_i = random deviation of the i th worker's true $\mu_{y,i}$ exposure

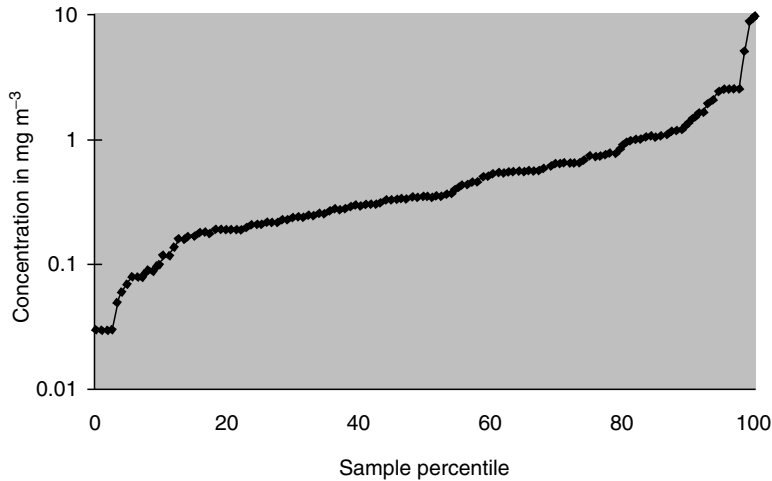


Figure 11.3 Log-normal probability plot of the same measurements as shown in Fig. 11.2.

from μ_y , and ε_{ij} = random deviation of the i th worker's exposure on the j th day, from the worker's true exposure, μ_y , i .

Under this model, both b_i and ε_{ij} are normally distributed with zero means: $b_i \sim N(0, \sigma_B^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_W^2)$. Also b_i and ε_{ij} are assumed to be statistically independent of each other. The parameters σ_B^2 and σ_W^2 are referred to as the components of the total variance $\sigma_T^2 = \sigma_B^2 + \sigma_W^2$. The estimated variance components (${}_T S_y^2$, ${}_B S_y^2$ and ${}_W S_y^2$ respectively) are used to estimate the total, between- and within-worker geometric standard deviation (${}_T S_g$, ${}_B S_g$ and ${}_W S_g$ respectively):

$${}_T S_g = \exp({}_T S_y^2)^{0.5} \quad (11.6)$$

$${}_B S_g = \exp({}_B S_y^2)^{0.5} \quad (11.7)$$

$${}_W S_g = \exp({}_W S_y^2)^{0.5} \quad (11.8)$$

It is also possible to estimate the ratios of the 97.5th and 2.5th percentiles of the total, the between- and within-worker distribution of exposures of each group of workers respectively (Rappaport, 1991):

$${}_T R_{0.95} = \exp[3.92({}_T S_y^2)^{0.5}] \quad (11.9)$$

$${}_B R_{0.95} = \exp[3.92({}_B S_y^2)^{0.5}] \quad (11.10)$$

$${}_W R_{0.95} = \exp[3.92({}_W S_y^2)^{0.5}] \quad (11.11)$$

These ratios provide information on the ranges of exposures experienced overall, between and within workers (i.e. from day to day). Rappaport (1991) has suggested that when the difference in exposure between the highest and the lowest exposed workers within a particular group (${}_B R_{0.95}$) is less than 2, this group can be defined as uniformly exposed.

Example

The estimated variances (${}_T S_y^2$, ${}_B S_y^2$ and ${}_W S_y^2$) for shipyard 1 were 1.13, 0.182 and 0.94 respectively. From this, the following GSDs could be calculated: ${}_T S_g = 2.89$, ${}_B S_g = 1.53$ and ${}_W S_g = 2.64$. The ratios ($R_{0.95}$) for the between- and within-worker distribution of exposures were 5.3 and 45 respectively. This implies that workers had their yearly average exposure within a factor of 5.3, whereas day-to-day exposures could vary up to a factor of 45. From this, it can be inferred that the welders at this shipyard did not belong to a uniformly exposed group.

Non-detectable exposures

It often happens that in an exposure measurement survey some exposure levels are so low that they cannot be accurately quantified. Such measurements are called 'non-detectable' or below 'method limit of detection' (MLOD). In statistical jargon, non-detectable values result in left censoring (or truncation) of exposure frequency distribution, because it appears to be cut off at the level corresponding to the MLOD. The MLOD is calculated from field blanks that yield 'analytical

limit of detection', and description of the sampling method. One example of the field blanks are particle filters that were taken to a workplace in sampling cassettes, but no air was drawn through them. Three standard deviations of weight change in such field blanks can be taken as an indicator of minimum change in weight that can be detected on these particle filters. The number of field blanks should be of the order of 10–20% of measurements made. All non-detectable dust measurements in such a study should be presented as '< MLOD mg m⁻³' in tables and reports.

Example

In the survey among workers in shipyard 1, the analytical limit of detection based on field blanks varied somewhat between time periods and was of the order of 0.10 mg. This value can be used to calculate MLOD, under the assumption that 21 min⁻¹ of air was drawn through an average particle filter for 480 min, yielding a method limit of detection of $0.10 \text{ mg} / (21 \text{ min}^{-1} \times 480 \text{ min} \times 0.0011 \text{ m}^{-3}) = 0.1 \text{ mg m}^{-3}$.

An obvious solution to the problem of non-detectable exposure measurements is to use a more sensitive exposure measurement tool in the future, but this is not always possible, and does not help one to get the best information from a series of measurements with non-detectable samples. There are several rules for dealing with non-detectable values that should be kept in mind.

- Non-detectable samples should not be discarded from calculations because this would produce a bias (distortion) in mean exposure towards a higher value.
- Non-detectable values are not zero (i.e. they do not necessarily imply absence of the monitored substance from the work environment).
- Non-detectable values should always be reported by analytical laboratories as less than 'analytical limit of detection', where 'limit of detection' is a number (e.g. < 0.01 mg dust/filter).
- The best way to deal with non-detectable values in statistical analysis is to replace them with a reasonable guess of what the non-detectable values should have been if we had a more accurate measurement device.

In most situations, replacing non-detectable measurements with MLOD/2 is expected to work well in producing unbiased estimates of the mean (Hornung and Reed, 1990). However, replacing all non-detectable values with a single value always produces standard deviations that are too small, and this has impact on statistical tests about the mean of exposure distribution. More advanced statistical methods can overcome this problem (Little and Rubin, 1987; Taylor *et al.*, 2001), but for now these are not readily available.

A complication arises when too many measurements in a survey are non-detectable (e.g. > 50%). In such situations, it is best to avoid calculating means and standard deviations. A frequency histogram may adequately describe the data, and either (1) a more sensitive measurement technique should be used in the future or (2) the exposures may be deemed too low to warrant further investigation. Professional judgement, rather than statistics, is the best guide on how to proceed when more than 50% of measurements are non-detectable.

Compliance with exposure limits

Traditional strategies and their limitations

Over the years, several approaches have been propagated to test compliance with occupational exposure limits (OELs). The most infamous of these approaches is called the compliance or Occupational Health and Safety Administration (OSHA)/National Institute for Occupational Safety and Health (NIOSH) evaluation based on a single measurement (Leidel *et al.*, 1977). In it, workers sampled are those considered to have the highest exposure (worst-case approach). The evaluation of measurement results is based on the individual data. A one-sided upper confidence limit is calculated at the 95% confidence level for each sample from the coefficient of variation of the sampling/analytical method. If the upper confidence limit is below the exposure limit, the exposure is considered to be below the limit. More sophisticated methods try to estimate the probability of exceeding the exposure limit (the percentage

of days that the actual shift-long average exposure will be above the OEL. The actual estimation of the probability of exceedance (γ) is rather straightforward:

$$Z_{\text{observed}} = [\ln(\text{OEL}) - m]s^{-1} \quad (11.12)$$

$$\gamma = 1 - \Phi(Z_{\text{observed}}) \quad (11.13)$$

where ‘ $\ln(\text{OEL})$ ’ is the natural logarithm of the OEL, ‘ m ’ is the mean and ‘ s ’ is the standard deviation of the log-transformed concentrations (see Equations 11.1 and 11.2), ‘ Z_{observed} ’ is the standard normal variable [$Z \sim N(0, 1)$], and ‘ $\Phi(Z_{\text{observed}})$ ’ is the probability that $Z \leq Z_{\text{observed}}$. The probability of the Z-score being less than or equal to the observed Z-score (Z_{observed}) can be derived from a table of the cumulative normal distribution that gives the area under the standard normal curve from $-\infty$ to Z_{observed} . These tables can be found in most introductory textbooks of statistics (e.g. Snedecor and Cochran, 1989).

Example

As shown before in Fig. 11.2, the concentration of the welding fumes among 32 workers in shipyard 1 had a GM of 0.40 mg m^{-3} ($m = -0.91$) and a total GSD ($_{T}S_g$) of 2.89 ($s = 1.06$). With an OEL of 3.5 mg m^{-3} , the Z_{observed} can be estimated to be: $Z_{\text{observed}} = [\ln(3.5) - (-0.91)] \times 1.06^{-1} = 2.03$. Looking this up in a cumulative normal frequency distribution yields as area from $-\infty$ to Z_{observed} : 0.979. In this case we have $\gamma = (1.00 - 0.979) \times 100 = 2.1\%$ chance of exceeding the OEL of 3.5 mg m^{-3} . The actual measurement series had three (out of 129) 8-h time-weighted average concentration measurements that exceeded this value (2.3%).

In both these approaches, two assumptions play an important role. The first is the notion of the homogeneous exposure group of workers. The idea behind it is that workers performing the same tasks (or having the same job title) in a given location (factory, department, etc.) will also share the same exposure concentrations. In other words, the assumption is such that in the long run these workers will experience the same average exposure. Therefore, it is not necessary to sample more than one worker because, under this assumption sampling, 10 workers on one day will result in the same estimate of the average concentration of this

group and day-to-day variability as sampling five workers on 2 days. The other assumption is that by focusing on groups with assumed high exposures, the sampling strategy will be more effective (worst-case sampling). The idea is that when even the highest exposed group stays under the acceptable probability of exceedance, there will be no need to sample lower exposed groups of workers. This assumption stands and falls with the ability of experts to assess the level of exposure qualitatively or semiquantitatively in a meaningful way. Research in the reliability and accuracy of subjective methods of exposure assessment has shown that

this assumption is not met in most circumstances (Kromhout *et al.*, 1987; Hawkins and Evans, 1989). Experts appear to have a tendency to overestimate exposure levels and often fail to identify worst-case exposure situations.

Efficient measurement strategies for hazard control

To overcome the problems presented above, a new measurement strategy for evaluating exposures with health effects due to chronic exposure has been introduced by Rappaport *et al.* (1995) and Lyles *et al.* (1997). Central in this approach is the notion that exposures tend to vary in more than one dimension: not only from day to day, but also between workers within the same group. The drive to develop this strategy came from a comprehensive review of a database with repeatedly sampled groups of workers with data from North America and Europe. These analyses showed that only 25% of all groups of workers considered, defined by job and location, had their individual average concentration within a factor of 2 (${}_B R_{0.95}$). The other 75% of groups could not be defined as uniformly exposed, whereas in 10% of the groups individual average exposures differed by a factor 50 or more (Kromhout *et al.*, 1993). Recently, it was shown that the distributions of day-to-day and between-worker variability in dermal exposure levels were very similar compared to those found for airborne exposure (Kromhout and Vermeulen, 2001). In

conclusion, by assuming that all workers within a group have the same mean exposure, some workers with distinctly higher average exposure might not be adequately protected.

Therefore, a simplistic log-normal model of exposure concentrations defined by only two parameters, one for central tendency (GM) and one for variability (GSD), had to be replaced by the random effects ANOVA model. In this model, there is still one measure of central tendency (overall mean exposure of the group), but individual workers are allowed to have a distinct personal mean exposure. Furthermore, there is the day-to-day component of exposure variability, which in most cases will be larger than the between-worker differences. Other than estimating the probability of exceedance (the change in a single measurement being above the OEL), this strategy focuses on the probability of overexposure (chance that a randomly selected worker's mean exposure is greater than the OEL) (Tornero-Velez *et al.*, 1997). The formula for probability of overexposure (Θ) is similar to the formula for exceedance, but focuses on the between-worker distribution of mean exposures:

$$Z_{\text{observed}} = [\ln(\text{OEL}) - (m + 0.5_w S_y^2)] \times ({}_B S_y^2)^{-0.5} \quad (11.14)$$

$$\Theta = 1 - \Phi(Z_{\text{observed}}) \quad (11.15)$$

ANOVA-based assessment of the probability of overexposure can be implemented via the Microsoft Excel macro called SPEED (available through www.iras.uu.nl).

Example

SPEED was used to evaluate the 129 inhalable dust (welding fumes) measurements from 32 workers during the 1-year period in shipyard 1. The estimated parameters (assuming a random effects ANOVA model) were as follows: GM = 0.40 mg m⁻³; GSD_{ww} = 2.64 and GSD_{bw} = 1.53. The exposure pattern was evaluated against the Dutch OEL for welding fumes of 3.5 mg m⁻³. A point estimate of the probability of overexposure could be calculated as follows: $Z_{\text{observed}} = [\ln(3.5) - (-0.91 + 0.5 \ln(2.64)^2)] \times \ln(1.53)^{-1} = 3.97$. Looking this up in a cumulative normal frequency distribution yields as area from $-\infty$ to Z_{observed} : 0.99996. In this case we have $\Theta = (1.00 - 0.99996) \times 100 = 0.004\%$ chance of an individual worker having a mean exposure greater than the OEL of 3.5 mg m⁻³.

(Continued)

Example (Continued)

In addition to the point estimate, the SPEED program also takes into account the power of the tests when evaluating if the exposure situation is acceptable. The power of the test is dependent, among others, on the number of measurements. In this situation, SPEED evaluated the exposure as acceptable. If the power had been insufficient, the program would have requested more measurements to be taken. The following figures are output from the program SPEED. Figure 11.4 presents the individual logged exposure concentrations for each of the workers in the group. Figure 11.5 shows the goodness of fit of the random effects model to the measurements within this group and, finally, Fig. 11.6 shows the uniformity of the workers' exposure. The last figure shows that all estimated mean exposures are within a factor of 4 from each other and well below the OEL.

Measurements for epidemiological studies

Studies of the effect of occupational exposure on health frequently lack sufficient exposure information, especially in retrospective case-control or

cohort designs. Few companies collect exposure data on a routine basis, and the current trend is away from collecting exposure data in favour of using generic risk assessment tools. Therefore, retrospective studies often rely on the use of semi-

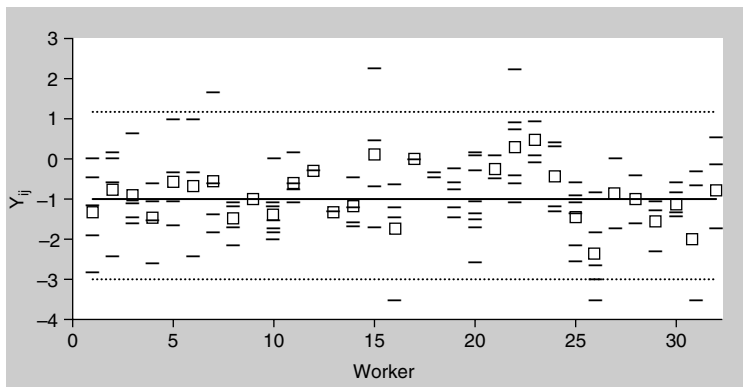


Figure 11.4 Distribution of the 129 inhalable dust measurements: the short horizontal lines represent the individual (logged) exposure data (Y_{ij}); the open squares represent estimated mean values of the logged exposures (Y_i); the solid line represents group mean (logged) exposure; and the dashed lines represent one standard deviation (SD) above or under the group mean.

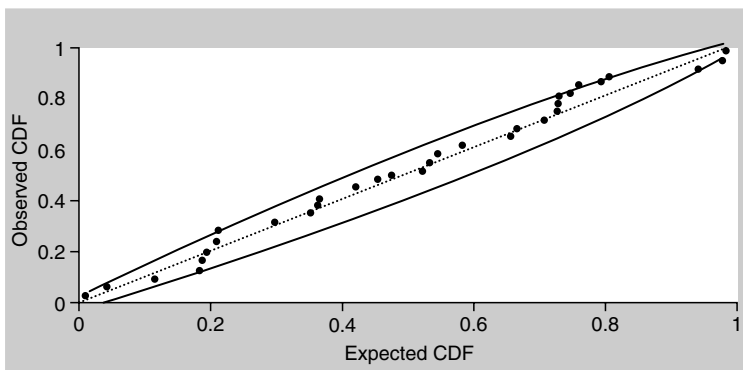
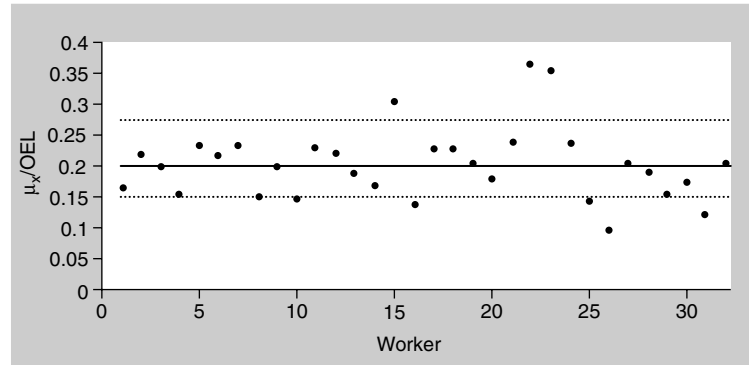


Figure 11.5 Graphical assessment of the fit to the random effects model. The expected and observed cumulative distribution functions (CDFs) are presented in the format of p - p plots. The individual points represent the ranked workers in the sample. The dashed line represents a perfect fit to the model, when observed probability equals expected probability. The two solid curves are error bands representing ± 1 SD of the expected CDF.

Figure 11.6 Assessment of uniformity: points represent predicted mean exposures for the workers expressed as a proportion of the OEL. The solid line represents the estimated mean exposure for the group (μ_x). With μ_x at 0.71 mg m^{-3} and an OEL for welding fumes at 3.5 mg m^{-3} , the line crosses the y-axis at $0.71/3.5 = 0.20$. The two dashed lines represent the approximate 90% confidence interval (CI) for μ_x .



quantitative (high, medium or low) or qualitative ('yes/no' exposed) occupational classifications, although some examples exist of retrospective studies when quantitative exposure estimates have been obtained after successfully modelling exposure (e.g. Burstyn *et al.*, 2003).

In prospective and cross-sectional studies of diseases, we have the ability to incorporate collection of quantitative data from the very beginning of the study and during follow-up. Owing to budgetary limitations, it is not usually possible to measure exposure for each individual at all times when exposure occurs. The only exception to this is the systematic assessment of exposure among workers potentially exposed to ionizing radiation. In most occupational studies, workers are grouped on the basis of similarities in exposure, using information on their job description and place of work, and exposures of a selection of workers from each group are measured.

During the 1950s, sampling strategies were designed by Oldham and Roach (1952) and Ashford (1958) to collect data on coal dust exposure for a longitudinal study of pneumoconiosis among coal-miners in the UK. Ashford (1958) defined stratum or occupational groups based on occupation, place of work and shift, and a large number of possible work shifts were selected randomly to be included in the measurement programme. The number of measurements allocated to each occupational group depended upon a number of factors, such as standard deviation of exposure, the number of workers in each group and the duration of employment within a group. Cumulative exposure for each individual miner was estimated, based

on the results of the exposure measurements and the duration of employment within a group. Only fairly recently have these methods for designing sampling strategies for occupational exposure assessment in epidemiological studies been rediscovered (Kromhout *et al.*, 1995; Sauleau *et al.*, 2003).

The strategies of Oldham and Roach (1952) and Ashford (1958) assumed that all groups were homogeneously exposed, that is exposure for all miners within each group could be described by a single distribution. As described before in this chapter, several studies have shown that within groups comprising apparently similarly exposed workers large differences in individual mean exposure can occur (Kromhout *et al.*, 1993; Rappaport *et al.*, 1993). These differences can be due to the inappropriate grouping of employees, but are also often due to individual differences in working practices. Especially in this day and age of multi-tasking within industry, it has become increasingly difficult to identify groups of individuals who will carry out the same activities, day-in and day-out, under the same conditions.

The performance of an assessment strategy in an epidemiological study does not only depend on the homogeneity of exposure groups, but is also determined by the contrast in exposure between the groups and the standard error of the mean exposure in a category or precision (Prais and Aitchison, 1954; Kromhout and Heederik, 1995). The importance of contrast in exposure is intuitively understood, as it is much easier to determine the effect of exposure when the difference between the low- and high-exposed groups is large (say a factor 100) compared with when this is only marginal.

In the latter case, you would need much more precision in the exposure estimates and more measurements will be required.

Lack of contrast in exposure, precision and homogeneity in exposure groups can result in bias in the exposure–response relationships. Equations have been formulated by Kupper and published by Kromhout *et al.* (1996) and Tielemans *et al.* (1998), which estimate the attenuation (the bias towards zero) of the exposure–response relationship and the standard error of the exposure–response slope. These equations are only applicable for linear exposure–response associations, using normally distributed exposure estimates in the absence of any confounding or modifying variables. Also, the equations assume that both within- and between-worker variance components are constant for all workers and groups. Therefore, these equations cannot be applied to adjust exposure–response associations, but should only be used as a tool to develop efficient exposure classification systems.

Contrary to general perception, it has been shown in a number of occupational situations that the bias towards zero effect is larger when using individual results (each worker is assigned his or her own measured exposure) compared with applying grouping strategies (each worker is assigned mean exposure of the group to which they belong). This has been demonstrated clearly by Tielemans *et al.* (1998) and is consistent with well-established statistical theory of measurement error.

Generally, when grouping workers a priori on the basis of similarities in working conditions and exposure, there is little bias in the exposure–response slope, even when some of the groups are not homogeneously exposed (van Tongeren *et al.*, 1997, 1999). However, the precision of the slope is affected by inappropriate grouping of workers, and considering the trend towards lower occupational exposures (Kromhout and Vermeulen, 2000; van Tongeren *et al.*, 2000; Vermeulen *et al.*, 2000), it is important to investigate the variance components of exposure in relation to the contrast among exposure groups to determine if the study has sufficient power.

In summary, there is no widely accepted model to determine the number of measurements and the allocation of measurements. However, some gen-

eral rules are available, suggesting that the measurement effort should be focused on largest groups, on groups with longest durations of employment and on groups with highest variability of exposure.

Identifying determinants of exposure

This section is devoted to an overview of exposure measurement surveys that an occupational hygienist may wish to conduct in order to identify factors that influence exposure levels in a workplace. Burstyn and Teschke (1999) give a comprehensive review of all the issues raised below.

Selecting determinants of exposure to document and study

Each exposure measurement occurs in a particular context. Thus, when measuring exposures, occupational hygienists not only sample concentrations of chemicals in the workplace, but also the circumstances under which those exposures occurred. These circumstances (i.e. context) become the determinants of exposure, if they can be used to predict exposure level. Consequently, it is paramount that the occupational hygienists take as much care documenting determinants of exposure as they do accurately measuring exposures. For example, it can be of vital importance for the interpretation of measurements to know that they were collected during disturbance in the process or other atypical production conditions. Some of the key contextual information that must be collected and registered (i.e. stored securely along with exposure measurements) during exposure survey can be found in Table 11.1.

The choice of the specific determinants of exposure to document and the degree of detail in documenting them influences an occupational hygienist's ability to interpret and apply results. In epidemiological studies, it may be sufficient to place measurements into the correct department and time period. However, in order to design effective exposure-reducing measures, more detailed documentation is needed (e.g. controls already in place). Thus, in deciding how to document determinants of exposure, occupational hygienists must

Table 11.1 Contextual information to be collected and registered during exposure measurements (adapted from Kromhout, 2002).

Category	Information
Strategy	Worst case or randomly chosen worker Worst case or randomly chosen time periods Task or full-shift based Reason for collecting measurements Duration Sampling method Analytical method
Location	Type of industry Department Number of employees in the department Calendar date of measurement
Worker	Personal identification code Gender Age Worker behaviour (e.g. tasks performed) Seniority Hand of preference Personal protective equipment use Machines and tools used Pace of work Degree of training Mobile or stationary worker
Process	Level of automation Continuous or intermittent Control-/exposure-reducing measures
Environment	Indoors or outdoors Temperature, atmospheric pressure, relative humidity Weather conditions (for outdoor work) General ventilation Room volume (e.g. confined space < 50 m ³) Day or night shift
Agent	Likely sources (e.g. composition of raw materials) Physical characteristics (e.g. powder versus pellets versus liquid)

judge how the observed determinants can be used to improve working conditions if they are found to be important predictors of exposure level.

Observational and experimental study designs

Most studies of determinants of exposure are observational in nature. This means that measure-

ments and observations are made on the working conditions that exist outside of the investigator's control. Such surveys can be very informative in identifying both hot spots of exposure (where intervention is needed) and effective exposure controls that are already in place. However, in some circumstances it might be advantageous to adopt an experimental approach, where some machines and work practices are tested to see how they influence exposure levels. Although apparently attractive, experimental studies are at a considerable disadvantage with respect to observational studies, because in an experiment one cannot test all the conditions that may arise in a workplace. Therefore, results of experimental studies should always be validated in real workplaces. Recently published papers on design and evaluation of interventions in a workplace can provide helpful guidance in the design of observational studies that test specific control measures (Lazovich *et al.*, 2002a,b). The number of measurements that have to be collected depends on the number of potential determinants of exposure, but in a typical factory-wide survey one can expect to gather 50–200 individual exposure measurements (Burstyn and Teschke, 1999).

Documenting determinants of exposure in observational studies

Observational studies must document in detail those worker and workplace characteristics with a potential to influence exposure levels during the measurement period. Some of these determinants of exposure are stable and can be documented at any point in time during walk-through surveys or by interviews with company personnel (e.g. engineering exposure controls). Other potential determinants are associated with daily activities (e.g. time spent on each task, machines and materials used, use of exposure controls) and therefore must be documented during exposure monitoring. Systematic direct observation of workers is the most reliable technique that can accomplish this. During direct observations, all relevant determinants of exposure are documented on standard data sheets at regular time intervals (e.g. every 5–15 min). Such approach is labour intensive and is often not practical, for example when work is carried out in

confined spaces or when one occupational hygienist is observing more than 10 workers. Consequently, task-profile diaries, worker interviews, questionnaires at the end of sampling and video-

taping have been developed as alternatives. These alternative methods of ascertaining determinants of exposure must always be validated against direct observations in pilot studies.

Example

In the study in the Royal Navy Shipyard, the measurements were carried out using the measurement scheme as shown above with very limited input of professional hygienists. The hygienist assisted only during the first period. A local trained technician consequently collected the samples in later periods. The sampled workers kept logs of activities (welding, welding-related activities like grinding and gouging, other), of places where activities were performed, of type of welding (gas, manual arc, MIG, MAG, TIG, etc.), of controls and personal protective devices used and whether they smoked during the sampling. This semi-self-assessment scheme (with randomly selected workers and randomly selected days within the four measurement periods) aimed to obtain 213 samples, but resulted only in 129 usable measurements of inhalable dust concentrations: 84 planned measurements failed due to no show of the worker (10%), sick leave (8%), holidays (7%) and pump failures and fraud (14%).

Analysis of data: multiple linear regression

Data arising from an observational study aimed at identifying determinants (X) of exposure (Y) are typically analysed via multiple linear regression. The equation fitted to the data takes the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \text{error} \quad (11.16)$$

where β_0 is the model intercept (exposure in reference group) and β_1 is the regression coefficient for determinant of exposure X_1 , etc.; intercept and regression coefficients are estimated from the data. The determinants of exposure can be dichotomous (1/0 = factor present or absent) or continuous (e.g. increase in bitumen fume concentration per increase in degrees centigrade of asphalt temperature). Consequently, regression coefficients can reflect either change in exposure due to the presence or absence of a dichotomous determinant or change in exposure due to increase or decrease in level of a continuous determinant. The model assumes that both Y and error term are normally distributed and independent of each other. The degree to which a regression model explains

observed data is characterized by the coefficient of determination (R^2), which reflects the proportion of variability explained by a model.

It is typical for multiple regression models to explain on the order of 50% of variability in exposure level. In selecting a model that best describes the data, it is advisable to retain determinants of exposure that have statistically significant regression coefficients (e.g. the chance that β_i is not different from zero should be less than or equal to 10%, $P < 0.10$). Most common statistical packages have automated model selection procedures, such as those that consider all possible models and select the ones with the highest R^2 , adjusted for the number of determinants in the model. Multiple linear regression models are fitted to the data using a method that minimizes the square of the difference between predictions of the model and observed values. Any regression model's assumptions should be tested using standard techniques such as residual plots (Kleinbaum *et al.*, 1988). An extension of this approach is mixed effects models that were used in the following example, and will be explained in more detailed in the following section.

Example

Even although the exposures to inhalable dust were considered to be acceptable for the workers in shipyard 1, collection of auxiliary data during the measurements enabled estimation of the contribution of determinants of exposure to the measured levels of inhalable dust. A mixed model explained 33% of the total variance. The factors shown in Table 11.2 appeared to be important. From the results of the mixed model it can be seen that all workers from shipyard 1 have a background exposure to inhalable dust of at least 0.13 mg m^{-3} . It also appears that there is a seasonal effect with higher concentrations in the winter and spring. In the spring period (April), exposure concentrations were 1.8 times higher than in September (dry and hot, with more general ventilation). MIG and MAG welding led to a higher exposure by factors of 1.9 and 3.7 respectively. Use of local exhaust ventilation lowers the exposure during MAG welding by almost a factor 3. Remarkably, manual arc welding with local exhaust ventilation increases exposure concentrations. The reason for this is not clear, but might be related to the intensity of welding, with or without local exhaust ventilation. Secondary activities, like gouging and grinding, lead to higher exposures with factors of 1.5 and 3.8 respectively. The duration of the (welding) activities seems to be an important factor. This is not a surprise given the character of the work: maintenance of ships, rather than continuous production. The environment (confined space versus outside the ship or outdoors) is an important determinant of exposure as well. This multivariate model predicts, for a worker in the spring season, who performed MAG welding for 241 min inside a ship, a median exposure to inhalable dust of $\exp(-2.01 + 0.57 + 1.15 + 241 \times 0.0016 + 0.65) = \exp(-2.01) \times \exp(0.57) \times \exp(1.15) \times \exp(241 \times 0.0016) \times \exp(0.65) = 0.13 \times 1.77 \times 3.16 \times 1.47 \times 1.92 = 2.05 \text{ mg m}^{-3}$. Had this worker used local exhaust ventilation, the estimated median exposure would have been: $\exp(-1.03) \times 2.05 = 0.73 \text{ mg m}^{-3}$.

Because occupational exposure levels are typically log-normally distributed, we often use the logarithm of measured exposure in multiple linear regression modelling (i.e. $Y_{ij} = \ln(X_{ij})$, where X_{ij} = measured exposure). As a result, models fitted to the data actually have multiplicative form when transformed to the original scale (e.g. mg m^{-3}):

$$E = \exp[\beta_0] \times \exp[\beta_1 X_1] \times \exp[\beta_2 X_2] \times \dots \quad (11.17)$$

Some of the predictions of such models can become quite unrealistic outside the range of the data on which they were built. Hence, caution is advised in applying such models in a context not described by a statistical model. Given the pure statistical and not physical nature of these models, utmost care should be taken when using these models to predict exposure.

Example

Using the same mixed model as before, one could estimate the median exposure of a worker who performed MAG welding under the same conditions, but now for the full shift of 480 min to be as follows: $\exp(-2.01 + 0.57 + 1.15 + 480 \times 0.0016 + 0.65) = 3.09 \text{ mg m}^{-3}$. The result of 3.09 mg m^{-3} is not unrealistic for an environment in which welding takes place, but it is unrealistic for the shipyard because full-time welding does not take place here.

Table 11.2 Results of a mixed model for inhalable dust in shipyard 1 ($n = 86$).

Factor	β	Multiplier [exp(β)]	95% confidence interval for the multiplier
<i>Season</i>			
Spring	0.57	1.78	1.19–2.66
Summer	0.06	1.06	0.67–1.68
Fall	0.00	1.00	–
Winter	0.25	1.29	0.87–1.91
<i>Activity</i>			
Gas welding	–0.19	0.82	0.45–1.51
Manual metal arc welding	0.16	1.18	0.72–1.91
Manual metal arc welding with LEV	0.50	1.65	1.03–2.63
Gouging	1.32	3.75	1.72–8.19
MAG welding	1.15	3.15	1.76–5.62
MAG welding with LEV	–1.03	0.36	0.14–0.93
MIG welding	0.55	1.73	0.86–3.48
Plasma cutting	–0.41	0.67	0.25–1.81
Grinding	0.40	1.50	1.05–2.14
Grinding with LEV	0.19	1.21	0.65–2.26
TIG welding	–0.21	0.81	0.38–1.76
TIG welding with LEV	0.17	1.19	0.54–2.60
Duration (per min)	0.0016	1.0016	1.0003–1.0028
<i>Environment</i>			
Inside a ship	0.65	1.91	1.14–3.20
Outside a ship	0.37	1.44	0.93–2.26
Outdoors	0.00	1.00	–
<i>Intercept (background)</i>	–2.01	0.13 mg m ^{–3}	0.07–0.25 mg m ^{–3}

Bold values indicate statistical significance at $P < 0.05$.

This also emphasizes the need to validate empirical exposure models. One can validate a model by setting aside some portion of the data (e.g. 20–30%, test subset) that is not used to construct exposure models. Subsequently, predictions of a model are compared to the test subset of data. If a model agrees well with the test subset of data, all measurements can be used to estimate the final model. Hornung (1991) provides formulae for evaluation of how well the exposure model fits the data:

$$\text{Bias} = \left[\sum_{i=1}^n (\hat{Y}_i - Y_i) \right] / n \quad (11.18)$$

$$\text{Precision} = \left\{ \sum_{i=1}^n [(\hat{Y}_i - Y_i) - \text{bias}]^2 / (n - 1) \right\}^{0.5} \quad (11.19);$$

where n = number of pairs of predicted (\hat{Y}_i) and measured (Y_i) values.

Bias is a measure of systematic deviation of the model from the data, and precision is a measure of random error in model predictions (i.e. standard deviation of bias). Plots of predicted versus measured values can also be very informative of the model fit, as under perfect fit we expect points (\hat{Y}_i, Y_i) to fall on to a straight line with slope of one and the intercept passing through the origin.

Controlling for all sources of variability in exposure

As we have already demonstrated in this chapter, levels of occupational exposure are influenced by a multitude of factors. Some of these factors are related to working environment itself (e.g. machines used, task performed, air movement through the workplace), whereas others are a

reflection of peculiarity in which each individual does their job (e.g. safety training, hand of preference, attitude towards risk-taking behaviour). In a sense, workers can be seen as creating their own unique work environment. Thus, in both design of exposure measurement surveys and in their analysis, it is important to distinguish between exposures that are unique to a given worker and exposures that are attributable to a job or equipment shared by several workers.

We have already demonstrated that it is essential to control systematic differences in exposure levels among workers by collecting repeated measurements on each worker. When considering only one group of workers, we have described how data with repeated exposure measurements could be analysed using ANOVA in order to reveal mean exposure of each individual, rather than the whole group. We have also described how regression analysis can be used to identify determinants of exposure, demonstrating how working environments differ in mean exposures. In such analyses, it is important to be sure that differences in exposure arising due to the use of two types of machines, for example, are not attributable to peculiarities of people who used these machines (Peretz *et al.*, 2002). To do this, we merge both ANOVA and multiple linear regression into a single statistical model.

Mixed effects models combine features of both ANOVA and multiple linear regression (e.g. Samuels *et al.*, 1985; Symanski *et al.*, 1996; Rappaport *et al.*, 1999). The name of this family of statistical models reflects the fact that they can estimate simultaneously the effects of factors for which we want to make predictions (fixed effects, e.g. machinery used) and the effects of nuisance factors that simply introduce noise into the data (random effects, e.g. worker identity). Let us consider statistical relationships between determinants of exposure (X) and exposure concentrations (Y), taking into account systematic differences in exposure among workers (z). The mixed effects model that is useful in solving this problem has the following general form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + b_1 z_1 + b_2 z_2 + \dots + \text{error.} \quad (11.20)$$

The part of the model ' $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$ ' represents the *fixed effects*, and this is analogous to multiple linear regression (where $\beta_0, \beta_1, \beta_2 \dots$ are regression coefficients); the part ' $b_1 z_1 + b_2 z_2 + \dots$ ' represents the *random effects*, and this is analogous to ANOVA, where ' b_i ' is the random effect of worker ' z_i ' (1/0 variable). The between-worker variability not explained by the model is reflected by differences among coefficients for random effects. Error term reflects the within-worker variability not explained by the model. Random effects can be estimated under different assumptions but the most common ones are that both between-worker and within-worker variability are (1) the same across all fixed effects (X) and (2) independent of each other. Mixed effects models can be easily estimated using a variety of statistical packages (e.g. PROC MIXED in Statistical Analysis System, SAS Institute, Cary, NC, USA); the estimating procedure is not fundamentally different from that used in estimating more familiar ANOVA and regression models. The majority of surveys that are carried out with the aim of identifying exposure controls and estimating exposure levels for a variety of purposes now use mixed effects models (as in the example of welders in the shipyard). Therefore, it is important for the occupational hygienist to be familiar with the ideas that these models are founded upon, in order to help them keep up with publications relevant to the professional practice of occupational hygiene.

Understanding exposure to mixtures

One of the fundamental problems in occupational hygiene is that exposures to pure substances rarely occur. More commonly, the workplace yields a complex mixture of chemicals that is potentially hazardous to workers. For example, numerous aromatic and aliphatic compounds are emitted during spray-painting. It is impossible to measure all individual compounds in such emissions. Two solutions exist for this problem: one is to measure representative constituents of the mixture that are deemed to be toxicologically relevant and the other is to assume that the composition of the

emissions is constant and to use a chemically non-specific method to characterize them (e.g. total volatile organic compounds). The first method can be wasteful if all individual constituents are highly correlated to each other. However, the second method can be based on incorrect assumptions, and thus miss important changes in composition of the mixture, which do not influence total amount of emissions.

A statistical technique called *factor analysis* can help occupational hygienists to (1) characterize the composition of mixtures and (2) identify factors that influence their composition. In order to apply factor analysis, we have to measure numerous constituents of a mixture. For example, for each sample we can determine multiple chemical agents representative of the range of chemicals likely to be present in the mixture.

In this application, factor analysis allows us to identify groups of chemicals that are correlated. This is accomplished by assuming that measured concentrations of chemicals (*manifest variables*) are correlated with measures of independent components of the mixture, which cannot be measured directly (*latent variables*). The observed correlation coefficients are used to make an inference about latent variables that they represent. For example, if concentrations of toluene, xylene and benzene are positively correlated, we can say that these three hydrocarbons (manifest variables) represent aromatic hydrocarbons in the mixture (latent variable). In the same study, concentrations of hexane, nonane and decane can also be correlated, signifying that these manifest variables represent aliphatic hydrocarbons in the mixture.

Therefore, factor analysis is nothing but an automated and systematic examination of correlation among constituents of a mixture, aimed at identifying independent sets of ingredients. In its simplest form, it identifies independent factors that explain the maximum amount of multiple correlations. Each one of these independent factors (latent variables) can be represented by a numerical score, which is a weighted sum of the manifest variables, with weights proportional to strength of the relationship between latent and manifest variables. Each j th factor's $[F(j)]$ eigenvector is the column of p weights $w_{(j)i}$ used to form the

factor from the observed variables $X_i \in (X_1, X_2 \dots X_p)$, such that:

$$F(j) = w_{(j)1} \times X_1 + w_{(j)2} \times X_2 + \dots + w_{(j)p} \times X_p \quad (11.21)$$

All of the factors are not correlated by definition and are selected to maximize the variability in multiple correlation that they explain. The simplest form of factor analysis used in occupational hygiene is called *principal component analysis*, with factors termed 'principal components'. Factor and principal component analysis can be fitted with most statistical packages, and has statistical properties that are well understood and explained in textbooks (e.g. Kleinbaum *et al.*, 1988).

Scores associated with each factor $[F(j)]$ can be used as a dependent variable in regression analysis in order to identify factors in the workplace that contribute to an increase in a particular constituent in a mixture, e.g. the use of solvent-based paints may be responsible for the increase in the proportion of aromatic hydrocarbons in emissions during spray-painting.

Examination of measured exposure that contributes to each factor also indicates which ingredients in the mixture should be measured in order to fully characterize the mixture. For example, it might be sufficient to measure only one aromatic and one aliphatic hydrocarbon in order to characterize exposure of painters to solvents. Such knowledge can result in a more efficient allocation of resources available to an occupational hygienist, for example by monitoring more different people on different days instead of analysing samples for a large number of chemicals. Thus, we recommend that factor analysis should be used as part of a pilot study that precedes development of any monitoring programme for mixed exposures.

Exposure surveys and models for risk assessment

The process of risk assessment for human health effects involves hazard identification and characterization, exposure assessment and risk characterization. Chapter 10 has already provided a detailed description of the various methods for

risk assessment purposes, and therefore only some general comments will be made here regarding the collection of exposure measurements and contextual data, as well as modelling of exposure levels in this process.

Risk assessments are generally carried out by government agencies for regulatory purposes. In addition, producers of chemicals or other products will often conduct risk assessments for scenarios of use of their products by their customers, as is the case for pesticides. Occupational hygienists are mostly concerned with collecting data for risk assessment within a company or factory. In such applications, the exposure measurement strategy and analytical approaches already discussed in this chapter generally apply. Sometimes, when dealing with exposures that have potential acute health effects, it may be necessary to adapt the previously described exposure measurements strategy. For example, in certain chemical process industries, high levels of exposure can occur during relatively rare and short activities, such as maintenance or cleaning work. Random allocation of samples may be inefficient in these situations, especially if the exposures during normal activities are very low. After identifying relevant determinants and circumstances of exposure, especially those that affect the within-day and day-to-day variance components, it should be possible to develop a task-based sampling strategy. Together with information on duration and frequency of these tasks, results of task-specific exposure intensity measurements can provide relevant inputs into the risk assessment process.

In cases of risk assessments carried out for regulatory purposes, exposure estimates are required for a large number of scenarios and sufficient exposure data will seldom be available. Similarly, in cases of risk assessment for new products, typically no field data are available. In such situations, exposure assessment is carried out using predictive models. The Health and Safety Executive in the UK has developed the EASE model for this purpose, based on the exposure data from the UK contained in the National Exposure Database (Friar, 1996). However, as the validity of this deterministic model has only been tested for a limited number of scenarios (Brendiek-Kämper, 2001; Kromhout,

2002) with relatively poor outcomes, the results of applying this model need to be interpreted with extreme care. In addition, this is a general model applicable to all workplaces and therefore does not include determinants of exposure for specific workplaces, activities or industries. Finally, the EASE model is designed for individual chemical products and does not deal with mixtures of exposure nor with exposures generated in the production process such as welding fumes, wood dust and the like. However, most occupational exposures occur in mixtures.

Others have advocated the use of probabilistic models for risk assessment purposes (e.g. van Drooge and van Haelst, 2001). To develop these models, information on determinants of exposure must be collected. The best way to do this is by analysing available exposure data using regression techniques described earlier in this chapter. Ideally, these models will also need to be validated using data that have not been used for the development of the model. Information on determinants of exposure (such as task, use of control measures, etc.) in certain exposure scenarios and measures of the associated uncertainty (e.g. range of task duration) can then be used to predict levels of exposure, just like in the example associated with Table 11.2. These predicted exposure levels could then be related to existing exposure–response relationships to determine the degree of risk.

In probabilistic models, rather than using point estimates of certain predictors of exposure, distributions of predictors of exposure are used (e.g. the distribution of duration of a certain task, the amount of substance used or the effectiveness of local exhaust ventilation). One can obtain a predicted distribution of exposure levels by repeatedly sampling from these distributions of exposure determinants and entering these into the models (equations). However, in even the best models of exposure, substantial unexplained variability in exposure concentrations will remain. This may produce considerable bias when applying probabilistic models to situations for which we have little or no exposure measurement data. As a result, occupational hygienists should follow a rule of thumb that deterministic and/or probabilistic modelling of exposure levels carries more

uncertainty and error than measuring exposure levels directly. It is often simpler and more cost-effective to measure rather than model exposures, especially when large capital investments for control measures are at stake.

As a consequence of this complication with generalizing predictions of exposure models, we would also like to emphasize that exposure measurements for risk assessment purposes should be carried out under realistic, rather than idealized, conditions. Thus, exposure assessment for risk assessment carried out in a plant that uses best available technology operated by well-trained staff may not be valid for a facility with poorly implemented exposure control measures where actual production or use takes place. This indicates that for informative risk assessment, exposure models and measurements should be aimed at accurately characterizing not just central tendency (as in compliance testing) but also the shape and width of exposure distributions. This implies that it may be necessary to oversample extreme (high and low) exposure situations in risk assessment.

Summary of recommendations

As we have seen in this chapter, the design of measurement strategies and the statistical analysis of exposure measurements resulting from them are no sinecure. Over the years, statistical models for describing the exposures of workers have evolved considerably. At present, anybody collecting exposure measurements should at least consider the following:

- conduct pilot studies that can help to plan long-term exposure surveys;
- collect repeated measurements on the same individuals;
 - enables estimation of between- and within-worker variability;
 - makes the estimation of both the probability of exceedance and overexposure possible;
- collect auxiliary information during the measurements;
 - enables the detection of factors affecting exposure (determinants of exposure) in the work environment

- enables the use of measurements' results for epidemiological purposes;
- gives direction for choice of exposure control measures.

Whenever possible, exposure measurement surveys should be designed in such a way as to be able to serve multiple purposes, such as the evaluation of compliance, exposure assessment in epidemiology, identification of effective exposure controls and risk assessment.

References

- Ashford, J.R. (1958). The design of a long-term sample programme to measure the hazard associated with an industrial environment. *Journal of the Royal Statistical Society*, 3 (Series A), 333–47.
- Brendiek-Kämper, S. (2001). Do EASE scenarios fit workplace reality? A validation study of the EASE model. *Applied Occupational Hygiene*, 16, 182–7.
- Burstyn, I. and Teschke, K. (1999). Studying the determinants of exposure: A review of methods. *American Industrial Hygiene Association Journal*, 60(1), 57–72.
- Burstyn, I., Bofetta, P., Kaupinnen, T., Heikkila, P., Svane, O., Partanen, T., Stucker, I., Frentzel-Beyme, R., Ahrens, W., Merzenich, H., Heederik, D., Hooiveld, M., Langard, S., Randem, B.G., Jarvholm, B., Bergdahl, I., Shaham, J., Ribak, J. and Kromhout, H. (2003). Estimating exposures in the asphalt industry for an international epidemiological cohort study of cancer risk. *American Journal of Industrial Medicine*, 43, 3–17.
- Devore, J.L. (1982). *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole, Monterey, CA.
- Friar, J.J. (1996). *The Assessment of Workplace Exposure to Substances Hazardous to Health. The EASE Model*. HSE, London.
- Hawkins, N.C. and Evans, J.S. (1989). Subjective estimation of toluene exposures: a calibration study of industrial hygienists. *Applied Occupational and Environmental Hygiene*, 4, 61–8.
- Hornung, R.W. (1991). Statistical evaluation of exposure assessment strategies. *Applied Occupational and Environmental Hygiene*, 6, 516–20.
- Hornung, R.W. and Reed, L.D. (1990). Estimation of average concentration in presence of non-detectable values. *Applied Occupational and Environmental Hygiene*, 5(1), 46–51.
- Kleinbaum, D.G., Kupper, L.L. and Muller, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods*. PWS-Kent, Boston.
- Kromhout, H. (2002). Design of measurement strategies for workplace exposures. *Occupational and Environmental Medicine*, 59, 286, 349–54.

- Kromhout, H. and Heederik, D. (1995). Occupational epidemiology in the rubber industry: implications of exposure variability. *American Journal of Industrial Medicine*, 27, 171–85.
- Kromhout, H. and Vermeulen, R. (2000). Long-term trends in occupational exposure: are they real? What causes them? What shall we do with them? *Annals of Occupational Hygiene*, 44, 325–7.
- Kromhout, H. and Vermeulen, R. (2001). Temporal, personal and spatial variability in dermal exposure. *Annals of Occupational Hygiene*, 45, 257–73.
- Kromhout, H., Oostendorp, Y., Heederik, D. and Boleij, J.S. (1987). Agreement between qualitative exposure estimates and quantitative exposure measurements. *American Journal of Industrial Medicine*, 12, 551–62.
- Kromhout, H., Symanski, E. and Rappaport, S.M. (1993). A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *Annals of Occupational Hygiene*, 37, 253–70.
- Kromhout, H., Loomis, D.P., Mihlan, G.J., Peipins, L.A., Kleckner, R.C., Iriye, R. and Savitz, D.A. (1995). Assessment and grouping of occupational magnetic field exposure in five electric utility companies. *Scandinavian Journal of Work, Environment and Health*, 21(1), 43–50.
- Kromhout, H., Tielemans, E., Preller, L., and Heederik, D., (1996). Estimates of individual dose from current exposure measurements. *Occupational Hygiene*, 3, 23–39.
- Lazovich, D., Murray, D.M., Brosseau, L.M., Parker, D.L., Milton, F.T. and Dugan, S.K. (2002a). Sample size considerations for studies of intervention efficacy in the occupational setting. *Annals of Occupational Hygiene*, 46, 219–27.
- Lazovich, D., Parker, D.L., Brosseau, L.M., Milton, F.T., Dugan, S.K., Pan, W. and Hock, L. (2002b). Effectiveness of a worksite intervention to reduce an occupational exposure: the Minnesota wood dust study. *American Journal of Public Health*, 92, 1498–505.
- Leidel, N.A., Busch, K.A. and Lynch, J.R. (1977). *Occupational Exposure Sampling Strategy Manual*. National Institute of Occupational Health and Safety, Cincinnati, OH.
- Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.
- Lyles, R.H., Kupper, L.L. and Rappaport, S.M. (1997). A lognormal distribution-based exposure assessment method for unbalanced data. *Annals of Occupational Hygiene*, 41(1), 63–76.
- Oldham, P.D. (1953) The nature of variability of dust concentrations at the coal face. *British Journal of Industrial Medicine*, 10, 227–34.
- Oldham, P.D. and Roach, S.A. (1952) A sampling procedure for measuring industrial dust exposure. *British Journal of Industrial Medicine*, 9, 112–19
- Peretz, C., Goren, A., Smid, T. and Kromhout, H. (2002). Application of mixed-effects models for exposure assessment. *Annals of Occupational Hygiene*, 46(1), 69–77.
- Post, W., Kromhout, H., Heederik, D., Noy, D. and Smit Duijzentkunst, R. (1991). Semiquantitative estimates of exposure to methylene chloride and styrene: the influence of quantitative exposure data. *Applied Occupational and Environmental Hygiene*, 6, 197–204.
- Prais, S.J. and Aitchison, J. (1954). The grouping of observations in regression analysis. *Journal of the International Statistics Institute*, 22, 1–22.
- Rappaport, S.M. (1991). Assessment of long-term exposures to toxic substances in air. *Annals of Occupational Hygiene*, 35(1), 61–121.
- Rappaport, S.M., Kromhout, H. and Symanski, E. (1993). Variation of exposure between workers in homogeneous exposure groups. *American Industrial Hygiene Association Journal*, 54, 654–62.
- Rappaport, S.M., Lyles, R.H. and Kupper, L.L. (1995). An exposure-assessments strategy accounting for within- and between-worker sources of variability. *Annals of Occupational Hygiene*, 39, 469–95.
- Rappaport, S.M., Weaver, M., Taylor, D., Kupper, L. and Susi, P. (1999). Application of mixed models to assess exposures monitored by construction workers during hot processes. *Annals of Occupational Hygiene*, 43, 457–69.
- Samuels, S.J., Lemasters, G.K. and Carson, A. (1985). Statistical methods for describing occupational exposure measurements. *American Industrial Hygiene Association Journal*, 46, 427–33.
- Sauleau E.A., Wild, P., Hours, M., Leplay, A. and Bergeret, A. (2003). Comparison of measurement strategies for prospective occupational epidemiology. *Annals of Occupational Hygiene*, 47, 101–10.
- Snedecor, G.W. and Cochran, W.G. (1989). *Statistical Methods*. Iowa State University Press, Ames, IA.
- Symanski, E., Kupper, L.L., Kromhout, H. and Rappaport, S.M. (1996). An investigation of systematic changes in occupational exposure. *American Industrial Hygiene Association Journal*, 57, 724–35.
- Taylor, D.J., Kupper, L.L., Rappaport, S.M. and Lyles, R.H. (2001). A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics*, 57, 681–8.
- Tielemans, E., Kupper, L.L., Kromhout, H., Heederik, D. and Houba, R. (1998). Individual-based and group-based occupational exposure assessment: some equations to evaluate different strategies. *Annals of Occupational Hygiene*, 42, 115–19.
- Topping, M. (2001). Occupational exposure limits for chemicals. *Occupational and Environmental Medicine*, 58, 138–44.
- Tornero-Velez, R., Symanski, E., Kromhout, H., Yu, R.C. and Rappaport, S.M. (1997). Compliance versus risk in assessing occupational exposures. *Risk Analysis*, 17, 279–92.
- van der Woord, M.P., Kromhout, H., Barregård, L., and Jonsson, P. (1999). Within-day variability of magnetic fields among electric utility workers: consequences for measurement strategies. *American Industrial Hygiene Association Journal*, 60, 713–19.

- van Drooge, H.L. and van Haelst, A.G. (2001). Probabilistic exposure assessment is essential for assessing risks – summary of discussions. *Annals of Occupational Hygiene*, **45**, S159–S162.
- van Tongeren, M.J.A., Gardiner, K., Calvert, I.A., Kromhout, H. and Harrington, J.M. (1997). Efficiency of different grouping schemes for dust exposure in the European carbon black respiratory morbidity study. *Occupational and Environmental Medicine*, **54**, 714–19.
- van Tongeren, M.J.A., Gardiner, K., Kromhout, H., Calvert, I.A and Harrington, J.M. (1999). An assessment of the sensitivity of estimating average exposure using various exposure grouping schemes on the relationship between exposure to dust and lung function parameters. *American Journal of Industrial Medicine*, **36**, 548–56.
- van Tongeren, M.J.A., Kromhout, H. and Gardiner, K. (2000) Trends in levels of inhalable dust exposure, exceedance and overexposure in the European carbon black manufacturing industry. *Annals of Occupational Hygiene*, **44**, 271–80.
- Vermeulen, R., de Hartog, J., Swuste, P. and Kromhout, H. (2000) Trends in exposure to inhalable particulate and dermal contamination in the rubber manufacturing industry: effectiveness of control measures implemented over a nine-year period. *Annals of Occupational Hygiene*, **44**, 343–54.

Chapter 12

Retrospective exposure assessment

Tom J. Smith, Patricia A. Stewart and Robert F. Herrick

Introduction

Basic concepts

- Dose and exposure indices for epidemiological studies

- Characteristics of historic job records

- Extrapolating past occupational exposures

 - Source–receptor model

 - Task-specific TWA model of exposures associated with a job title

- Characteristics of historic exposure measurement data

- Exposure groups – issues in exposure variability among workers

Approaches to retrospective exposure assessment

- Traditional exposure classification approaches

 - Estimates based on expert judgement

 - Semiquantitative expert estimates

- Quantitative estimates from calibrated expert judgement

- Approaches for cohort studies

- Grouping similar jobs together

- Estimation of mean exposures

- Extrapolation of exposure over time

- Statistical extrapolation approaches

- Extrapolation with deterministic physical exposure models

- Approaches for case–control studies

- Reliability and validity issues

- Summary and recommendations

- Acknowledgements

- References and suggested reading

- General reference

Introduction

In the investigation of disease in the workplace, one criterion for causality is the demonstration of an exposure–response relationship. This chapter focuses on the extrapolation of past exposures for studies of disease. Chronic occupational diseases, such as cancers and some lung diseases, develop over long time periods with exposure. Direct measurements of each subject’s exposures over the whole time period of interest would give the most accurate assessment of exposure for an epidemiological study. However, this ideal has not and is not likely to be achieved. Moreover, because of the long exposure or latency periods associated with these effects it is rare that there are exposure data covering the entire period. Consequently, epidemiological studies of these diseases require the estimation of past exposures by a process called ‘retrospective exposure assessment’.

With the development of risk assessment, there has been a strong need to develop exposure esti-

mates for exposure–response studies, a variety of approaches has been developed. Some of these are technically difficult to use and may be costly in time and resources. Without extensive personal exposure data it is not feasible to make individual exposure estimates. Exposure estimates are generally job based and personalized to some degree by using each subject’s personal job history. As will be shown below, there are a number of elements that must be present for detailed quantitative retrospective estimation to be feasible. These approaches range from the simplest separation of job titles into broad exposed and unexposed categories based on judgement to the most elaborate statistical models and estimation strategies that predict a unique quantitative exposure for every job held by the subjects. The approaches also vary with the type of epidemiological study. Regardless of the strategy, the common goal of retrospective exposure assessment is to develop the most accurate and unbiased estimates of exposure within the limitations of the resources.

Basic concepts

Dose and exposure indices for epidemiological studies

Our goal in epidemiological studies is to approximate the dose to the target tissue as closely as possible because it is the actual cause of the adverse effect observed in the epidemiological outcome. A *dose index* is a single number intended to summarize a part or all of a subject's exposure history that is aetiologically relevant to the risk of an adverse outcome, such as total dose of the suspected agent received by the subject (Smith, 1992). Even although a subject may be exposed for 20 or 30 years, all of that period may not be relevant to the risk of a disease, and the average of all daily exposures may not be relevant. The relevant duration and intensity depend on the mechanism of the effects. For example, lifetime exposure to a complete carcinogen is usually relevant, but only the 5 or 10 years before the onset of the disease are relevant for a cancer promoter. Likewise, exposure intensity may need to exceed a minimum level to cause some types of chronic effects, and exposures less than the minimum are irrelevant to risk; repeated brief, 'peak' exposures can cause chronic effects (see Chapter 14 for more on the effects of peaks). The epidemiological dose index that has been most widely used is the *cumulative exposure*, which is the mean exposure in a job times the duration in the job summed over all jobs held. This has been a useful measure in many studies of chronic disease from asbestos, lead, cadmium and other agents (Checkoway, 1986). Other dose indices may also be important for the risk of a particular disease, such as the occurrence of peak exposures. The choice of an optimum dose index depends on the mechanism of the disease (Smith, 1992). Whatever dose index is used, it is important to recognize that exposure is not equivalent to dose. Some writers have used exposure and dose as interchangeable terms and created considerable confusion as a result.

Characteristics of historic job records

Long-term exposures generally must be assigned to each subject, based on his or her *work history*

(a chronological listing of date started, job title and department or work site for each job held in a company). If an individual has worked for several companies then the work history from each should be obtained. The fundamental exposure assessment problem is one of converting job titles, department names and an industry name at a specific time period into exposure estimates (composition and intensity).

The *job title* is the most common way to assign an exposure to a subject in an epidemiological study of long-term effects. A job title usually has a defined set of work activities (*tasks*) an individual has to perform at one or more locations. These activities are specified by the needs of the industrial or commercial process. Unfortunately, job titles are not standardized and have little intrinsic meaning for exposure, and can vary among companies and can change across time when activities are reorganized. For example, a job title such as 'clerk', which is usually associated with low exposure, can be misleading because clerks can be located in production areas and be near emission sources. Conversely, the job title 'machine operator' may have little exposure because it refers to a Teletype operator. The task activities and work locations are determined by the nature of the manufacturing or commercial activity. However, the aggregation of those tasks and work locations under the definition of a job title is somewhat arbitrary and can vary across time, across plant site and across companies. As a result, the evaluation of a job's tasks and work locations is a critical part of retrospective exposure assessment.

An example of the tasks and work locations associated with the job title 'gasoline truck driver' are given in Table 12.1. Some tasks have high exposure potential, such as loading, and some have none, such as paperwork. The nature of the work site where the task is performed is also critical, such as delivering gasoline to large underground tanks with remote venting versus delivering gasoline to small tanks vented directly into the operator's breathing zone.

Tasks may require less than 1 min or several days to perform, and may be performed at a wide range of frequencies: many times per day or less than once per month. Tasks may also vary widely

Table 12.1 Example of tasks and work locations associated with the job title of gasoline truck driver.

<i>Work location</i>	<i>Task activity</i>	<i>Duration</i>	<i>Frequency</i>
<i>Truck cab</i>	Driving	5–60 min	2–12 per day
<i>Loading facility</i>	Loading truck tanks		
Top loading (no vapour control)		15–30 min	2–12 per day
Bottom loading (vapour recovery)		15–30 min	2–12 per day
<i>Customer tanks</i>	Delivery		
Underground tanks (remote venting)		10–20 min	2–6 per day
Above-ground tanks (vent in breathing zone)		5–15 min	2–10 per day
<i>Office/café</i>	Paperwork and breaks	10–45 min	2 or 3 per day

Note: a typical situation will involve a mix of these tasks.

in their exposure intensity. All of a job's tasks will contribute to the individual's long-term exposure, and all should be considered in a retrospective exposure assessment. Rare tasks may be difficult to assess but can be an important part of a job's exposures, such as cleaning the vinyl chloride polymerization tanks by chemical operators, who later developed angiosarcomas. Variability in both task frequency and exposure intensity over time is a fundamental characteristic of jobs.

Epidemiological studies require estimates of exposure for all job titles in the subjects' work histories. Where there have been exposure surveillance programmes, it is rare that all job titles and tasks with exposure potential have been characterized. Thus, estimates must be made for both current jobs that have not been sampled, and for past exposures prior to measurements.

Extrapolating past occupational exposures

An extrapolation rationale is needed that is compatible with the present and past data available to estimate exposures. The objective is to relate the composition and intensity of exposure to deterministic factors that can be evaluated or estimated from current and historic records. There are two simple paradigms that can guide this evaluation in a given situation:

- 1 the *source-receptor model* to describe the exposure process for a task; and
- 2 the *task-specific time-weighted average (task-TWA) exposure* to estimate a worker's overall average exposure by combining exposure esti-

mates for each task activity (task-TWA = average task exposure multiplied by duration of each task, summed over all tasks, divided by total time).

Evaluation of a job title in an epidemiological study can use both of these models to identify the deterministic factors associated with the worker, the tasks and the work environment that determine exposures. A partial listing of the task factors is given in Table 12.2. If there are historical changes in these factors then there are likely to be changes in the composition and/or intensity of exposures for a task. The task-TWA exposure calculation allows us to combine data on tasks associated with a job title and determine their contributions to the overall TWA exposure for the job. When calculated over longer time periods, this can also include infrequent, high-exposure tasks and changes in job description.

Source-receptor model

This model of an exposure is based on a concept borrowed from air pollution modelling, which describes the atmospheric transport of an air contaminant from an emission source to a receptor, e.g. an exposed population (see Chapter 30). In general, an industrial operation and its raw materials and products will define the sources, output strength and composition of airborne emissions (potential agents of effects). For example, a scrap brass refining operation requires scrap feed material, which may have some lead content, and the use of certain furnaces operated at defined temperatures over a specified production cycle, which in turn define the

Table 12.2 Exposure process model: factors affecting exposure in a task.

<i>Source</i>	<i>Transport</i>	<i>Worker</i>	<i>Setting</i>
Process	Air	Location relative to source(s)	<i>Physical</i>
Materials	Surface	Duration in area	Room size
Output rate	Radiative	Energy demands	Sources and locations
Worker influence		Work habits and techniques	General controls
Source controls		Personal exposure controls	<i>Management</i>
			Pressure for production
			Concern for health and safety

quantities of emissions of lead fumes. The intensity of exposure is determined by source output strength and configuration, air movements that transport and dilute the contaminants, effectiveness of ventilation and the worker's proximity, which is defined by his or her work task. Variability in the worker's exposure over time is a function of variability of source output composition and strength, variability of the transport processes (e.g. turbulent mixing), variability in the worker's position relative to the source and the effects of exposure controls at either the source or the worker (see Chapter 11). In many cases, the worker's job activities control or contribute to the source, e.g. welding or sweeping. Information about historical changes in materials, the process or the work site configuration can be used in the model to evaluate historical changes in exposure.

Although the potential complexity of the factors determining an exposure are daunting, it is not necessary to fully characterize the deterministic relationship for all possible factors to use this approach. Identification of the major factors and their likely effects on exposure can be carried out qualitatively by examining the schematic model for a given task and work location. Then the effects of changes in a factor, such as the addition of local exhaust ventilation (LEV), can be described by multipliers. For example, the ratio of mean exposures for a task before LEV and after LEV is installed gives a multiplier to estimate the effects of LEV in this setting without evaluating all of the component parts. It can also be argued that this ratio will apply to similar changes in other settings, which have not been measured. Thus, the multipliers associated with past changes may be estimated from existing

exposure measurements. Schneider and co-workers (1991) have developed this idea most extensively but others have also used it. Thus, the model approach provides an explicit powerful argument by analogy to estimate past operations that were never measured or characterized, if the emission sources and the exposure situation are similar to an exposure that has been measured. In most work settings, exposure is primarily determined by a small number of major factors, such as three or four of the following: the nature of the source (point, area), composition of emissions, emission strength, source isolation and emission controls, worker time spent in close proximity to the source, and worker task activities that may generate exposure.

One limitation of the source-receptor model is that it works best for a single task performed near a defined source. In some cases a job is associated with a single task, such as a 'packer', who loads mineral wool products off the end of a production line, or a 'data clerk', who works at a computer terminal all day. However, a job usually involves more than one task and the model does not provide guidance about how the tasks may be combined to give an overall estimate of the mean for the job title. The task-specific TWA model provides this link to estimate the mean by properly weighting samples collected during various single task activities or during mixtures of tasks.

Task-specific TWA model of exposures associated with a job title

The worker's job title and work location are the link between the exposure assessment and the epi-

demioleological evaluation. The task-TWA analysis provides two important insights: (1) it provides a method for extrapolating the exposure effects of historic changes in the tasks included under a job title and (2) it provides a method for appropriately weighting short-term samples collected to characterize tasks with high exposure potential. For example, an occupational hygienist may collect 10 samples, five during 'normal' activities and five during 'worst case' task activities with high exposure potential. If the high-exposure tasks represent only 10% of the total activity time, the simple average of these 10 samples will overestimate (bias) the estimate of the mean exposure for the job title.

The mathematical form of the task-specific TWA model is shown below:

$$\text{Task-TWA} = \frac{\sum_{i=1}^N X_i \times t_i}{\sum_{i=1}^N t_i}$$

The mean exposure for each task, X_i , is weighted by the total duration spent on the task, t_i , to obtain the correctly weighted mean for the job title.

It is important to note the differences in this model from the common time-weighted average exposure, which is measured directly in an 8-h personal sample. The expression above is intended to cover all of the time period variations in tasks and exposures that will occur for a job title, not just 8 h. Some tasks are very infrequent, such as a periodic 6-month cleaning of the vinyl chloride reactor, but these tasks may be very important contributors to health risk. Furthermore, as one considers increasing time spans in a chronic exposure study, changes may occur in the definition of a job when some tasks are excluded and new ones added. Changes in job title definitions have rarely been examined in epidemiological studies, but recent studies have identified important changes in component tasks and in time required to accomplish tasks (Smith *et al.*, 1993; Quinn, *et al.*, 2001). Effects of changes in process and production rate may increase or decrease task means.

The task weighting also provides a mechanism for utilizing occupational hygiene samples col-

lected to describe peak exposures and low-level area exposures without distorting the overall distribution. A limitation of the task-TWA approach is that only limited task data may be available. Another limitation is that usually only tasks with high exposure potential have been measured by hygienists. However, low-exposure tasks can frequently be estimated with area samples that describe 'background' exposure levels in a work area.

Characteristics of historic exposure measurement data

Exposures must have been measured at some point in time to anchor models for extrapolating past exposure. The estimation process is easiest and most accurate if there are measurements across time. Company surveillance and other data can be extremely useful, but they have several important limitations that must be addressed in building models or making exposure estimates:

1 Exposures to mixtures of chemicals were usually not characterized by historical sampling. Available data are often only rough indicators of exposure intensity for a specific agent. For example, total dust samples (mg m^{-3}) may be used to indicate the relative exposure intensity for a toxic component of the dust exposure. However, composition of the total dust and possibly its size distribution can change with work area, especially between high- and low-exposure areas. Assumptions about exposure composition should be stated clearly and carefully reviewed (see Chapter 11).

2 Historical exposures have been most often measured with older techniques specified by regulations, but these techniques are no longer used and may not be directly relevant to current hypotheses about health risks. A classic example of this is using impinger counts of total dust (particles $> 1 \mu\text{m}$) to characterize exposure for an effect caused by respirable crystalline silica dust (particles $< 3.5 \mu\text{m}$). A conversion factor may exist in some situations but the relationship is weak. In general, older methods are not similar to current approaches and technology. Studies may be needed to characterize the relationship between the old and new methods.

3 Historic samples were often fixed location or area samples. These cannot always be used to directly estimate personal exposure, but they may be useful when combined with time–activity data on jobs during periods when both area and personal data were collected. For example, in a study by Smith and co-workers (1980) of cadmium smelter workers, only area measurements were available for the 1940s to 1974, but there was an overlap of these samples with personal measurements for one year. The authors compared the area and personal measurements and used the ratio for each work area to adjust the area samples and make estimates of personal exposure.

4 Workers in some jobs may have used personal protective equipment, such as respirators. It is probably inappropriate to assume that this equipment dramatically reduces internal doses. Although laboratory studies have shown that well-fitted respirators may reduce inhalation exposure by 10-fold or more, under normal day-to-day usage the protection factors are much less (see Chapter 32).

5 The sampling strategy for determining compliance with standards requires the occupational hygienist to measure exposure on those days that include situations or tasks with the highest exposure potential (see Chapter 11). Consequently an unweighted mean of all full-shift samples for a job may give a biased estimate of overall mean exposure. Data on the time workers spend in various situations and doing specific tasks may be obtained by interview and used to weight the full-shift and task sampling data if the samples have adequate information on the sampling conditions.

Exposure groups – issues in exposure variability among workers

Estimation of individual exposures for all of the subjects in an epidemiological study is generally not possible because it requires measurements across time for each subject. Alternatively, some strategy for forming exposure groups must be used. The goal is large exposure differences between groups and small variability within groups. When that can be achieved it is more efficient than individual data and produces less attenuation in exposure–response relationships (van Tongeren *et al.*, 1997).

Epidemiological studies of long-term exposures have generally assumed that useful exposure groups can be formed by all individuals with the same job title, and have assigned the same mean exposure to everyone holding that job. This was based on the assumption that the primary source of variation in exposure samples was day-to-day variations in conditions experienced by all of the workers, and that averaging samples across time could control that variability. However, between-worker variability for individuals with the same job title has been found to be large in a number of settings (Rappaport *et al.*, 1993; Fig. 12.1). Many factors can cause systematic differences in exposure among workers with the same job title, such as differences in the subset of tasks performed during monitoring, training, technical skill, work habits, posture or body size. As the variability between workers' means increases, epidemiological exposure groups become more heterogeneous and statistical comparison of job groups by their overall sample means will underestimate the variability between groups, and overestimate differences between groups.

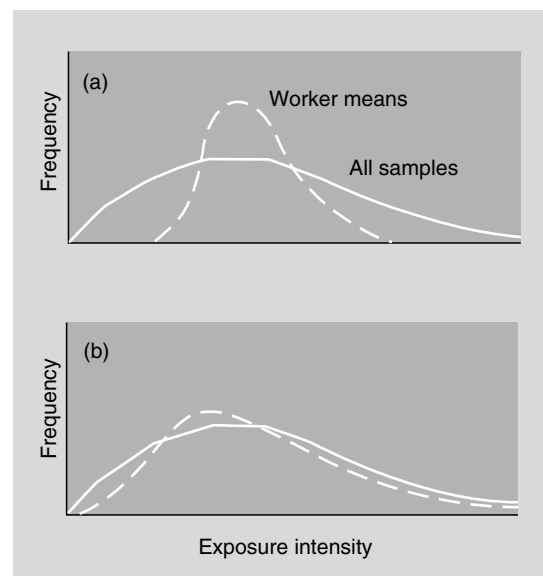


Figure 12.1 Hypothetical sampling distributions for two jobs with the same overall sample means but different variability in individual mean exposures: (a) low between-worker variability; (b) high between-worker variability.

There is little the occupational hygienist or epidemiologist can do to eliminate large between-worker variations in mean exposure. However, it can be an important source of misclassification and an explanation for the lack of apparent relationship between exposure and health effects. Where there are known large differences between workers, it may be necessary to personalize exposure estimates, such as through collection of interview data about differences in performing key high-exposure tasks.

Approaches to retrospective exposure assessment

The previous section developed the basic concepts needed to support the extrapolation process to estimate past exposures, both the composition and intensity. This section will take those concepts and illustrate their application to the common epidemiological study designs, industry- and plant-based cohort and population case-control studies, through the discussion of examples. We will emphasize the quantitative approaches because they are more useful for hypothesis testing and for development of dose-response relationships. However, we will begin by discussing the advantages and limitations of common traditional qualitative approaches, such as the 'ever/never' and duration of exposure classification schemes (Stewart and Herrick, 1991).

Traditional exposure classification approaches

In many broad epidemiological studies, the study subjects were classified by whether they had ever worked in a particular industry, or had a particular job title, which allowed the investigators to classify them into 'ever' worked in a job and 'never' worked in a job or industry. This type of classification may be highly accurate when ascertained from personnel records but, as noted above, employees within an industry or company or work area are likely to be exposed to a variety of chemicals at various levels, which often change over time. If this exposure misclassification is random or non-differential, it will result in a decrease in the

estimate of relative risk, and a causal association could be entirely missed. Even if an excess of some disease is identified, it is usually impossible to determine what workplace exposure may have been responsible for the excess without further studies to make a more detailed assessment of exposures.

As an attempt to sharpen the ever/never analysis, some occupational hygienists have asked study subjects whether they had exposures to specific substances used in an industry, company or job title. Broadly applied, this approach may achieve limited accuracy. However, it is unusual for everyone in a plant or job title to have exposure to a particular agent, and certainly they do not all have the same level of exposure. When applied in depth to a single plant with good records, better results may be obtained.

Another widely used surrogate for dose is duration of employment in an industry or a job, or duration of exposure to a particular agent. These have been widely used to investigate the existence of dose-response relationships when few exposure data were available. Duration has two major advantages: it is readily determined through a subject's work history obtained by interview or from the company records, and duration usually has reasonable accuracy. Duration may be a reasonable surrogate for cumulative exposure (intensity times duration), but only under certain conditions.

- 1 The exposure level is approximately the same for all workers in an exposure category.
- 2 Exposure levels have remained approximately the same over time.

Although these conditions are important, they are very difficult to verify and are commonly violated.

A study by Dement and co-workers (1983) illustrates a typical case in which these conditions were not met. Exposure monitoring data from a chrysotile asbestos plant were available back to 1930, and some examples of the range of estimated mean exposures for jobs within each department are presented in Table 12.3. Exposure levels varied widely both within and across the different departments. Grouping subjects who were 'exposed' in the fibre preparation operation with subjects who were 'exposed' in light weaving with the same

Table 12.3 Mean chrysotile asbestos levels (fibres per cm³) and range in an asbestos plant by department and time period*.

Department	1930	1936–39	1945–46	1965–66	1971–75
Fibre preparation and waste processing	26–78 [†]	– [‡]	8–24	6–17	–
Carding	11–13	5–11	2–5	4–9	–
Ring spinning	7–8	–	–	7–9	5–6
Mule spinning	5–7	–	–	–	–
Foster winding	10–21	4–8	–	–	–

*Abstracted from a report by Dement and co-workers (1983).

[†]The report contained means of specific jobs within departments; the range of these means is shown.

[‡]Dashes indicate no change from the earlier period.

duration (of employment or exposure) would result in misclassification of actual exposures and long-term doses. Table 12.3 also demonstrates that exposure levels may not remain static over time; if they change, they do not always drop and they do not remain at the same ranking relative to other jobs. Grouping subjects by duration alone can produce considerable misclassification.

The arguments presented here are not to suggest that analyses of ever/never exposure or duration of employment should not be performed. These measures can be useful in hypothesis-generating studies, particularly when using readily available records. Detailed retrospective exposure assessment may not be possible or may require more financial or time resources than are available to the investigator. Investigators should recognize, however, that relying upon ever/never or duration of employment as the sole measures of exposure will probably result in misclassification, which will decrease the probability of detecting true associations.

Estimates based on expert judgement

The limitations of the simple methods presented earlier for evaluating dose–response relationships have led some investigators to use a semiquantitative or quantitative approach based on expert judgement. Experts knowledgeable about the past conditions, such as plant hygienists, are used to create relative exposure categories, e.g. high, medium and low, based on their expert judgement, or quantitative estimates that have been ‘calibrated’ against limited measurement data. This type of analysis has been successful in finding associ-

ations, particularly in case–control studies. Unfortunately, few investigators have described in detail the procedures and rationale followed for the estimation of exposures.

Semiquantitative expert estimates

There are several drawbacks to using semiquantitative assessments without quantitative data. Dose analyses require that ranked jobs be assigned weights to allow analysis by cumulative exposure (sum of each exposure level times its duration). However, these weights are typically arbitrary, usually 1, 2 and 3, designating low, medium and high exposure levels. Such an assignment assumes that a job in the medium category has twice the exposure level as a job in the low category and two-thirds of the exposure level of a job in the high category. Other investigators have used geometric scales to quantify exposures. However, it is not known if these weights better reflect reality. In one study by Kromhout and co-workers (1987), air monitoring was conducted on various job tasks in five industries, and the sampling results were used to calculate an arithmetic mean for each task. These means were used to place each of the tasks into one out of four exposure categories and to derive an overall mean for the exposure category. Independent of the monitoring, two occupational hygienists classified the tasks into four exposure categories ranging from no exposure (1) to high exposure (4). This study suggests that using arbitrary weights provides less accurate weights than directly estimating exposure levels. More investigation, however, is needed in this area.

Semi-quantitative relative ranking of exposure has another limitation: the ranks may not be similar across sites and different companies. The relative ranks of job titles at a location may be appropriate at other locations, but the absolute levels may vary substantially so that in an extreme case the actual exposure of a 'high' job in one location may be equivalent to that in a 'low' job in another. Eisen and co-workers (1984) found that there was a fixed relative ranking of dust exposures for different job titles within granite sheds and large differences in absolute exposures for similar job titles across the sheds.

Quantitative estimates from calibrated expert judgement

Recently, investigators have explored hybrid schemes in which limited quantitative exposure data have been used to calibrate hygienist's professional judgement. Hawkins and Evans (1989) found that a reasonable accuracy could be achieved with limited data that was not possible without it for hygienists unfamiliar with a specific operation. Post and co-workers (1991) obtained similar findings. The hybrid approach makes good use of limited measurement data in combination with expert judgement.

Approaches for cohort studies

Industry-based cohort studies (cross-sectional, prospective or nested case-control) are designed to observe the disease experience of subjects chosen because they do or do not have an exposure of interest. These are studies in specific workplaces. With quantitative exposure estimates, they are useful for investigating exposure-response relationships. The exposed subjects are chosen because they all have a common exposure situation at a point in time, such as they all worked at a given company for a minimum of 1 year. As the focus is one or several companies, it is usually possible to obtain work history records from the companies or labour unions and other data on plant operations, materials used and changes over time. The companies also may have collected ex-

posure measurements through hygiene surveillance systems, which may be available.

In cohort studies the primary task is developing exposure estimates for all of the job titles in the work histories of the subjects. This problem can be separated into two components: exposures in jobs not measured during periods with measurements, and exposures in jobs in periods before there were measurements. Much more attention has been given to the second problem, for which a variety of approaches has been used.

Grouping similar jobs together

Lists of job titles abstracted from personnel records are commonly more extensive than needed to accomplish the plant operations. This occurs for a variety of personnel reasons unrelated to exposure. Consequently, many titles in the personnel records are synonymous with respect to exposure, and may be collapsed into a short list of 'generic' or standard titles based on an assessment of job activities, tasks and work locations associated with exposures (Quinn *et al.*, 2001). This can substantially reduce the list of unmeasured jobs. It will also ensure that the jobs grouped together have common features that affect exposure, such as working close to or distant from a local source of production emission. Broad grouping of jobs across industries can be done when the tasks and environment are judged sufficiently similar. This approach has been used to describe exposure zones (Corn and Esmen, 1979) or homogeneous exposure groups. For example, in a North Carolina study of dusty trade workers, all facilities within a commodity (several different companies) were considered to be similar enough to combine all measurements across those facilities (Rice *et al.*, 1984). Investigators have provided varying levels of detail in their justifications of their grouping schemes. If exposures are not similar within groups, combining jobs with heterogeneous exposures could result in considerable misclassification (Rappaport *et al.*, 1993). As the grouping scheme becomes broader and has less detail on specific job activities and work settings, then it becomes more likely that there are heterogeneous exposures within groups. When the data are limited,

directly estimating the between-worker variation within exposure groups may not be possible, although estimates by Kromhout *et al.* (1993) for broad types of work activities, settings and agents may provide some indication.

Estimation of mean exposures

When there are measurement data, the general approach is to assign everyone in a job group the simple mean of the exposure data. However, there are not usually equal amounts of data for every job group. Given the occupational hygienist's bias towards measuring when there are likely to be significant exposures relative to existing exposure standards, the unmeasured jobs are most commonly those with background exposures or those with little likelihood of production area exposures, such as office workers. Jobs in peripheral areas can be assigned observed background levels distant from sources or, based on the absence of sources, they may be assigned zero if they are in isolated areas.

Using historical measurement data to estimate TWA mean exposures can be limited by the methods used at the time. In some cases, many of the data are area measurements and/or short duration samples. The use of these data to directly represent full-shift personal samples is problematic. However, time-weighted models of exposure can be developed to use short-term or area samples to develop estimates that are representative of full-shift exposures; some investigators have weighed short-term monitoring results by time (Dement *et al.*, 1983; see Table 12.3). For example, 8-h TWA asbestos exposures were calculated for each job by summing the products of each zone's average exposure and the time spent in that zone for each task exposure in a job, and adding an increment in exposure based upon the tasks associated with each job, such as machine operator, clean-up and raw fibre handling (Dement *et al.*, 1983). This type of approach requires that the monitoring results be available for most tasks or areas in the study and that time spent in the zone or task can be estimated. Exposure effects of historic changes in job activities can be esti-

mated using the task-TWA approach as described earlier.

If there is an exposure that is parallel to the one of interest, measurements on that exposure may be used to predict the second exposure. This is only appropriate, however, when the relative level of exposure is expected to be the same for all of the jobs being estimated. This approach was used in an aluminium smelter that had benzene-soluble materials (BSM) measurements over the study period and benzo(a)pyrene (BaP) measurements after 1976 (Armstrong *et al.*, 1986). The authors derived a ratio using BaP and BSM measurements from 1976 to 1983 for 19 occupational groups. Assuming that the ratio remained the same over time, they applied this ratio to determine pre-1976 BSM levels.

In many situations, a single approach such as described above is insufficient to estimate exposures in all of the jobs because the assumptions needed are violated. It may be necessary to combine several approaches to make the estimates. For example, in a study of acrylonitrile workers, although there were 18 000 measurements available to the investigators, these measurements were from more recent time periods and for only small numbers of the 3500 jobs in the study. Because the type and number of data varied by job and year, a hierarchy of exposure assessment methods was developed using several estimation methods (Stewart, 1998). The methods included calculation of means based on personal monitoring results when they were available, using a ratio method when the ratio of exposures of some jobs is applied to other jobs, calculation of exposure means for homogeneous exposure groups using the measurements of all the jobs within the group, and use of the task-TWA approach. Criteria for using these methods, and a hierarchy of their use, were developed, based on their ability to predict the measurements. Each of the estimates was documented as to how it was derived and the assumptions made, which allowed reviewers to easily follow the decision-making process. A level of confidence was assigned to each estimate based on the availability of information and used to determine the effect of misclassification in the epidemiological analysis.

Extrapolation of exposure over time

Some investigators have attempted to develop an estimate for each year of exposure. Others have reduced the number of measurements necessary to specify an exposure by identifying time periods with no job or plant changes, within which it was assumed that exposures remained constant. The source–receptor model strongly implies that exposures may be reasonably assumed to remain stationary over time periods when evidence indicates that no changes in exposure determinants occurred. The method for determining these stationary time periods has varied. For example, in the North Carolina dusty trades study, the extensive silica monitoring results for each company were plotted by time and sample location (Rice *et al.*, 1984). Any point in time when all measurements were above or below all successive measurements was considered as evidence of a change in the workplace environment. Mean concentrations were then calculated for before and after the change. If the plot of measurements showed no such pattern, the mean concentrations from all the years were averaged. Other investigators have developed time periods based on changes in the workplace identified from interviews of workers or from engineering and other production reports (Smith *et al.*, 2001). Information on changes in exposure has also been developed with professional judgement to derive exposure levels (Armstrong *et al.*, 1986; 1986; Dodgson *et al.*, 1987).

Statistical extrapolation approaches

As noted above for silica exposure in North Carolina, statistical models based on available monitoring data can be used to predict exposure levels in unmeasured jobs and past exposures based on factors that define exposures, such as workroom size and ventilation. Measurements made across the jobs and years of a study can be used to make a model to complete the missing data cells. This approach has the advantage of being straightforward and it requires a minimum of assumptions.

Regression techniques have been used in several studies, for example a study of bitumen-paving workers (Burstyn *et al.*, 2000). Two problems may arise, however, in using statistical approaches. First, the exposure determinants (independent variables) in the statistical model need to be a small number relative to the number of samples collected. In a workplace with several hundreds to thousands of job titles, it is likely that most jobs will not have been monitored. Reduction of a large number of jobs to a small enough number for use in a statistical model may result in job categories with heterogeneous exposures, as noted earlier. A second problem is that models developed during periods with measurements are used to estimate exposures in earlier time periods when the conditions that created the model may not have existed. Unfortunately, this is a situation often encountered by investigators. In those situations, an unknown amount of misclassification may be present.

Extrapolation with deterministic physical exposure models

An approach with growing support is the deterministic model approach, which is closely related to the source–receptor model. In the deterministic model, the major physical factors controlling exposure are specified, an estimate of the multiplier associated with the factors is obtained from a standard model based on first principles, and the history of changes in the factors for a workplace is determined. Combining the history with recent measurements of exposure, the multipliers can be used to estimate past exposures, when the multipliers were estimated from measurement data before and after a change (Schneider *et al.*, 1991). Dodgson *et al.* (1987) estimated past man-made mineral fibre exposures using estimates of factors associated with the workplace changes. Uncertainties arise from the effects of workplace idiosyncrasies, such as placement of doors and windows, which may modify the effect of the standardized factors, such as emission rates. This approach has the advantage that it has a clear rationale for estimating past exposures because the multipliers are derived from first principles or experimental data.

Approaches for case-control studies

Population-based case-control studies are useful because they can investigate rare diseases which cannot practicably be studied in a reasonable-sized cohort. They also provide opportunities to study more efficiently the confounding and effect modification by non-occupational factors. The epidemiological analysis for case-control studies contrasts the histories of exposure for the patients who have the disease with those of the control subjects who do not to identify differences that might represent causal factors. Although this design is practical for the epidemiologist, it presents major difficulties to the exposure assessor because the subjects are chosen on the basis of their disease status, so they generally have highly varied work experience representing a wide variety of workplaces and job titles. Because the subjects are not drawn from a single workplace, it is usually not feasible to obtain company records of work histories. Job titles and other exposure-related data are gathered from hospital records, patient or surrogate interviews.

Broad job exposure matrices (JEMs) have been a preferred method for semiquantitative exposure assignments in case-control studies. A JEM is a cross-tabulation of job titles, or industries or job-industry combinations against a combination of general exposures and specific agents. For each job-agent cell in the table there is some type of exposure assignment, such as intensity, frequency or probability. The literature reviewed by the occupational hygienist to develop the exposure assignments is usually identified, but investigators have rarely described in detail how estimates of exposure have been derived. Because reports of occupational histories are prone to error, particularly when reported by next-of-kin, evaluating the quality of the reported information may help to identify occupational histories that are likely to contain errors.

The primary limitation of the JEM approach is that it is based on very limited data for each subject, which increases the likelihood of exposure misclassification. Use of most chemicals varies even within a single industrial workplace, and so few chemicals are found at every worksite within

the same industry. The probability that a person in a job is exposed varies with the process and its environmental characteristics, the chemical being assessed and the tasks being performed in the job. Asking direct questions of the respondent may allow a definitive evaluation of the probability exposure, but many workers do not know or cannot recall the substances that they used. If information still is not specific enough, the best estimate of the probability that exposure occurred may be based on the frequency of exposure in the population of workers holding the job in that industry. Such information could be derived from existing databases.

A major advance in assessing exposures in case-control studies has been described by Gerin *et al.* (1985). These investigators recognized that exposures are often idiosyncratic to the person holding the job, i.e. that everyone with the same job title does not necessarily have the same exposure. Gerin *et al.* therefore used information on job activities, equipment and materials used, and responses to occupation-specific questions for each individual study subject when assessing the exposures of each individual. Importantly, these questions were keyed to local industries and their histories of operations. This method substantially increases the accuracy of the assessments and has become increasingly popular.

Reliability and validity issues

The extrapolation of past exposures is not just a matter of having adequate measurement data to describe the composition and intensity of exposures. For a given study and exposure situation, a variety of data can be used, providing varying amounts and quality of information. Evaluation of past exposures requires the investigator to take advantage of all of these data sources and blend them into an overall picture of historical exposures. As a result, within a given study the quality of information about exposures, its reliability and validity can vary from job to job and over time periods.

Some researchers are uncomfortable with this approach and point out the probability of introdu-

cing errors when estimates are not based on actual measurements. This concern is valid and, undoubtedly, some estimates result in the misclassification of subjects. It is believed that the critical issue is not whether a quantitative approach results in misclassification but whether the misclassification is greater than it would have been using some other approach, i.e. ever/never exposed, or duration of employment versus imprecise quantitative estimates. The authors believe that evaluating each job or job task for its possible exposure level, taking into account the relative differences between jobs, is likely to ensure a better estimation of exposures and, therefore, less misclassification of subjects than other approaches.

Evaluation of the validity and reliability of the exposure estimates is important to the credibility of the study. Validity of estimates of past exposure are very difficult to determine because there are no data with which to check them. A few cohort studies have held some data aside to check the quality of estimates made with quantitative models (Dodgson *et al.*, 1987; Griefe *et al.*, 1988; Smith *et al.*, 1993). This has generally shown that the model approach produces reasonable estimates, generally within a factor of 2 of the measured exposures. However, it is not clear that the quality of the estimates for current or recent past exposures is equivalent to that of more distant past exposures. Even less work has been done on the validity of estimates for case-control studies. A few studies have examined intra- and inter-rater reliability, which have shown that moderate agreement can be obtained (case-control studies: Goldberg *et al.*, 1986; Hayes *et al.*, 1986; cohort study: Stewart *et al.*, 2000). Tielemans and co-workers (1999) compared the agreement of three methods for assessing exposure in a case-control study: self-reports, JEM and individual exposure measurements. They also compared the results to urinary biomarkers of exposure for chromium and several solvents. Overall, the agreement among the measures was poor (none of the kappa statistics exceeded 0.3). However, the personal exposure assessments were much better than the self-reports, and somewhat better than the JEM estimates.

Exposure misclassification is most likely to be non-differential in nature, i.e. the same proportion is misclassified in the diseased and non-diseased. Non-differential misclassification between adjacent exposure categories can have an attenuating effect on an exposure-response trend (Checkoway *et al.*, 1991). Even when the exposure misclassification rate is only 20%, the estimate of risk among the exposed can be substantially attenuated relative to the true relative risk. A misclassification rate larger than 20% would not be at all surprising in epidemiological studies, particularly when exposures must be estimated based upon a historical reconstruction for some members of a study population.

Summary and recommendations

Estimation of past exposures to potential health hazards is one of the most difficult problems for occupational hygiene research. Although it is very difficult, it is not impossible. It is important to recognize that the steps in extrapolation of past exposures have variable magnitudes of uncertainty. Quantitative estimates of exposure intensity for the distant past generally have the highest uncertainty. However, large uncertainty in the intensity of exposure does not mean that qualitative exposures are equally uncertain. It is important to consider a variety of exposure measures because less quantitative measures derived from information on the nature of operations and job activities can be very useful in some cases. Because a small amount of misclassification can substantially reduce the strength of an apparent exposure-response relationship, it is crucial to improve the quality of exposure assessments.

Acknowledgements

The authors wish to acknowledge the many contributions of others to the development of these ideas, especially our colleagues: Susan Woskie, Katharine Hammond, David Kriebel, Marilyn Hallock and Margaret Quinn.

References and suggested reading

- Armstrong, B.G., Tremblay, C.G., Cyr, D. and Theriault, G.P. (1986). Estimating the relationship between exposure to tar volatiles and the incidence of bladder cancer in aluminum smelter workers. *Scandinavian Journal of Work, Environment and Health*, **12**, 486–93.
- Burstyn, I., Kromhout, H., Kauppinen, T., Heikkila, P. and Boffetta, P. (2000). Statistical modelling of the determinants of historical exposure to bitumen and polycyclic aromatic hydrocarbons among paving workers. *Annals of Occupational Hygiene*, **44**, 43–56.
- Checkoway, H. (1986). Methods of treatment of exposure data in occupational epidemiology. *Medico Lavoro*, **1**, 48–71.
- Checkoway, H., Savitz, D.A. and Heyer, N.J. (1991). Assessing the effects of nondifferential misclassification of exposures in occupational studies. *Applied Occupational and Environmental Hygiene*, **6**, 528–33.
- Corn, M. and Esmen, N.A. (1979). Workplace exposure zones for classification of employee exposures to physical and chemical agents. *American Industrial Hygiene Association Journal*, **40**, 47–57.
- Dement, J.M., Harris, R.L., Symons, M.J. and Shy, C.M. (1983). Exposures and mortality among chrysotile asbestos workers. Part I: Exposure estimates. *American Industrial Hygiene Association Journal*, **4**, 399–419.
- Dodgson, J., Cherrie, J. and Groat, S. (1987). Estimates of past exposure to respirable man-made mineral fibers in the European insulation wool industry. *Annals of Occupational Hygiene*, **31**, 567–82.
- Eisen, E.A., Smith, T.J., Wegman, D.H., Louis, T.A. and Froines, J. (1984). Estimation of long term dust exposures in Vermont granite sheds. *American Industrial Hygiene Association Journal*, **45**, 89–94.
- Gerin, M., Siemiatycki, J., Kemper, H. and Begin, D. (1985). Obtaining occupational exposure histories in epidemiological case-control studies. *Journal of Occupational Medicine*, **27**, 420–6.
- Goldberg, M.S., Siemiatycki, J. and Gerin, M. (1986). Interrater agreement in assessing occupational exposure in a case-control study. *British Journal of Industrial Medicine*, **43**, 667–76.
- Griefe, A.L., Hornung, R.W., Stayner, L.G. and Steenland, K.N. (1988). Development of a model for use in estimating exposure to ethylene oxide in a retrospective cohort mortality study. *Scandinavian Journal of Work, Environment and Health*, **14** (Suppl. 1), 29–31.
- Hawkins, N.C. and Evans, J.S. (1989). Subjective estimation of toluene exposures: a calibration study of industrial hygienists. *Applied Industrial Hygiene*, **4**, 61–8.
- Hayes, R.B., Raatgever, J.W., deBruyn, A. and Gerin M. (1986). Cancer of the nasal cavity and paranasal sinuses, and formaldehyde exposure. *International Journal of Cancer*, **37**, 487–92.
- Kromhout, H., Oostendorp, Y., Heederik, D. and Boleij, J.S.M. (1987). Agreement between qualitative exposure estimates and quantitative exposure measurements. *American Journal of Industrial Medicine*, **12**, 551–62.
- Kromhout, H., Symanski, E. and Rappaport, S.M. (1993). A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *Annals of Occupational Hygiene*, **37**, 253–70.
- Post, W., Kromhout, H., Deederick, D., Noy, D. and Smit-Duijzentkunst, R. (1991). Semiquantitative estimates of exposure to methylene chloride and styrene: the influence of quantitative exposure data. *Applied Occupational and Environmental Hygiene*, **3**, 197–204.
- Quinn, M.M., Smith, T.J., Youk, A.O., Marsh, G.M., Stone, R.A. and Buchanich, J.M. (2001). Historical cohort study of US man-made vitreous fiber production workers: VIII. Exposure-specific job analysis. *Journal of Occupational and Environmental Medicine*, **43**, 824–34.
- Rappaport, S.M., Kromhout, H. and Symanski, E. (1993). Variation exposure between workers in homogeneous groups. *American Industrial Hygiene Association Journal*, **54**, 654–62.
- Rice, C., Harris, R.L., Lumsden, J.C. and Symons, M.J. (1984). Reconstruction of silica exposures in the North Carolina dusty trades. *American Industrial Hygiene Association Journal*, **45**, 689–96.
- Schneider, T., Olsen, I., Jorgensen, O. and Lauersen, B. (1991). Evaluation of exposure information. *Applied Occupational and Environmental Hygiene*, **6**, 475–81.
- Smith, T.J. (1992). Occupational exposure and dose over time: limitations of cumulative exposure. *American Journal of Industrial Medicine*, **21**, 35–51.
- Smith, T.J., Anderson, R.J. and Reading, J.C. (1980). Chronic cadmium exposures associated with kidney function effects. *American Journal of Industrial Medicine*, **1**, 319–37.
- Smith, T.J., Hammond, S.K. and Wong, O. (1993). Health effects of gasoline exposure: I. Exposure assessment for US distribution workers. *Environmental Health Perspectives*, **101** (Suppl. 6), 13–21.
- Smith, T.J., Smith, T.J., Quinn, M.M., Marsh, G.M., Youk, A.O., Stone, R.A., Buchanich, J.M. and Gula, M.J. (2001). Historical cohort study of US man-made vitreous fiber production workers: VII. Overview of the exposure assessment. *Journal of Occupational and Environmental Medicine*, **43**, 809–23.
- Stewart, P.A. and Herrick, R.F. (1991). Issues in performing retrospective exposure assessment. *Applied Occupational and Environmental Hygiene*, **6**, 421–7.
- Stewart, P.A., Zaebst, D., Zey, J.N., Herrick, R.F., Dosemeci, M., Hornung, R., Bloom, T., Pottern, L., Miller, B.A. and Blair, A. (1998). Exposure assessment for a study of workers exposed to acrylonitrile. *Scandinavian Journal of Work, Environment and Health*, **24**, 42–53.

- Stewart, P.A., Carel, R., Schairer, C. and Blair, A. (2000). Comparison of industrial hygienists' exposure evaluations for an epidemiologic study. *Scandinavian Journal of Work, Environment and Health*, **26**, 44–51.
- Tielemans, E., Heederik, D., Burdorf, A., Vermeulen, R., Veulemans, H., Kromhout, H. and Hartog, K. (1999). Assessment of occupational exposures in a general population: comparison of different methods. *Occupational Environmental Medicine*, **56**, 145–51.
- van Tongeren, M., Gardiner, K., Calvert, I., Kromhout, H. and Harrington, J.M. (1997). Efficiency of different

grouping schemes for dust exposure in the European carbon black respiratory morbidity study. *Occupational and Environmental Medicine*, **54**, 714–19.

General reference

- Armstrong, B.K., White, E. and Saracci, R. (1992). *Principles of Exposure Measurement in Epidemiology*. Oxford University Press, New York.

Chapter 13

Biological monitoring

Tar-Ching Aw

Introduction	
Definitions and terminology	
Biological monitoring	
Biological effect monitoring	
Health surveillance	
Indications for biological monitoring	
Types of biological samples	
Urine	
Blood	
Breath	
Hair and nail	
Fat	
Practical aspects	
Timing of sample collection	
	Selection of the correct container for the biological samples
	Contact with the laboratory
	Interpretation of results
	Reference values
	Specificity of metabolites
	Units of expression for results
	Interference by other chemicals
	Other practical, legal and ethical issues
	Training of staff to collect biological samples
	Notification of biological monitoring results
	Storage of biological monitoring results
	Conclusions
	References
	Further reading

Introduction

Biological monitoring is a process that is available in occupational health practice to provide a composite index of exposure and systemic absorption of chemicals in the workplace. It relies on the analysis of biological samples to indicate exposure from all routes – inhalation, ingestion and through the skin. Therefore, it complements ambient air monitoring as a tool for risk assessment and prevention of occupational ill health.

Definitions and terminology

Different definitions used for the term ‘biological monitoring’ can lead to confusion (Zielhuis, 1985). Some authors use the term to refer to any procedure used to monitor exposed workers, e.g. periodic radiographs, symptom enquiry or blood and urine tests. Others include tests indicating special effects, such as detecting the presence of DNA adducts in biological samples. The following definitions dis-

tinguish biological monitoring from biological effect monitoring and health surveillance.

Biological monitoring

This is ‘the measurement and assessment of workplace agents or their metabolites either in tissues, secretata, excreta, expired air or any combinations of these to evaluate exposure and health risk compared to an appropriate measure’. This definition restricts the term to the detection of chemical substances, or their breakdown products, in biological samples. It requires that there is an adequate and valid method for measurement, and that there is a means to decide on the extent of exposure and risk to health from the results obtained. Although this definition confines it to workplace agents, the methods and application may also be used for non-occupational environmental exposure to chemicals. Biological monitoring does not include detection of alterations in enzyme levels or other biochemical changes in such samples. This is covered by the term ‘biological effect monitoring’.

Biological effect monitoring

This term was proposed by Zielhuis and Henderson (1986) to refer to 'the measurement and assessment of early biological effects, of which the relationship to health impairment has not yet been established, in exposed workers to evaluate exposure and/or health risk compared to an appropriate reference'. The effect may not by itself be adverse to health but it would be an indication of a workplace agent causing some detectable biochemical alteration. Examples are the detection of free erythrocyte protoporphyrin (FEP) in blood, or δ -aminolaevulinic acid in urine (ALA-D), of workers exposed to inorganic lead, and β_2 -microglobulin (a low-molecular-weight protein) in the urine in cadmium-exposed individuals, and serum cholinesterase depression in workers exposed to organophosphate pesticides.

Health surveillance

This is the periodic physiological or clinical examination of exposed workers to detect early reversible health effects, so that measures can be taken to prevent occupational disease. Examples of physiological tests are spirometry and audiometry. Examples of clinical examination procedures include regular inspection of the skin and nose for chromate-exposed workers, and periodic review of respiratory symptoms for workers exposed to respiratory sensitizers.

Indications for biological monitoring

The main indications for biological monitoring are:

- 1 the occurrence of several routes of exposure and absorption of a chemical, e.g. organic solvents such as xylene can be absorbed by inhalation and through the skin;
- 2 the existence of valid laboratory methods for detecting the presence of the chemical or its metabolites;
- 3 the availability of reference values for interpreting the results obtained.

Biological monitoring can be used to confirm the efficacy of control measures, through demonstrat-

ing the absence or reduction of systemic absorption of a chemical. Biological monitoring is also useful when there are several sources of exposure to a chemical. These sources may be occupational and non-occupational. For example, in a painter working with paint stripper in a poorly ventilated garage warmed by use of a gas heater, blood carboxyhaemoglobin could be raised for several reasons. The poor ventilation, and therefore poor supply of oxygen, could lead to the inadequate combustion of carbon compounds, producing carbon monoxide in the garage. Methylene chloride in the paint stripper is metabolized to produce carboxyhaemoglobin. If the painter is a cigarette smoker, this will contribute to increased blood carboxyhaemoglobin. In this situation, environmental monitoring for carbon monoxide will not give a complete indication of the risk to health. Biological monitoring for carboxyhaemoglobin levels would be better.

Types of biological samples

A wide range of biological samples can, in theory, be obtained and analysed for the presence of workplace chemicals. However, for biological monitoring in occupational health practice, these samples are confined mainly to blood and urine, and possibly breath samples. Table 13.1 shows some examples of the types of biological samples that can be collected, and chemicals encountered in occupational settings and metabolites that can be identified in those samples.

Urine

Urine samples can be used to determine exposure by detecting the amount of the parent compound present, or the amount of metabolites of specific chemicals (Table 13.1). The main advantage of using urine samples is the ease of collection as the procedure is non-invasive. However, urine samples are open to external contamination, and clear instructions must be given to those participating in biological monitoring to minimize the likelihood of sample contamination.

Several factors have to be considered before deciding on using urine samples for biological

Table 13.1 Types of biological samples and analytes.

<i>Biological sample</i>	<i>Examples of parent compounds</i>	<i>Examples of metabolites</i>
Urine	Heavy metals, e.g. organic lead, mercury, cadmium, chromium, cobalt Metalloids, e.g. arsenic Ketones, e.g. acetone, methyl ethyl ketone, methyl isobutyl ketone Other compounds, e.g. fluorides, pentachlorophenol	Aromatic compounds, e.g. phenol (for benzene and phenol), hippuric acid (for toluene), methylhippuric acids (for xylenes), mandelic acid (for styrene and ethyl benzene) Chlorinated solvents, e.g. trichloroacetic acid (for trichloroethylene, perchloroethylene, 1,1,1-trichloroethane) Dialkylphosphates (for organophosphorous pesticides) 2,5-Hexanedione (for <i>n</i> -hexane)
Blood	Heavy metals, e.g. inorganic lead, mercury, cadmium, cobalt Aromatic compounds, e.g. toluene, styrene Chlorinated solvents, e.g. trichloroethylene, perchloroethylene, 1,1,1-trichloroethane	Carboxyhaemoglobin (for methylene chloride and carbon monoxide) Trichloroethanol (for trichloroethylene)
Breath	Aromatic compounds, e.g. benzene, toluene, ethyl benzene Chlorinated solvents, e.g. trichloroethylene, perchloroethylene, 1,1,1-trichloroethane, carbon tetrachloride, methylene chloride	
Hair and nail	Arsenic, mercury	
Fat	Polychlorinated biphenyls	

monitoring over other biological samples. These include the specific chemical of interest, whether it is organic or inorganic and different valency states if applicable, as well as the period and duration of exposure. The following examples illustrate some of these considerations.

1 For metallic mercury, urine is the biological sample of choice for assessing longer term (over a 3- to 4-week period) exposure. For recent, acute exposure (over 1 or 2 days), blood mercury gives a better indication of exposure.

2 In the case of exposure to inorganic lead, blood lead is preferred over urinary lead. However, for exposure to organic lead compounds the situation is reversed, with urinary lead regarded as a better index than blood lead.

3 With exposure to chromium compounds, metabolism results in chromium being excreted in the urine in the trivalent form, regardless of whether exposure and absorption includes hexa-

valent as well as trivalent forms. Hence, total chromium in the urine will not indicate the relative exposures to different species of chromium compounds of different valency states.

Blood

Blood samples can also be collected to quantify parent compounds or their metabolites (Table 13.1). The disadvantages of using blood samples for biological monitoring include possible poorer compliance because of the discomfort of the procedure, especially if venepuncture is required on a regular basis. Where participants have veins that are not prominent, a degree of skill and experience is required by the phlebotomist to avoid creating a haematoma that can discourage future participation in biological monitoring.

As with urine samples, prevention of external contamination of the blood sample is important.

This is especially true if the sample is to be analysed for the parent compound. The site of venepuncture has to be adequately cleaned. Also, all precautions should be taken by the phlebotomist to minimize the risk of acquiring blood-borne infections. Practitioners who are regularly involved in collecting blood samples for biological monitoring should be fully immunized against hepatitis B, and comply with universal precautions in handling blood samples. There should be adequate provisions for the proper disposal of all equipment used for collection of blood samples.

Breath

The model for breath sample analysis is the detection of breath alcohol by use of a breathalyser. The same principles can be used for detecting and quantifying other volatile organic compounds that are excreted in the breath. The process involves either breathing out into a direct reading instrument or into a glass pipette, aluminium tube, sampling bag or other collection device before dispatching the collected sample to a laboratory for analysis. The advantages of breath analysis are that the collection of samples is non-invasive, and repeated samples can be taken within a short period of time. The disadvantages include current lack of agreement on how samples have to be collected, standardization on the type of collecting device used and laboratory methods for analysis. Analytical instruments for breath samples that can be used at the workplace or for field studies are not widely available.

The American Conference of Governmental Industrial Hygienists (ACGIH) distinguishes between mixed-exhaled air and end-exhaled air, and refers to the difference between these two types of breath samples in amounts of solvent detected during exposure and after exposure. Those with impaired lung function may not provide a suitable breath sample for such analysis. Cigarette smoking and endogenous compounds from dietary sources can also affect the breath levels of some compounds. Acetaldehyde concentrations in breath are higher in smokers than in non-smokers. Ammonia in breath can be derived from protein metabolism.

Hair and nail

Analysis of hair and nail samples has been used to assess exposure to arsenic and mercury. However, this has been used more for forensic purposes than in occupational health practice for biological monitoring. The limitations include cost and consistency of the analysis. External contamination of the hair and nail may give an erroneous indication of the amount absorbed systemically. Contamination may occur from occupational and non-occupational activity, and procedures to wash and clean the samples before analysis may not adequately remove contaminants adsorbed onto the surface of the samples. The differences between amounts in head and pubic hair or in fingernail and toenail samples may indicate the extent of surface contamination. There are also limitations in the interpretation of the results obtained, particularly for individuals rather than groups.

Fat

Samples of body fat have been used for determining the extent of exposure to polychlorinated biphenyls (PCBs). PCBs are present in industrial transformers and have been documented as causing chloracne and liver damage. Besides the toxicity of PCBs and their contaminants, they are not easily biodegradable and therefore persist in the environment, as well as in adipose tissue. Fat samples are obtained by needle biopsy or by surgical excision. The quantity of fat required for the assay of PCB content is several hundred grams. The amount of PCB is measured in parts per billion or parts per trillion quantities. Any contamination or error will cause a considerable difference in the estimate from the true value. It is not practical to collect fat samples periodically; nor is the procedure likely to attract many volunteers. This method, therefore, has considerable limitations for practical biological monitoring.

Practical aspects

Timing of sample collection

For some substances, especially those with long half-lives, the timing of collection of the biological

Table 13.2 Reference values for biological monitoring: examples for blood, urine and breath samples.

<i>Chemical</i>	<i>HSE</i>	<i>ACGIH</i>	<i>DFG</i>
Blood lead	25 µg per 100 ml for women of reproductive age; 40 µg per 100 ml for those aged < 18 years; and 50 µg per 100 ml for all others	30 µg per 100 ml	30 µg per 100 ml for women < 45 years old; 40 µg per 100 ml for others
Blood carboxyhaemoglobin for carbon monoxide exposure	—*	3.5% of haemoglobin at end of shift	5% at end of exposure or end of shift
Urine inorganic mercury	20 µmol mol ⁻¹ creatinine	35 µg g ⁻¹ creatinine for pre-shift sample	100 µg l ⁻¹
Urine 4,4'-methylene dianiline	50 µmol mol ⁻¹ creatinine	—*	—*
Urine mandelic acid for styrene exposure	—*	800 mg g ⁻¹ creatinine for end-of-shift sample; or	600 mg g ⁻¹ creatinine at end of exposure or end of shift (urinary mandelic acid plus phenylglyoxylic acid)
Tetrachloroethylene in end-exhaled air	—*	5 ppm for sample prior to last shift of work week	300 mg g ⁻¹ creatinine for sample taken prior to next shift
Carbon monoxide in end-exhaled air	30 ppm (post-shift sample)	20 ppm for end-of-shift sample	—*

*Dashes indicate that no specific reference value has been specified.

For a complete up-to-date list of values for other substances, refer to the most recent publications from Health and Safety Executive (HSE), American Conference of Governmental Industrial Hygienists (ACGIH) Inc. and Deutsche Forschungsgemeinschaft (DFG).

sample is not critical. This applies to blood lead and urinary cadmium. The half-life of a substance refers to the time required for clearance of 50% of the substance from the medium. Examples of chemicals for which the timing of sample collection during the working week is critical are shown in Table 13.2. The reference values for interpreting the results take into account when samples are collected.

Selection of the correct container for the biological samples

Precautions have to be taken to use a suitable container for the biological sample. The container must be able to hold a sufficient amount of the biological sample needed for analysis, and this

will be indicated by the laboratory. For urinalysis, 20–25 ml of urine is usually adequate for a spot sample; 24-h urine samples are seldom used for occupational health purposes because of the logistics of obtaining such samples for individuals at work. As little as 5 ml of blood may be sufficient for blood samples for most compounds. For breath samples, the volume collected depends upon the device provided by the laboratory. Most laboratories provide or recommend appropriate containers to use for biological samples.

Chemicals such as mercury can be adsorbed onto the surface of some polypropylene or polyethylene containers and this could result in the detection of an artificially lower urinary mercury level. For such chemicals, the use of a glass container would be preferable to a plastic bottle.

Chemical plasticizers used for plastic containers or bottle caps can interfere with the analysis for PCBs. Hence, in the collection of blood or adipose tissue samples for determining PCB levels, plastic receptacles are best avoided. Because the quantities of PCBs determined are small, eliminating any interference with the analysis is critical to avoid considerable errors in the results.

If serum samples are required, no anticoagulant is needed. If whole blood is required, then heparin or ethylenediamine tetra-acetic acid (EDTA) should be present as an anticoagulant. However, as EDTA chelates metals, it would not be appropriate, for example, for determining chromium levels in red blood cells. Preservatives are used for urine samples if there is likely to be some delay between collection and delivery to the laboratory for subsequent analysis.

Contact with the laboratory

The choice of a suitable laboratory for analysis of biological samples is essential for obtaining valid results. Laboratories should have sufficient experience in analysing samples from occupational health sources, they should belong to a quality control scheme and should have a quality assurance programme. In the USA, a list of laboratories for blood lead determinations is available from the Occupational Health and Safety Administration (OSHA). In the UK, the Health and Safety Laboratory of the Health and Safety Executive (HSE), independent toxicology laboratories and those within the National Health Service (NHS) provide analysis of biological samples for a range of chemicals. Advice from the laboratory should be obtained on the quantity of the sample required, whether any special precautions are needed and how samples should be stored and delivered to the laboratory. Certain samples have to be kept at a low temperature and dispatched to the laboratory as soon as possible after collection. All samples should be adequately and securely labelled and packed. A large 'pooled' sample of urine due to breakage from poor packing, or a number of labels separated from their samples, is to be avoided.

Interpretation of results

Results from biological monitoring of similarly exposed workers are subject to intra- and inter-subject variation. Factors to consider in the interpretation of results include those related to:

- 1 the *individual* (age, sex, body mass index, genetic differences in the metabolism of compounds, pregnancy state, exercise and physical activity, smoking, medication, consumption of alcohol and other dietary factors, and the presence of any existing lung, liver or kidney disease or other illness);
- 2 the *exposure* (timing and intensity of exposure in relation to timing of collection of the biological sample, mixed exposures that may affect the metabolism of the compounds absorbed, and routes of exposure);
- 3 the *chemical of interest* (and its biological half-life, and where and how it is metabolized and excreted).

Reference values

In the UK, occupational exposure limits for airborne substances are published annually in an HSE document, EH40 (Health and Safety Executive, 2002). There is a section within this publication that provides biological monitoring guidance values for a limited number of chemicals. These values are categorized into (1) health guidance values that are set on the basis of available scientific evidence indicating no adverse effects on health and (2) benchmark guidance values, which, although not based on health effects, are set at the 90th percentile of biological monitoring values obtained from workplaces with good occupational hygiene practice. The Health and Safety Laboratory of the HSE also has an internal handbook with some additional information on interpreting biological monitoring results. Scientific staff of the laboratory and employment medical advisers can be consulted for advice on interpretation of biological monitoring data. HSE Guidance Notes (Medical Series) are another useful reference source. Other analytical laboratories may provide some guidance on the interpretation of the results.

In the USA, biological exposure indices (BEIs) are published annually by the ACGIH. BEIs are described as representing ‘the levels of determinants which are most likely to be observed in specimens collected from healthy workers who have been exposed to chemicals to the same extent as workers with inhalation exposure at the threshold limit values (TLV)’. They provide an indicator of exposure, and are not meant to be used to measure adverse health effects or to diagnose occupational disease.

A list of BEIs appears together with TLVs in the annual ACGIH booklet on TLVs and BEIs (American Conference of Governmental Industrial Hygienists, 2002). The 2002 edition of this reference includes BEIs for around 40 different chemicals or groups of chemicals. For each substance there is an indication of when the sample should be collected, whether the parent compound or the metabolite should be determined and what the BEI is. Where applicable, notations are also provided to indicate increased susceptibility for some individuals, the presence of background levels of some determinants in non-occupationally exposed groups, non-specificity of some findings and the semiquantitative nature of some substances measured. The ACGIH, like the HSE in the UK, also indicates substances for which there is an intention to establish or change the indices. The German Commission for the investigation of health hazards of chemical compounds in the work area (Deutsche Forschungsgemeinschaft, 2001) promulgates maximum permissible concentrations for airborne chemicals in the workplace (MAK values) and biological tolerance values (BAT values) for biological monitoring. BAT values are defined as the maximum permissible quantity of a chemical compound, its metabolites or any deviation from the norm of biological parameters induced by these substances in exposed humans. Nearly 50 different values have been published providing reference standards for both biological monitoring and biological effect monitoring. Table 13.2 provides a comparison of some values published by different organizations. It demonstrates the variation in values, units of expression and timing of sample collection. The philosophy behind the setting

of reference values for each organization has to be appreciated in order to use the values appropriately.

Specificity of metabolites

In the interpretation of the results from biological monitoring, it is essential to be aware of whether the compound detected in the biological sample is specific to the exposure of concern or whether it may result from several sources of exposure. Some compounds produce one main specific metabolite of importance, e.g. xylene is metabolized to methylhippuric acid in the urine, and this is not derived from other sources. However, there are three different isomers of xylene (*ortho*-, *meta*- and *para*-) and it is possible to detect three related isomers of methylhippuric acid in urine from exposure to a mixture of isomers of xylene. Other compounds produce two or more metabolites, for example styrene is metabolized to mandelic acid and phenylglyoxylic acid. Phenylglyoxylic acid is also a metabolite of ethyl benzene. Hippuric acid in the urine can result from exposure to toluene and also from ingestion of foods and drinks that contain benzoic acid as a preservative. Both trichloroethylene and perchloroethylene (tetrachloroethylene) exposure produces the metabolite trichloroacetic acid in the urine.

Units of expression for results

The use of different units for expressing the results from biological monitoring can add confusion to the interpretation of results. Blood results are normally expressed in micrograms (μg) or milligrams (mg) per litre (l). Urinary levels of compounds are often expressed in milligrams per unit volume of urine, which has the disadvantage of being affected by urinary concentration or dilution. Expression of urine results in terms of the specific gravity of the urine sample has been suggested. However, creatinine correction is now used most widely to overcome urine concentration/dilution limitations. Creatinine is a protein that is excreted in the urine fairly independently of urine concentration or dilution and is relatively constant at

15–20 mg kg⁻¹ for females and 20–25 mg kg⁻¹ for males. Biological monitoring reference values for chemicals or their metabolites in urine are now mainly expressed in terms of milligrams per gram (mg g⁻¹) of creatinine, or millimoles per millimole (SI units) of creatinine. As a rough guide for converting millimoles per millimole (mmol mmol⁻¹) of creatinine to milligrams per litre (mg l⁻¹), the following formula can be used:

$$(\text{amount in mmol}^{-1} \text{ creatinine}) \times (\text{molecular weight of the substance}) \times 10 = \text{amount in mg l}^{-1}$$

This is based on the assumption that 1 l of urine contains about 10 mmol of creatinine.

Interference by other chemicals

Consumption of alcoholic beverages can interfere with the biological monitoring for organic solvents. In addition to causing additive, synergistic or antagonistic clinical effects, ethanol may delay the rate of metabolism of compounds by competing for liver enzymes such as dehydrogenases, catalase and mixed-function oxidases. The competition for acetaldehyde dehydrogenase between ethanol and trichloroethylene delays the metabolism of trichloroethylene and causes the clinical phenomenon ‘degreaser’s flush’ which is seen in trichloroethylene-exposed workers who have recently consumed alcohol. Competition by ethanol for liver enzymes can also increase the blood level of xylene, toluene and trichloroethylene, and decrease the amount of urinary metabolites of these solvents.

When several metabolites are produced from exposure to one chemical, ethanol may have a differential effect on the rate of production of these metabolites. Alcohol ingestion with concomitant styrene exposure leads to a greater reduction in mandelic acid than in phenylglyoxylic acid (both are metabolites of styrene) in blood and urine. Poorly metabolized compounds, such as perchloroethylene and 1,1,1-trichloroethane, are less affected by alcohol ingestion.

Regular alcohol consumption can cause non-specific induction of microsomal enzymes in the liver, leading to speeding up of the metabolism of organic solvents such as styrene. In such situations, the timing of collection of biological samples in

relation to the period of exposure can be critical in interpreting the results of biological monitoring. Because of the variation in the nature and extent of the effect of alcohol on the metabolism of chemicals, it is preferable that alcohol consumption is avoided on days when biological monitoring for organic solvents is to be performed.

Cigarette smoking can affect enzymatic activity in lung parenchymal cells and in alveolar macrophages, and may alter the rate of metabolism of inhaled compounds. This may explain the difference in acetaldehyde levels in the breath in exposed smokers compared with non-smokers.

Aspirin consumption can affect the metabolism of xylene. It reduces methylhippuric acid concentration in the urine by competing for glycine. This amino acid is essential for conjugation in the metabolism of both aspirin and xylene. Other medications, such as barbiturates, are known to induce liver enzymes, and can alter the rate of hepatic breakdown of absorbed chemicals.

Other practical, legal and ethical issues

Training of staff to collect biological samples

Occupational health practitioners who have to carry out biological monitoring should have adequate training in the process and precautions necessary for obtaining valid samples and interpreting the results. This includes the need to explain the reasons for the procedure, indicate what tests will be done on the samples and reassure the workers on what tests will not be performed, e.g. HIV testing on blood samples. In the UK, biological monitoring is usually performed by an occupational physician or occupational health nurse. Occupational hygienists and other occupational health practitioners with relevant training can acquire the necessary competence to participate in biological monitoring. However, it is essential to recognize when outside assistance may be required.

Notification of biological monitoring results

The individual who has been tested is entitled to his or her own results with an explanation of what

they mean; this is the responsibility of the occupational health professional who is carrying out the monitoring. With the patient's consent, the results should also be communicated to the family physician for inclusion into the health records, as the patient may raise this with his or her doctor on subsequent visits to the surgery. Management and unions can be provided with grouped data, with the proviso that confidentiality is maintained by the removal of specific identifiers from the grouped data. Feedback should also be provided to other occupational health and safety practitioners and, when required, to the appropriate regulatory agencies.

Storage of biological monitoring results

The UK Control of Substances Hazardous to Health (COSHH) regulations require that when biological monitoring is performed, the results are kept for at least 40 years from the date of last entry. If such records are properly collected, recorded and stored they can be of value in future epidemiological studies. When companies cease operations, they are advised to offer the collected results of biological and environmental monitoring to the Health and Safety Executive for safekeeping.

Conclusions

Biological monitoring will be used to a greater extent for assessing occupational exposure to chemicals as new methods are developed for detecting and quantifying substances or metabolites in biological samples. Advances in laboratory technology will allow smaller quantities of chemicals to be detected in biological samples with greater sensitivity and specificity. Reference values will have to be agreed to allow clear and uniform interpretation of results. Recognition of the limitations of the process and the many factors that can affect the results is essential. More occupational health practitioners may endeavour to carry out biological monitoring to complement environmental monitoring, but the precautions needed and the skills required have to be considered. In the future, the techniques of biological monitoring may be

extended from occupational exposures to non-occupational environmental exposures. This would aid in risk assessment in occupational and environmental settings, and would be a useful tool for the prevention of occupational and environmental ill health from chemical exposures.

References

- American Conference of Governmental Industrial Hygienists (2002). *2002 Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices*. ACGIH, Cincinnati, OH.
- Deutsche Forschungsgemeinschaft (2001). *List of MAK and BAT Values 2002*. Commission for the Investigation of Health Hazards of Chemical Compounds in the Work Area, report no. 37. Wiley-VCH, Weinheim.
- Health and Safety Executive (2002). *Occupational Exposure Limits 2002*, (EH40/2002). HSE Books, Sudbury.
- Zielhuis, R.L. (1985). Biological monitoring: confusion in terminology (Editorial). *American Journal of Industrial Medicine*, 8, 515–16.
- Zielhuis, R.L. and Henderson, P.T. (1986). Definitions of monitoring activities and their relevance for the practice of occupational health. *International Archives of Occupational and Environmental Health*, 57, 249–57.

Further reading

- Aw, T.C. (1999). Health surveillance. In *Occupational Health: Risk Assessment and Management* (eds S.S. Sadhra and K.G. Rampal). Blackwell Science, Oxford.
- Borak, J., Sirianni, G., Cohen, H., Chemerynski, S. and Jongeneelen, F. (2002). Biological versus ambient exposure monitoring of creosote facility workers. *Journal of Occupational and Environmental Medicine*, 44: 310–19.
- Health and Safety Executive (1997). *Biological Monitoring in the Workplace, a Guide to its Practical Application to Chemical Exposure*, HSG 167. HSE Books, Sudbury.
- Lowry, L.K. (1986). Biological exposure index as a complement to the TLV. *Journal of Occupational Medicine*, 28, 578–82.
- Mraz, J., Galova, E., Nohova, H., Vitkova, D. and Tichy, M. (1999). Effect of ethanol on the urinary excretion of cyclohexanol and cyclohexanediols, biomarkers of the exposure to cyclohexanone, cyclohexane and cyclohexanol in humans. *Scandinavian Journal of Work, Environment and Health* 25: 233–7.
- Notten, W.R.F., Herber, R.F.M., Hunter, W.J., Monster, A.C. and Zielhuis, R.L. (eds) (1988). *Health Surveillance of Individual Workers Exposed to Chemical Agents*. Springer-Verlag, Berlin.

Oliveira, G.H., Henderson, J.D., and Wilson, B.W. (2002). Cholinesterase measurements with an automated kit. *American Journal of Industrial Medicine*, 2 (Suppl.): 49–53.

Wilson, H.K. and Baxter, P.J. (2000). The role of laboratory techniques in the prevention and diagnosis of occupational and environmental diseases. In *Hunter's Diseases of Occupations*, 9th edn (eds P.J. Baxter, P.H. Adams, T-C. Aw, A. Cockcroft and J.M. Harrington). Arnold, London.

Chapter 14

Epidemiology

J. Malcolm Harrington

Introduction

- Definition
- Uses

Sources of data

- National records (vital statistics)
 - Death certificates
 - Birth certificates
- Morbidity
- Local records
- Ad hoc records

Measures of exposure and health outcome

- Exposure
- Health outcome
 - Measures of occurrence
 - Measures of frequency
 - The relevance of time, place and person
 - Measures of risk

Causation or association

Study design

Goals

- Validity
- Precision
- Cost

Options

- Timing
- Cross-sectional studies
- Longitudinal studies
- Control of confounding factors
- Subject allocation
- Data handling

Which type of study do I use?

- Practical aspects of the field survey
- Shortcomings in the epidemiological method
- Appraising an epidemiological study
- Conclusions
- Reference
- Further reading

Introduction

Definition

Epidemiology may be defined as the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems. Whenever consideration is given to the health of groups of people – and this is invariably so in occupational health – the principles of epidemiology must be understood. Therefore, the hygienist needs not only to be conversant with the methods of epidemiological investigation, but also to be able to incorporate some of these concepts into his or her work. Even if hygienists never wish to undertake an epidemiological study, at the very least they must be able to evaluate such studies. This chapter aims to cover these areas.

Consideration of the definition given above implies that epidemiology concerns a study of the distribution of the disease and a search for the determinants of the observed distribution. Hygienists will be concerned mainly with the latter. Once again, the team approach to occupational health studies is vital. If the physician suspects a health risk that is related to the workplace, further study of the workers may reveal the distribution of this risk – real or apparent – in the health status of the employees. Many variables, including age, sex, occupation and race, will need to be considered. Collaboration with the hygienist will be necessary to determine the occupational factors that may be involved in the observed ill health, for example dust concentrations, work cycles, geographical distribution of the disease in the factory and timing of the onset of disease in relation to new or altered processes. All of these factors could be important

and the hygienist's expertise will greatly enhance the physician's ability to tease out those that are most relevant to the pathological processes underlying the ill health discovered.

The process of elucidating disease causation using epidemiology involves three types of investigation: (1) a description of the current status of health of the 'at-risk' group (descriptive epidemiology); (2) ad hoc studies to test aetiological hypotheses in order to get closer to the likely cause (analytical epidemiology); and (3) the design and execution of a study that aims to alter exposure to the putative risk factor in order to assess whether this leads to an altered disease rate (experimental epidemiology). The last type of study is frequently difficult to design and execute for ethical reasons but, in the final analysis, epidemiology is concerned with preventing ill health by establishing the causes of disease and removing them.

Uses

In occupational health, epidemiology has five main uses:

- 1 the study of disease causation;
- 2 the study of the natural history of a disease;
- 3 the description of the health status of a specified population;
- 4 the evaluation of intervention in health-related issues;
- 5 the development of hygiene standards from epidemiological studies of exposure and health outcome.

Sources of data

Population census and death registrations were first introduced for political and legal reasons. They have, nevertheless, been a never-ending source of data for the investigation of occupational causes of disease. Pension records, sick benefit schemes and treatments records have likewise been used. In fact, epidemiologists will use almost any source of personal health record to further their researches, although few have been collected initially with such researches in mind! However, the emergence of powerful personal data protec-

tion legislation in many countries threatens the very existence of some types of epidemiological research.

National records (vital statistics)

During an individual's lifetime, major milestones are recorded for various purposes. Birth, marriage, divorce, death are all recorded nationally in developed countries. The registrations are undertaken locally and stored centrally.

Death certificates

Death marks the final event in an individual's health record. It is readily verified, and in countries where registration is complete it provides a reasonably accurate and quantifiable measure of life-threatening illness in the community. Inaccuracies in such countries apply only to the cause of death, not to the fact of death. Internationally agreed coding systems exist for death certificate information – the latest is ICD-10 – and this enables death certificates to be analysed by underlying cause. It also allows for more valid between-study comparisons.

It has been established with reasonable certainty that death certificates are reasonably accurate and, if based on broad diagnostic categories, the 'true' cause of death is correctly coded in over 80% of cases. This accuracy varies with the disease – sub-categories of cardiovascular and respiratory disease can be notoriously difficult to define, whereas a disease such as leukaemia is much more accurately defined, as it requires a precise pathological diagnosis prior to the institution of treatment.

Occupation is recorded on the death certificate as the 'last known' occupation. Although uniformly applicable, it leads to a statistic of limited value when studying retired decedents or persons whose death was actually caused by a previous occupation. For example, the occupational categories 'retired', 'housewife' and 'civil servant' are epidemiologically useless. Similar inaccuracies are noted when pneumoconiosis is seen to be the cause of death in a car park attendant; this is not due to the dust generated as cars drive in and out of

the parking area past the man's booth, but due to the fact that the attendant is probably a pensioned-off coal-miner. Such inaccuracies can sometimes be circumvented by the use of factory pension scheme records (see below).

Birth certificates

In the past, these records were primarily used for the establishment of denominators for the calculation of infant disease rates. The recent advent of recording congenital malformations and pregnancy complications, as well as birth weight and duration of pregnancy, has afforded an opportunity of using these records when studying the effects of the mother's as well as the father's occupation. The strictures regarding diagnostic accuracy are, however, similar to those for death certificates.

Morbidity

Nationally acquired morbidity records regarding health and safety at work are available in some countries. They are less accurate than mortality records and for epidemiological purposes would require supplementation with ad hoc recording in order to make them acceptable. In the UK, industrial accidents and certain diseases are reported to the Health and Safety Executive, whereas industrial injury benefit claims and prescribed diseases are reported to the Department of Work and Pensions. Errors in diagnosis, failure to report accidents, illness and injury, and incomplete coverage by law all militate against these sources of data as ideal epidemiological tools. Sickness absence data are particularly deficient regarding female employees and notoriously inaccurate, except for the broadest diagnostic groupings. Nevertheless, such national statistics can indicate gross secular changes and may highlight new hazards. For diseases that are not life-threatening, these are crucial measures of occurrence.

Newer surveillance schemes based on voluntary reporting of certain diseases by occupational health practitioners and hospital specialists are proving to be useful adjuncts to the national statistics.

Local records

These may be acquired through hospitals, family doctors, factories, schools, pension schemes, insurance policies, professional associations and trade unions. All such records have been used at one time or other by occupational health epidemiologists. However, they are all collected for purposes unrelated to epidemiological study, and their accuracy, completeness, comparability and relevance are always doubtful.

Ad hoc records

Some large industrial organizations now maintain continued surveillance of the 'high-risk' workers not only as they move from job to job with the company, but also if and when they leave or retire. These exposure registers can be of inestimable benefit later on in assessing the health status of workers in various exposure circumstances many years previously. At present, there are few industries in which retrospective exposure data of any worth are available.

Notification and registration of certain specific diseases have been made compulsory from time to time. Early examples include various infectious diseases such as whooping cough or measles. More recently, cancer registries have been established in a number of countries. The more efficient ones now boast a diagnostic accuracy in excess of 98%, with a high level of enumeration. Other examples include registers set up to monitor specific diseases such as mesothelioma, angiosarcoma of the liver, the pneumoconioses, adverse reactions to certain drugs, specific congenital malformations and certain disabilities, such as blindness.

Despite this welter of health records, the epidemiologist frequently has to search several sets to obtain only a portion of the information required regarding an employee's health. It may still be necessary to contact the employee or his or her next-of-kin for further data. Even if a total picture of that employee's health from the cradle to the grave is acquired, it may still be too imprecise about possible occupational hazards and their relationship to the worker's health, owing to the

paucity of exposure data at the factory or factories at which the person worked.

Measures of exposure and health outcome

Information relating to hazard exposure and health outcome acquired from the above data sources has to be expressed in terms that permit comparisons between and within populations. This section deals with some of those measures and the concepts underlying their use.

Exposure

For the foreseeable future, this will remain the least accurate measure and therefore the weakest link in the chain between cause and effect. The majority of epidemiological studies investigating occupationally related causes of disease falter when it comes to establishing, with any degree of accuracy, the exposure histories of the populations of workers under study. The whole area of retrospective exposure assessment is now a major research priority in occupational hygiene and is described in Chapter 12. This is the area where hygienists can make a major contribution to future occupational health research.

In the final analysis, the ideal epidemiological study will show a dose–response relationship between a suggested cause and the disease outcome. This adds great strength to the association being causative and also materially assists in establishing safe (or relatively safe) working conditions for future generations of employees.

At present, past exposures are frequently classified as low, medium and high or merely expressed in years of exposure of whatever degree. This is most unsatisfactory. Ideally, past exposure information should be of high quality and measured with great accuracy, using standard or comparable instrumentation. It should not only provide information on, say, the concentrations of the toxic material potentially (and realistically) absorbable by the workers, but also provide accurate data on variations in that concentration during the work cycle, daily, weekly, monthly and its duration.

In addition, data should be available on other relevant exposures. These include changes in the physical and chemical formulation of the toxic substance in the worker's immediate environment as well as an assessment of possible interactions between various noxious elements, whether they be additive, multiplicative, synergistic or even negative.

At present such data are rarely available and if available are virtually never complete. Thus, they furnish epidemiologists with a continuing source of major error in their investigations.

The range of possible exposure groupings is:

- 1 ever/never employed in the industry;
- 2 length of service in the industry;
- 3 job categories by precise division or task duties (qualitative);
- 4 job categories ranked ordinally by exposure intensity;
- 5 quantitative exposure intensity categories;
- 6 quantitative dose categories.

Health outcome

Measures of occurrence

In epidemiology the commonest measures of occurrence are the *incidence* and the *prevalence*. These are commonly expressed as rates per 'person-periods' (usually person-years). The 'incidence' of a disease relates to the occurrence of new cases and the 'incidence rate' relates to the number of new cases that occur in a given population over a given period of time. [Strictly speaking, the 'incidence rate' (or incidence density) should be distinguished from the 'cumulative incidence', which relates to the proportion of persons developing the disease. Also, incidence (and prevalence) can be expressed as spells of disease rather than persons.]

The 'prevalence' of a disease concerns the existing number of cases of the disease at one point in time or over a period of time. Prevalence is, therefore in strict terms, a ratio of the number of existing cases by the population at risk at a given time over a given period.

Prevalence and incidence are related to each other with reference to a given disease through

the duration of disease. The relationship can be expressed thus:

$$\text{Prevalence} \propto \text{incidence} \times \text{duration} \quad (14.1)$$

Pictorially, this can be conceived as a reservoir (Fig. 14.1) supplied with water from streams above and released through the dam below. The quantity of water in the lake (the prevalence) is dependent upon the amount flowing into it from the streams above (the incidence) and the amount of water leaving the lake below the dam (those people with the disease who cease to have the disease – they recover or die).

In practical terms, incidence is an unsuitable measure of occurrence for chronic diseases with a vague or prolonged onset; it is a better measure when the disease is acute and has a clearly defined onset.

Measures of frequency

These are rates of one sort or another. Although rates are established to allow comparability be-

tween populations, they can be confusing if a clear idea of their limitations is not established. In essence, there are three measures of frequency of a given event (death, disease, accident, etc.):

- 1 crude;
- 2 adjusted;
- 3 standardized.

These and other epidemiological definitions are lucidly defined in Last's *A Dictionary of Epidemiology* (see Further reading).

For illustration purposes, we will concentrate on measuring the frequency of death, but any health-related event can be so measured.

$$\text{Crude death rate} = \frac{\text{total deaths in the population at risk}}{\text{population at risk in person-years}} \quad (14.2)$$

Such a death rate is far from ideal as the deaths relate to both sexes, and, more importantly, all age groups. Comparison of two crude death rates could lead to erroneous conclusions. For example, the crude death rate for town A (an

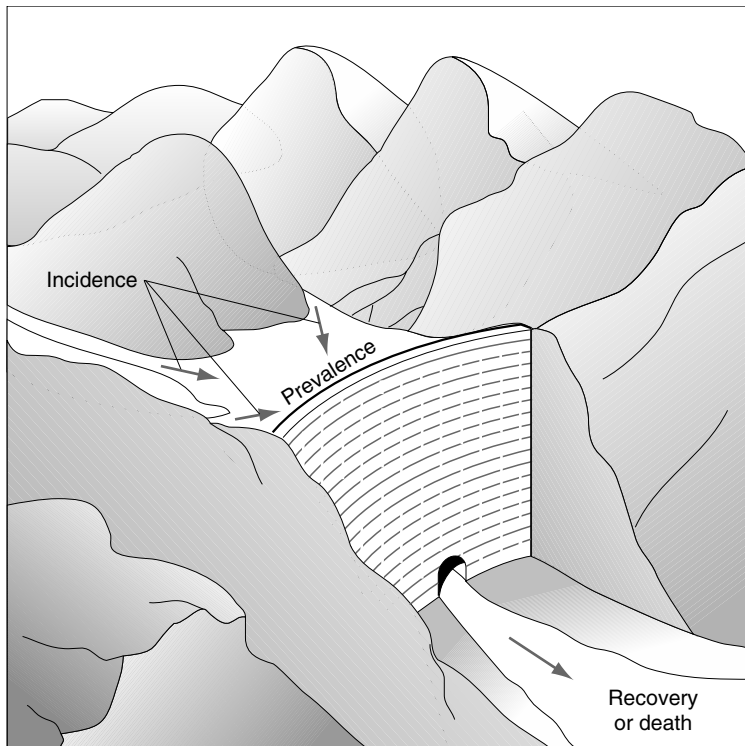


Figure 14.1 Relationship between incidence and prevalence.

industrial centre) is 12 per 10⁶ per year and the rate for town B (a seaside resort) is 15 per 10⁶ per year. The implication is that town B is a less healthy place than town A. This is because no account has been taken of the age breakdown of the populations. Town A has a younger population and the age-specific deaths are all higher than town B, but town B has an older population. The way round this dilemma is to calculate an adjusted death rate. This removes the latent weighting inherent in the crude rates by another set of weights – in this case, the age-specific rates. If this is done for towns A and B, the comparison now shows town A to have higher age-specific death rates than town B and summary statistics can be calculated to reflect this.

Although widely used, the standardized mortality rates (SMRs) do have pitfalls in interpretation for the unwary. First, as the technique involves an indirect age adjustment, SMRs calculated for two or more study populations, using the same standard population death rates for calculating expected deaths, cannot themselves be compared with each other. This is because the two index populations may have different age structures.

The second point is that the magnitude of the SMR is dependent upon the choice of the standard population. As the standard population is frequently the national population, it is, by definition, less healthy than a working (occupational) population. This is because the national population contains people who do not work because they are sick, disabled or dying. Therefore, if one assumes that the index population is exposed to no serious occupational hazards, the SMR for that occupational group, calculated using national data as the standard, should be invariably less than 100%. In practice, this 'healthy worker effect' means that an unexposed occupational group, when compared with the national population should have an SMR of about 80–90. In Britain, the Office of Population, Censuses and Surveys is attempting to get round some of this comparison bias by providing national population data for employed persons – the so-called 'Longitudinal Study'.

The third factor is that SMRs gloss over age-specific differences in the working population. Not all workers are exposed equally to the putative

hazard – it might be more severe in the youngest group, or more noticeable by its cumulative effect in the older groups. A way round this is to consider age-specific SMRs.

A fourth factor that could be relevant is socio-economic class. There are differences in SMR by socio-economic class and the proportion of each in the two populations compared may not be equal. Allowances can be made for this also. Although socio-economic class is a convenient grouping, it is made up of inter-related factors that include income, education and way of life (such as housing and site of house) as well as occupation.

Finally, it is necessary to mention proportionate mortality ratio (PMR). This statistic, although not as robust as the SMR, is useful particularly when populations at risk are not accurately known. The PMR is the proportion of deaths from a given cause in the index population divided by the proportion of deaths from that cause in the standard population.

The relevance of time, place and person

From what has gone before, the reader should, by now, have realized that numbers, the currency in which epidemiologists deal, have to be continually reviewed in the light of factors that could lead to comparison problems. In short, epidemiologists strive to compare apples with apples, not apples with oranges.

Assembling the data for an epidemiological investigation is a bit like piecing together the crucial elements in a criminal investigation. The questions are the same:

- 1 To whom?
- 2 Where?
- 3 When?
- 4 By what? ('Why?')
- 5 How?

The sixth question that is vital to health prevention can be added to these – 'So what?'

The person, place and time questions were asked by John Snow before discovering that the cholera epidemic of 1854 in Soho affected only people who drank water from the Broad Street pump during a few weeks in August of that year. The great London smog of 1951 killed the very young and the

very old in early December in central London, and most of them died of cardiorespiratory disease or terminal bronchopneumonia upon a serious underlying disease such as cancer. Similarly, a study of epidemic adult asthma in Barcelona was linked in time, susceptible person and place to soya bean dust generated in the docks when vessels carrying the beans were unloaded and the wind was blowing in a particular direction.

The main characteristics affecting these three factors are summarized in Table 14.1.

Measures of risk

Risk estimation is primarily a function of data analysis but can be conveniently considered here. Two measures are commonly used:

- 1 relative risk;
- 2 attributable risk.

Relative risk is the ratio of disease rate in exposed persons divided by the disease rate in non-exposed persons. *Attributable risk* is the difference between disease rates in exposed persons and disease in non-exposed persons.

The magnitude of the relative risk is a measure of the strength of the association between the risk factor and the disease – the magnitude of the statistical significance is not related to this strength.

Case-control studies (see below) do not usually permit relative and attributable risk values to be obtained as described above, as such studies do not usually permit direct measurement of disease rate, but merely measure the frequency of risk factor exposure.

Table 14.1 The main characteristics affecting the person, place and time.

<i>Person</i>	<i>Place</i>	<i>Time</i>
Age	Natural boundaries	Day
Sex	Political boundaries	Month
Ethnic group	Urban/rural boundaries	Year
Social class	Place of work in factory	Season
Occupation	Environment	
Marital status	Climate	Secular and cyclic
Family	Migrant status	
Genes		

Causation or association

Before considering specific types of epidemiological studies, one further concept needs to be emphasized: the statistical association of a risk factor with a disease does not necessarily prove causation. The association could be spurious or it may be indirect through other known or unknown variables. Epidemiological techniques can never prove that A causes B, but they can often provide considerable support (or denial) for a causal hypothesis.

Such support can be conveniently considered under nine headings [for the original exposition of these nine factors, the reader is referred to Bradford Hill (1965)]:

- 1 *Strength of association* – is the disease more common in a particular group of workers, and if so, by how much?
- 2 *Consistency* – has the association been described by more than one researcher and preferably using different methods of enquiry?
- 3 *Specificity* – is the disease restricted to certain groups of people and to certain sites?
- 4 *Time* – does the suspect cause always precede the disease and is the time interval reasonable?
- 5 *Biological gradient* – is there a good dose-response relationship?
- 6 *Biological plausibility* – does the association seem reasonable or is it absurd?
- 7 *Coherence* – do all aspects of the causality hang together in a logical and feasible way?
- 8 *Experimental evidence* – can the causality be tested experimentally or does experimental evidence support causality?
- 9 *Analogy* – has a similar suspect cause been shown for related causes or effects?

Rarely will all nine points be present in the proof of a hypothesis, nor do they all carry equal weight.

However, the more there are the stronger the association and the more likely it is that there is a causal relationship. But, as Bradford Hill says in this paper:

All scientific work is incomplete – whether it is observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time.

Study design

Consideration of the design of an epidemiological study requires advance planning. It does not just happen. Having said that, it is usually impossible to stick rigidly to a classic study design as practical considerations will modify such an ideal situation. What is fundamental to all epidemiological studies is the need to have a question that demands an answer.

Study design consideration can be divided into two parts: (1) goals and (2) options.

Goals

The ultimate aim of an epidemiological study is to obtain accurate information about the object of the study. Practical restrictions present before, during and after the investigation may limit the feasibility of obtaining the most accurate picture. The balance between these two opposing forces can be denoted as the efficiency of the operation as a whole (Fig. 14.2).

Validity

Validity is related to the general (external) and the specific (internal). In the general sense it is concerned with how the study results could be extrapolated in a more general context. For example, if a study showed that farmers in a particular area had a higher prevalence of fractured legs

than office workers, can these results be extrapolated to all farmers? Or are there specific circumstances in the study area that show that the farmers or the landscape, climate, their tractors, whatever are significantly different from the population of such workers as a whole to militate against such generalizations? (Perhaps the control group is inappropriate and therefore it is the factor that vitiates generalization.)

In a specific sense, the study groups may be biased. Bias can take many forms and some can be controlled at the design stage. Three broad groups can be distinguished:

1 Selection bias concerns the way the study populations were assembled and includes the validity of choosing the chosen, for example 'were they volunteers?'; 'were they lost to follow up?'; 'were they a survivor population?'; and 'did they all come from the same hospital/district?' and so on.

2 Information bias relates to the quality and accuracy of the data gathered. It includes errors by the interviewer or the interviewee in the diagnosis or the exposure measures and so on.

3 Confounding is a factor that independently influences both the exposure and the outcome and thereby suggests a spurious direct relationship between the two. The classic example is age. The older the worker, the more likely he or she is to have been significantly exposed to the occupational hazard and the more likely to have the illness in question – given that most diseases are age dependent.

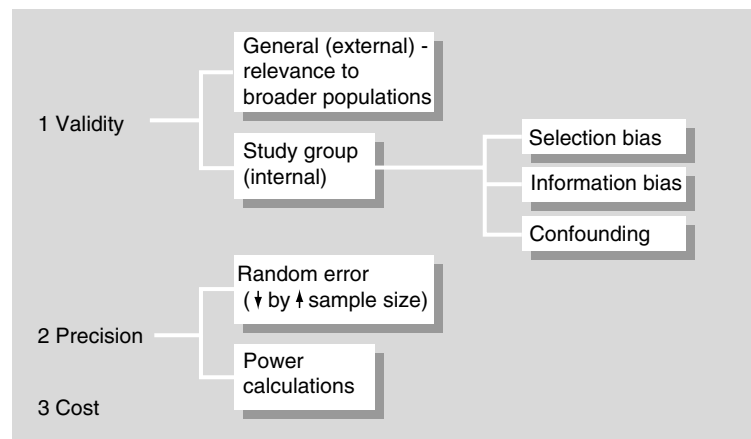


Figure 14.2 Study design: goals.

Precision

Precision is another aspect to be considered. If a study plan, involving a particular size of population (or measurements), were implemented repeatedly and independently an infinite number of times, the results would be grouped about a mean value. Departure from this mean value would give an estimate of random error. A small random error would indicate high precision. Information regarding the precision of the study is augmented by increasing the sample size or by increasing the number of times the measurements are made. Wherever possible, it is valuable to calculate the likelihood of discovering a real effect given the study population size. Formulae are available to undertake these so-called power calculations.

Cost

The cost of the investigation can be measured in terms of time, effort and personnel. The efficiency of the study is a measure of the value of the information gained against the cost of the study.

Options

The goals of rational study design require that certain choices are made in terms of the way the investigation is executed (Fig. 14.3).

Timing

A choice of prime importance is the timing of the investigation. Figure 14.4 depicts the life history of a factory population on a calendar – time/age–time format – and can be used to illustrate these options in timing. Real-life events are, of course, more complex than in this example but the principles remain the same. The factory concerned opened in 1945 with a population of workers aged 18–40 years. The passage of time (horizontal scale) is, of course, accompanied by the ageing of the population (vertical scale). The population is assumed not to alter and therefore progresses diagonally across the figure. In 1960, a major expansion programme with the advent of new processes necessitates enlarging the workforce with predominantly

younger men. This new cohort ages *pari passu* with the extant group. (An alternative model is the introduction and cessation of a particular process in 1945 and 1960 respectively.) If we wish to investigate this process and/or the population, we have several options with regards to timing and the direction of the study. Referring to Fig. 14.3, the study could be cross-sectional (vertically orientated) or longitudinal (horizontally orientated).

Cross-sectional studies

The decision to undertake a cross-sectional study generally means a quick, cheap opportunity to study the problem in hand. These advantages are offset by the limitations imposed by having to assess the population at risk in a narrow time frame. The narrowness of the time interval means that the investigators cannot look at exposure and outcome as a time-dependent relationship. The cross-sectional study tends to be outcome or exposure selective. All that may be feasible is the estimation of the prevalence of the exposure (or its outcome).

Longitudinal studies

Longitudinal investigations take longer to do and are more expensive but provide, by virtue of the study being concerned with a period of time rather than an instant, an opportunity for looking at an exposure and its outcome as a time-related chain of events. Two types of longitudinal study are commonly employed: (1) the case–control study (or, more accurately, case–referent study) and (2) the follow-up study (which includes cohort investigations) (Fig. 14.5).

Case–control studies tend to be retrospective. They begin with a definition of a group of cases and relate these and the non-cases (control subjects/referents) to the past exposure history. In occupational epidemiology, the main drawback here is the accurate ascertainment of exposure history going back anything up to 40 years.

Follow-up studies do not necessarily suffer such limitations if the exposure is defined and accurately known, and a group of exposed (and possibly

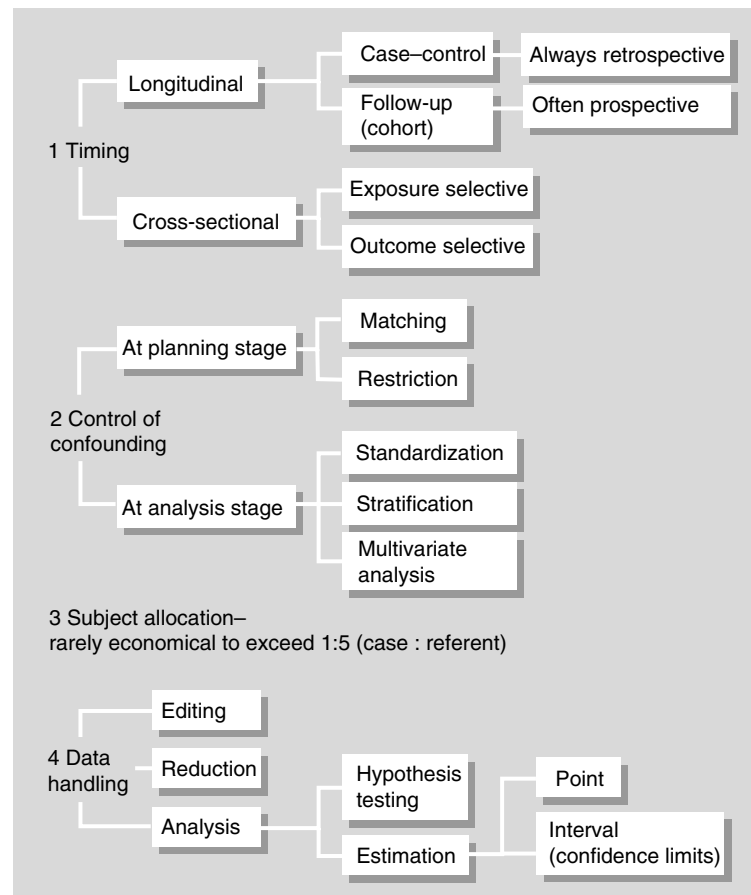


Figure 14.3 Study design: options.

non-exposed persons) are followed up to assess the eventual outcome of such exposure. Such studies are frequently, but not invariably, prospective in directionality. A cohort is a specific type of follow-up study in which a population, defined in advance for exposure characteristics, is followed for a period of time and the outcome subsequently measured. Follow-up studies are designed to observe incidence and, ideally, should span a period of time in excess of the maximum induction period for the exposure factor to produce a putative outcome.

Control of confounding factors

The control of confounding factors can be undertaken at the planning stage or during data analysis. During planning, matching the cases (or exposure

group) with persons without the characteristic essential for selection may reduce or eliminate such confounders. For example, age-matching eliminates the problem of inadvertently comparing old with young. The alternative to matching is the restriction of the cases and their referents to narrow strata, which effectively excludes unwanted confounding. Such stratification without restriction can be carried out during data analysis, or standardization procedures can be adopted instead. Complex inter-reactive confounders can be minimized by multivariate analysis.

Subject allocation

Studies are often published in which numbers in the referent (control) group exceed the cases by a factor of 2 or more. This can strengthen the valid-

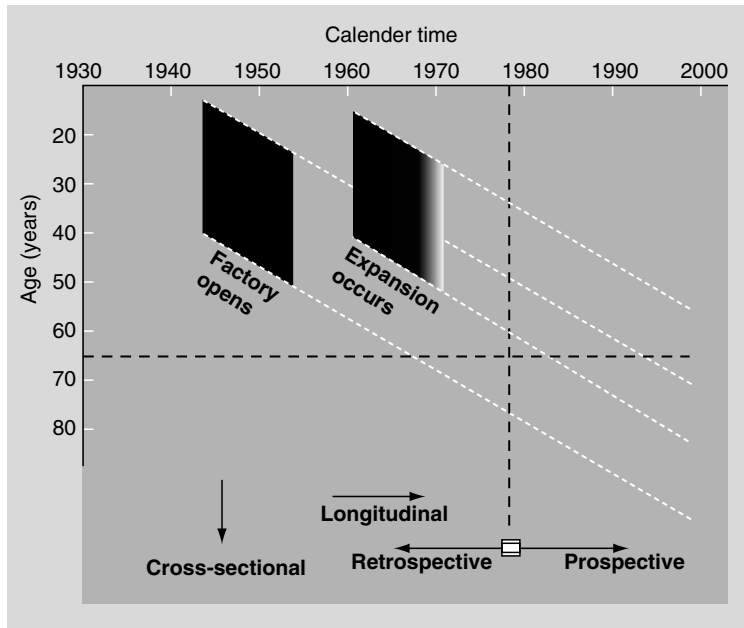


Figure 14.4 Some age–time factors in study design.

ity of the comparison and, thereby, the conclusions are drawn. The law of diminishing returns, however, begins to operate after the case–referent ratio exceeds 1:3, and 1:5 is rarely exceeded – largely for reasons of economy.

Data handling

Three main procedures are employed: (1) the editing of the collected data to a readily usable form; (2) its reduction to a manageable size; and, finally, (3) the analysis. Analytical procedures tend to test hypotheses propounded at the outset of the investigation or are employed to estimate various parameters, either to one point or, more commonly, to establish confidence limits for the calculated estimates.

Which type of study do I use?

For many years, follow-up studies have been considered to be the epidemiological study par excellence. Their reputation for accuracy is based partly on their inherently unbiased concept – a group or groups of people are chosen in terms of characteristics manifest before the appearance of the disease in question, and then these individuals are followed over time to observe the frequency of the disease. By definition, however, such studies tend to take a long time to complete and, in addition, are costly and frequently rather complex. Nevertheless, some epidemiologists nowadays feel that a

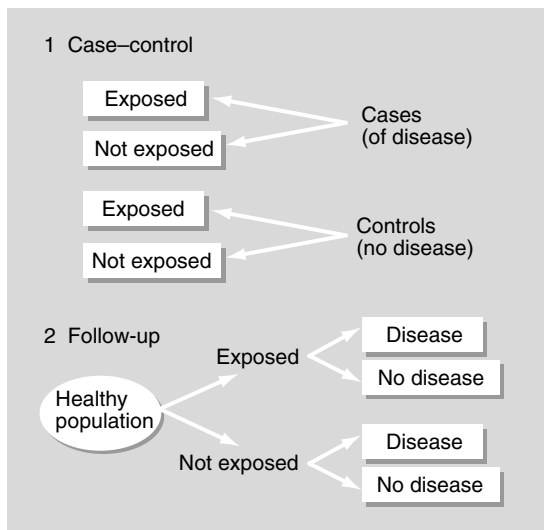


Figure 14.5 Longitudinal studies.

well-designed and efficiently executed case-control study 'nested' within a cohort carries the advantages of both studies while minimizing the disadvantages of each.

The choice between the two is dependent upon the question posed in the exposure/health outcome equation. This can be summarized as follows:

<i>Question</i>	<i>Study type</i>
A causes B?	Either Case-control (disease rare) Follow-up (exposure rare)
A causes B ₁ , B ₂ , B ₃ ?	Follow-up
A causes ?	Case-control
B caused by A ₁ , A ₂ , A ₃ ?	
B caused by A?	

Selecting cases (and control subjects) can, however, be difficult. Here are some questions that the researcher should ask before making a choice:

- 1 Have I collected information on the control subjects with the same zeal as the cases?
- 2 Are the response rates from both cases and control subjects similar?
- 3 How well are these referents matched to the case?
- 4 Have I matched all important potential confounders?
- 5 Have I 'overmatched'?
- 6 Do the referents come from a population similar to the cases?
- 7 Could I choose an equally good comparison group more cheaply or more quickly?

Practical aspects of the field survey

Most hygienists operate in the practical world of the workplace and spend most of their time attempting to measure the environment and controlling the risks they find there. Fieldwork in epidemiology also forms an important part of many of the epidemiological studies discussed above and it might be of value, therefore, to list the steps likely to be required in planning and executing an epidemiological field study.

- 1 Define study objectives and formulate hypotheses to be tested.
- 2 Review the literature.

- 3 Outline the study plan and assess feasibility.
- 4 Define the study population, preferably with the help of a statistician, especially if the population is large, and matching for comparison groups is envisaged.
- 5 Define study methods: cross-sectional, longitudinal, etc.
- 6 Decide on timing of the study.
- 7 Plan data processing and analysis.
- 8 Assess sample size and costs – including power calculations.
- 9 Write detailed protocol.
- 10 Publicize study plan to important, interested parties – especially the workforce – and obtain all necessary agreements (including ethical approval if necessary).
- 11 Prepare questionnaire, if any.
- 12 Undertake pilot study to test methods.
- 13 Recruit and train ancillary staff.
- 14 Approach study population(s); they should be aware of the aims of the study, the expected benefits, the sponsors and be assured regarding the confidentiality of any personal data they may be asked to provide.
- 15 Redefine study, if necessary, in the light of previous steps – diagnostic criteria, measurement assessments, questionnaire design, sampling procedures, error assessments, etc.
- 16 Main study.
- 17 Vigorously minimize 'non-response'.
- 18 Assess bias and errors.
- 19 Edit data.
- 20 Reduce data.
- 21 Analyse and test hypotheses.
- 22 Reach conclusions and report results to the participants and in the scientific literature.
- 23 ?Plan further research in the light of conclusions, including any opportunities for preventative action.

Shortcomings in the epidemiological method

By now it will be clear to the reader that the epidemiological method is not without its difficulties! In brief, the main problems can be summarized as follows:

- a healthy worker effect – the comparison group has a different general health status compared with the cases;
- a poor response rate;
- high turnover of study populations – selecting in (or out);
- latency between exposure and effect longer than study period;
- insufficient evidence of differing effects by differing exposures;
- poor quality of health effects data;
- poor quality of exposure data, multiple exposures;
- no effect of exposure noted – ‘does this imply a true negative result or merely a poor or small study (non-positive result)?’

Appraising an epidemiological study

Perhaps the reader has now been finally dissuaded from ever attempting an epidemiological study. That may be so. Indeed, it is not an essential part of the hygienist’s job description. But it is crucial that hygienists understand what epidemiologists do and equally vital that they can read a report of an epidemiological paper and know whether it is good, bad or indifferent. To help that process, the checklist below summarizes much of what has been described in this chapter:

- question clearly formulated;
- appropriate study design;
- good-quality health effects data;
- good-quality exposure data;
- valid population choice for cases and control;
- high response rate and good sampling strategy;
- confounders considered and allowed for;
- population large enough to detect an effect if present;
- correct statistical techniques;
- estimates of risk include measures of variability, e.g. confidence intervals;

- cause–association issues addressed;
- non-positive or negative study result reviewed;
- effect of results on current knowledge assessed.

Conclusions

It is too glib to describe epidemiology as ‘common sense made complicated’. It can be (and should be) a science applied to the study of diseases in populations. It is logical but can be complicated. It does not necessarily have to be undertaken by a cast of thousands studying populations of millions for decades. The principles apply equally to smaller scale, factory-based investigations undertaken by one researcher. Every occupational health specialist should be conversant with the tenets of epidemiology and capable of executing studies using its methods. It is hoped that a chapter such as this, although seemingly out of place in the eyes of many occupational hygiene specialists, would be viewed by the reader as relevant. Ideally, it should stimulate him or her to go back out on to the shop floor and view the workforce anew. If you find a working population in which there is an exposure–health outcome question to be answered, then you have an epidemiological study on your hands!

Reference

Hill, A.B. (1965). Environment and disease, association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295–8.

Further reading

- Last, J.M. (2002). *A Dictionary of Epidemiology*, 4th edn. Oxford University Press, New York.
- Monson, R.R. (1990). *Occupational Epidemiology*, 2nd edn. CRC Press, Boca Raton, FL.
- Rothman, K.J. (2002). *Epidemiology – An Introduction*. Oxford University Press, New York.

Part 4

**Environmental hazards: principles
and methods of their assessment**

Chapter 15

The sampling of aerosols: principles and methods

David Mark

Introduction	Personal samplers
Health-related aerosol sampling criteria	Static samplers
Fibrous aerosols	Multifraction samplers
Biological aerosols	Investigational instruments
Sampling strategy	Errors involved in taking an aerosol sample
The basics of an aerosol sampling system	Performance of the sampling head
The sampling head	Flow rate setting and control
The transmission section	Analytical errors: gravimetric
Size selectors	Analytical errors: chemical
Filters	Random errors: variability of exposure
Pumps	Special problems
Practical samplers for inhalable aerosol	Fibrous aerosol particles
Personal samplers	Bioaerosols
Static (or area) samplers	Two-phase compounds
Practical samplers for respirable aerosol	Future developments
Personal samplers	References
Static samplers	Further reading
Practical samplers for thoracic aerosol	

Introduction

The term ‘aerosol’ is defined as a disperse system of solid or liquid particles suspended in a gas, most commonly air. In the workplace, it applies to a wide range of particle clouds including: compact mineral and metallic particles produced in the process and manufacturing industries; agglomerated particles found in fumes during welding, metal smelting, etc.; fibrous particles such as asbestos and man-made mineral fibres produced for insulation purposes; droplets from electroplating processes and cutting oils; and a number of bioaerosols produced in the agricultural, food and biotechnology industries.

Generally, as explained in Chapter 7, the aerodynamic behaviour of all aerosol particles is controlled by the same physical processes. It is the aerodynamic behaviour that governs whether the aerosol particle remains airborne, how far it travels from the source of production, whether it

is captured by aerosol control systems and whether it reaches the human body. In the field of occupational hygiene, we are interested in controlling the levels of aerosol in the workplace air and thereby minimizing the risk to workers from adverse health effects due to exposure to the aerosol particles.

There are three main routes by which aerosol particles can reach the body and have the potential to cause harm. They are inhalation, skin deposition and the food chain. In the occupational field, exposure from particles depositing on food or drink is very rare, as meals are normally taken in specially provided areas away from the workplace. Skin deposition is known to be a significant route of exposure for droplet aerosols and some metallic particles, but the most important route of exposure for aerosols is generally considered to be by inhalation.

This chapter concentrates on the sampling of aerosols for estimating the risk from inhalation. It starts by providing a scientific framework for the

health-related sampling of aerosols by describing the sampling criteria and considering the sampling strategy to be employed. It then moves on to outline the basics of an aerosol sampling system and describes briefly the many systems that can be employed for the tasks specified. Sections describing procedures and errors involved in taking an aerosol sample follow, with special problems dealt with separately. With the present high international activity on standards for aerosol sampling (in Europe and the USA), and improvements in detector technology, it must be emphasized that this chapter describes the present state of affairs. A final section is included, therefore, that deals with future developments and how they may affect the occupational hygienist measuring aerosol concentrations in the workplace.

Health-related aerosol sampling criteria

It has long been realized that different aerosol-related health effects may be linked to different sizes of aerosol. Until recently, two sizes were considered to be important for health effects. Coarse particles, which could be deposited in the upper regions of the respiratory tract, were thought to be associated with toxic effects, whereas the finer particles, which could penetrate into the gas exchange region of the lung, were implicated as the cause of pneumoconioses, such as that found in coal-miners.

The sampling of coarse particles for health-related purposes was based, in the past, on the use of samplers for so-called 'total' aerosol. Although the implicit assumption in their use was that they collected a sample of all sizes of airborne particles with 100% efficiency, in practice early samplers were not designed with regard to their sampling efficiencies and measurements of 'total' aerosol would have varied greatly, dependent upon the instrument used. More recently, a form of standardization based on a physical rationale has been achieved by specifying a fixed mean air velocity entering the sampler of 1.25 m s^{-1} .

The sampling of fine aerosols has historically borne more relevance to those particles that are thought to be responsible for disease in the deep

lung, such as pneumoconiosis. In the late 1950s and 1960s, a number of definitions were proposed for the so-called 'respirable' fraction, representing particles that penetrated to the gas exchange region of the lung. They were the British Medical Research Council (BMRC) curve (Orenstein, 1960), the US Atomic Energy Commission (AEC) curve (Lippmann and Harris, 1962) and the American Conference of Governmental Industrial Hygienists (ACGIH) curve (American Conference of Governmental Industrial Hygienists, 1968). They were pragmatic curves, because not only did they fit the available deposition data reasonably well, but they were also matched by suitable samplers. However, they were different and therefore measurements made according to one criterion could not necessarily be compared with those made by the other.

Over the last 10–15 years, considerable progress has been made on the provision of internationally acceptable definitions for health-related aerosol fractions in workplace atmospheres. Collaboration between members of committees of the International Organization for Standardization (ISO), the Comité Européen de Normalisation (CEN) and the ACGIH has produced a set of agreed definitions for health-related aerosol fractions in both workplace and ambient atmospheres. The workplace set has been published by CEN as EN 481 (Comité Européen de Normalisation, 1993a) and is available from the British Standards Institution as BS EN 481. It is also published by ISO as IS 7708 (International Standards Organisation, 1996). The three main fractions – inhalable, thoracic and respirable – are provided for occupational hygiene use and they are shown in Fig. 15.1.

The *inhalable fraction* (E_T) is defined as the mass fraction of total airborne particles that is inhaled through the nose and/or mouth. It was derived from wind tunnel measurements of the sampling efficiency of full-size tailor's mannequins and replaces the very loosely defined 'total' aerosol fraction used previously. For industrial workplaces it is given by:

$$E_I = 50[1 + \exp(-0.06D)] \quad (15.1)$$

where D is the particle aerodynamic diameter.

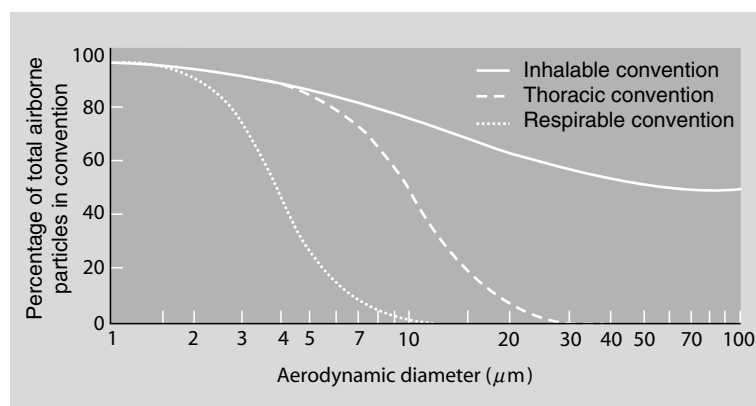


Figure 15.1 ISO/CEN/ACGIH sampling conventions for health-related aerosols.

The *thoracic fraction* (E_T) is defined as the mass fraction of inhaled particles penetrating the respiratory system beyond the larynx. It is given by a cumulative log-normal curve, with a median aerodynamic diameter of 11.64 μm and geometric standard deviation of 1.5.

The *respirable fraction* (E_R) is defined as the mass fraction of inhaled particles that penetrates to the unciliated airways of the lung (the alveolar region). It is given by a cumulative log-normal curve with a median aerodynamic diameter of 4.25 μm and a geometric standard deviation of 1.5.

These sampling conventions comprise the target specifications for the design of sampling instruments for the health-related sampling of aerosols in the workplace. However, these conventions cannot be consistently applied without an agreed testing protocol. This is presented in EN 13205 (Comité Européen de Normalisation, 2000) and provides guidance on the assessment of the suitability of samplers based upon the overall inaccuracy (bias and precision) by which the sampler measures the mass of a chosen aerosol fraction from a wide range of occupationally occurring size distributions. Pilot studies carried out on the development of the standard showed that some of the samplers used at present are able to meet the requirements of the new conventions and the test protocol.

The agreements reached in these new conventions have given renewed impetus to the move towards a new set of standards, based on the three fractions (inhalable, thoracic and respirable)

replacing the old ‘total’ and respirable aerosol combination. To see which aerosol types fall into which categories, it is necessary to reconsider the types of health effects associated with the deposition of particles of various types at the various parts of the respiratory tract.

- The deposition of some biologically active particles (e.g. bacteria, fungi, allergens) in the extrathoracic airways of the head may lead to the inflammation of sensitive membranes in that region, such as symptoms of ‘hay fever’ (e.g. rhinitis). Other types of particle (e.g. nickel, radioactive material, wood dust) depositing in the same region may lead to more serious local conditions, such as ulceration or nasal cancer. For the health-related measurement of all such aerosols, it is appropriate to sample according to a criterion based on the *inhalable* fraction.

- Some particles may provoke local responses in the tracheobronchial region of the lung, leading to such effects as bronchoconstriction, chronic bronchitis, bronchial carcinoma, etc. For the health-related measurement of all such aerosols, it is appropriate to sample the *thoracic* fraction.

- Lastly, those particles that deposit in the alveolar region may cause pneumoconiosis, emphysema, alveolitis and pulmonary carcinoma, etc. Asbestos fibres may cause mesothelioma in the nearby pleural cavity. In relation to these, *respirable* aerosol continues to provide the most appropriate sampling criterion.

Finally, as a general rule, standards should be specified in terms of the inhalable fraction for

aerosol substances that are soluble and are known to be associated with systemic effects (where toxic material can enter the blood after deposition in any part of the respiratory tract and be transported to other organs).

Fibrous aerosols

Fibrous aerosol particles – those with long aspect ratio such as asbestos and man-made mineral fibres – have, historically, been considered separately by the scientific community. The definition of what is ‘respirable’ for such particles is based not only on the aerodynamic factors that govern the deposition of fibres in the lung after inhalation, but also on their known dimension-associated health risks. For example, in the case of asbestos, long, thin fibres are thought to be more hazardous to health than short, fat ones. This is because they are capable of penetrating deep into the alveolar region of the lung, and the normal lung defence mechanisms are less able to eliminate long particles than isometric ones of similar aerodynamic diameter. Selection of the respirable fraction of the airborne fibres is therefore carried out, after they have been collected on filters, by sizing and counting under the microscope. Unlike for isometric particles, it is the number, not the mass, of particles that is measured.

The internationally agreed criteria for respirable asbestos fibres is that they should have an aspect ratio of 3:1, length > 5 µm and diameter < 3 µm. This was agreed in 1979 by the Asbestos International Association (AIA, 1979) and is still widely used today. Both sampling and microscopy procedures are prescribed in this document, many versions of which have been proposed in different countries. Although this prescriptive approach should lead to consistency in fibre measurements, it is not consistent with the more general conventions described above, which specify instrument performance and not particular instrument designs.

Biological aerosols

The recently agreed European Standard (EN 13098, 2001) has provided welcome guidance for

the measurement of airborne micro-organisms and endotoxins in the workplace. Strategies are given for measurement that include not only appreciation of the requirements of health-related sampling (EN 481), but also the additional problems of the fact that some of the particles only cause problems to the human body when alive. These particles (bacteria, viruses, moulds, etc.) are detected by culturing them on media such as agar, and so they must be kept alive and unharmed during the sampling process. For these particles it is the number rather than the mass concentration that is determined. EN 13098 has built on the many studies that were carried out in the 1990s to study the sampling of bioaerosols and tables of generic information are given to guide the reader to the most appropriate methodology for his specific environment and type of micro-organism.

Sampling strategy

A detailed discussion of sampling strategy is given in Chapter 17. Nevertheless, it is an essential part of the sampling process and will be discussed here briefly.

The most important question to ask before setting out to develop a sampling strategy is ‘Is it necessary to sample?’. In the UK, the Control of Substances Hazardous to Health (COSHH) Regulations state clearly that an assessment of the likely risk to health at the workplace should be carried out first. Only if the estimated risk may be significant is it recommended that a sampling programme should be instigated. There are then a whole number of questions (outlined in Chapter 17) that need to be answered before a reliable sampling strategy can be achieved.

If sampling is to be carried out to assess the true exposures of individual workers (or of groups of workers), one of the most important questions is whether a personal or static sampling strategy should be used (or a combination of both). In static (or area) measurements, the chosen instrument is located in the workplace atmosphere, and provides a measurement of aerosol concentration that is (hopefully) relevant to the workforce as a whole. For the case of personal measurements, the chosen

instrument is mounted on the body of the exposed subject and moves around with him or her at all times.

When choosing one or other of these, some important considerations need to be taken into account. For a few workplaces (e.g. some working groups in longwall mining), it has been shown that reasonably good comparison may be obtained using suitably placed static instruments and personal samplers. More generally, however, static samplers have been found to perform less well, tending to give aerosol concentrations that are consistently low compared with those obtained using personal samplers. One advantage with static samplers is that a relatively small number of instruments may be used to survey a whole workforce. If this can be shown to provide valid and representative results, it is a simple and cost-effective alternative. Furthermore, the high flow rates that are available for static samplers mean that, even at very low aerosol concentrations, a relatively large sample mass can be collected in a short sampling period.

The use of personal samplers is more labour-intensive. More instruments are deployed and this leads to greater effort in setting them up and in recovering and analysing the samples afterwards. By definition, personal sampling involves the direct cooperation of the workers themselves. Also, for such samplers, it is inevitable that the capacities of the pumps used will be limited by their portability. So, flow rates will usually be low (rarely $> 4 \text{ l min}^{-1}$). However, personal aerosol sampling is the only reliable means of assessing the true aerosol exposures of individual workers, so it is by far the most common method of aerosol measurement in workplaces.

A combination of both static and personal measurements should provide the most cost-effective and comprehensive sampling strategy. Personal samplers can be used to provide the detailed individual exposure information for regulatory purposes, for example, on one shift every month, whereas coverage of the other shifts may be achieved by using a strategically placed static monitor providing continuous assessment. Provided that the work process is relatively stable, an alarm monitor may be employed, set to trigger when the level is

reached at which personal exposure is expected to exceed the occupational exposure limit. This system would require calibration with personal sampling to set the appropriate trigger level.

The basics of an aerosol sampling system

An aerosol sampling instrument (or sampler) always comprises a number of components that contribute to the overall accuracy with which a sample is taken. These components are: the sampling head; the transmission section; the particle size selector (which is not always present); the collecting or sensing medium; calibrated flow monitoring and control; and the pump. A simple schematic diagram of these essential components is given in Fig. 15.2.

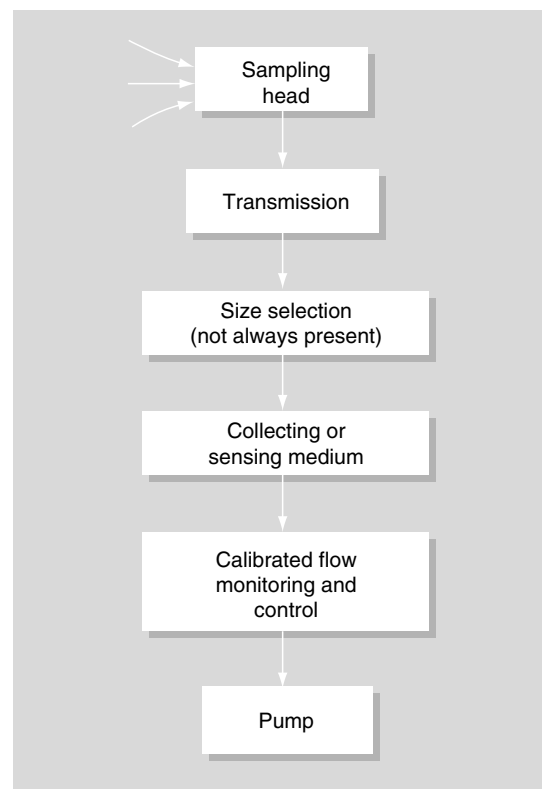


Figure 15.2 The basic components of an aerosol sampling system.

The sampling head

In the past, many occupational hygienists have overlooked the choice of sampling head. Instead, they have concentrated on the performance of the pump and the choice of filter and the analytical technique to be employed, and have used any filter holder of suitable size, without knowing its particle sampling efficiency. This approach may lead to large errors in the mass concentration measured. For example, there is a wide range of different personal sampling heads used for the measurement of so-called 'total aerosol'. Recent wind tunnel tests (Vincent and Mark, 1990) have shown that these samplers exhibit widely differing sampling efficiencies under typical workplace conditions and therefore would not be expected to give the same mass concentration values. However, this situation is now much improved because, with the approval of the new inhalable aerosol convention described above, instrument manufacturers will, and users must, demonstrate that their equipment meets the new convention if the result is to be used for regulatory purposes. To aid this process, work has been completed within Europe to provide a test protocol to demonstrate compliance of samplers to the new conventions (EN 12305) (CEN, 2000).

For the thoracic and the respirable fractions, the performance of the sampler entry should nevertheless be taken into account, although it is not as important as for the inhalable fraction. The most suitable approach for a sampler for the thoracic or respirable fraction (and one that mimics the way that particles from the ambient air arrive at their site of deposition) is to use an inhalable aerosol entry followed by a suitable size selector. Suitable entries are available for both personal and static samplers.

The transmission section

This concerns the transport of particles that have entered the sampler entry to the collecting or sensing region. Particle inertia and sedimentation forces can result in considerable deposition onto the internal walls of the transmission section, especially when the sampling head is remote from the sensing region and particles are conducted down narrow

pipework and bends. This results in a loss of particles reaching the sensing region and is dependent on both the aerodynamic diameter of the particles and their composition. Droplets and soft, sticky particles will stick to the walls once deposited, whereas hard, granular particles may bounce off the walls and become re-entrained into the airflow. Electrostatic forces may also be important when the particles are highly charged and the sampler is made from non-conducting material. In addition, therefore, the humidity of the sampled air may play a role in both increasing the conductivity of the surface and reducing particle bounce. For personal samplers, these wall losses can introduce a significant error into the measurement. Losses of up to 100% have been reported for some non-conducting 'total' aerosol samplers, and these losses cannot be easily accounted for as they depend upon the roughness of handling of the particle-laden sampling head. For static samplers, similar problems occur and are especially evident in continuous particle counters and particle size analysers.

Size selectors

For the thoracic and respirable fractions, some form of particle size selector is used to select the relevant portion of the sampled aerosol. Particles are generally selected by aerodynamic means using physical processes similar to those involved in the deposition of particles in the respiratory system. Gravitational sedimentation processes are used to select particles in horizontal and vertical elutriators; centrifugal sedimentation is used in cyclones; inertial forces are used in impactors; and porous foams employ a combination of both sedimentation and inertial forces.

Owing to their size and the requirement to be accurately horizontal or vertical for correct operation, elutriators are only used in static samplers. Cyclones, impactors and foams, however, can be used in either personal or static samplers.

Filters

A filter is the most common means of collecting the aerosol sample in a form suitable for assessment. Assessment might include gravimetric weighing on

an analytical balance before and after sampling to obtain the sampled mass. It might also include visual assessment using an optical or electron microscope, and/or a range of analytical and chemical techniques.

The choice of filter type for a given application greatly depends on how it is proposed to analyse the collected sample. Many different filter materials are now available, with markedly different physical and chemical properties. These include fibrous (e.g. glass), membrane (e.g. cellulose nitrate) and sintered (e.g. silver) filters. Membrane filters have the advantage that they can retain particles effectively on their surface (which is good for microscopy), whereas fibrous filters have the advantage of providing in-depth particle collection, and hence a high load-carrying capacity (which is good for gravimetric assessment).

Filters are available in a range of dimensions (e.g. 25–100 mm in diameter) and pore sizes (e.g. 0.1–10 μm). Collection efficiency is usually close to 100% for particles in most size ranges of interest. However, sometimes reduction in efficiency might be traded against the lower pressure drop requirements of a filter with greater pore size. For some types of filter, electrostatic charge can present aerosol collection and handling problems – in which case, the use of a static eliminator may (but not always) provide a solution. For other types, weight variations due to moisture absorption can cause difficulty, especially when being used for the gravimetric assessment of low masses. It is therefore recommended that the stabilization of filters overnight in the laboratory be carried out before each weighing, together with the use of blank ‘control’ filters to establish the level of variability. It is preferable that temperature and humidity be controlled in the balance room, especially when collected particle weights are low.

The chemical requirements of filters depend on the nature of the analysis that is proposed. As already mentioned, weight stability is important for gravimetric assessment. If particle counting by optical microscopy is required, then the filters used must be capable of being rendered transparent (i.e. cleared). Direct on-filter measurements of mineralogical composition (e.g. by infrared spectrophotometry, X-ray diffraction, scanning electron

microscope and energy-dispersive X-ray analyses, X-ray fluorescence) are often required. For these, filters must allow good transmission of the radiation used, with low background scatter. Collected samples may also be extracted from the filter prior to analysis, using a range of wet chemical methods, ultrasonication, ashing, etc., each of which imposes a range of specific filter requirements.

Pumps

Most samplers require a source of air movement so that particulate-laden air can be aspirated into the instrument. For personal and static (or area) sampling, the main difference in terms of pump requirements is the flow rate, which tends to be low for personal sampling (usually from 1 to 4 l min⁻¹), and larger (up to 100 l min⁻¹ and even higher) for static sampling. The main limiting factor for a personal sampling pump is its weight, as it must be light enough to be worn on the body (usually on a belt) without inconvenience to the wearer.

A wide range of lightweight, battery-powered pumps is available for personal sampling (and also static sampling, if desired). These instruments are based on diaphragm, piston and rotary pumping principles. Those in practical use are equipped with damping devices to reduce the effects of flow pulsations. The actual volumetric flow rate will depend first on sampling considerations (e.g. entry conditions to provide the desired performance), and then the amount of material to be collected for accurate assessment, analytical requirements, etc. Internal flowmeters, usually of the rotameter type or digital counters, are incorporated into most pumps, but these must always be calibrated against a primary flow rate standard (e.g. a bubble flowmeter or modern dry cal flowmeter). It should also be noted that the flow rate may vary with the resistance imposed by the filter and its collected aerosol mass. For this reason, flow rates should be checked periodically during sampling and adjusted if necessary. However, most modern pumps incorporate some form of flow control that eliminates the need for such regular attention during sampling. Finally, for sampling in potentially explosive atmospheres (e.g. coalmines, chemical plants), intrinsically safe or flame-proof pumps should be used.

Performance requirements for personal sampling pumps are specified in European Standard EN 1232 (Comité Européen de Normalisation, 1993b).

Practical samplers for inhalable aerosol

Personal samplers

There are many small sampling heads used to measure 'total' aerosols. However, not all of them will be suitable for measuring personal exposures to the inhalable fraction. Tests carried out in moving air and calm air by a number of workers have shown that there is not one sampler that fully complies with the inhalability criterion. The samplers with the best overall performance include the 2 l min^{-1} Institute of Occupational Medicine (IOM) personal inhalable aerosol sampler and the 3.5 l min^{-1} GSP conical inhalable sampler. These are shown in Fig. 15.3. The IOM sampler (Mark and Vincent, 1986) features a 15-mm-diameter circular entry that faces directly outwards when

the sampler is worn on the torso. The entry is incorporated into an aerosol-collecting cassette, which, during sampling, is located behind the face-plate. This cassette also houses the filter, and the whole cassette assembly (tare weight of the order of a few grams) is weighed before and after sampling to provide the full mass of aspirated aerosol. This system eliminates the possibility of errors associated with internal wall losses. In addition, the lips of the entry protrude outwards slightly from the faceplate in order to prevent over-sampling associated with particle blow-off from the external sampler surfaces.

Static (or area) samplers

There are a number of high-flow-rate samplers deployed at present as static samplers for coarse (or total) aerosol in the nuclear and chemicals industries (e.g. galley samplers). However, the sampling performance of these samplers does not follow the inhalability convention and so their

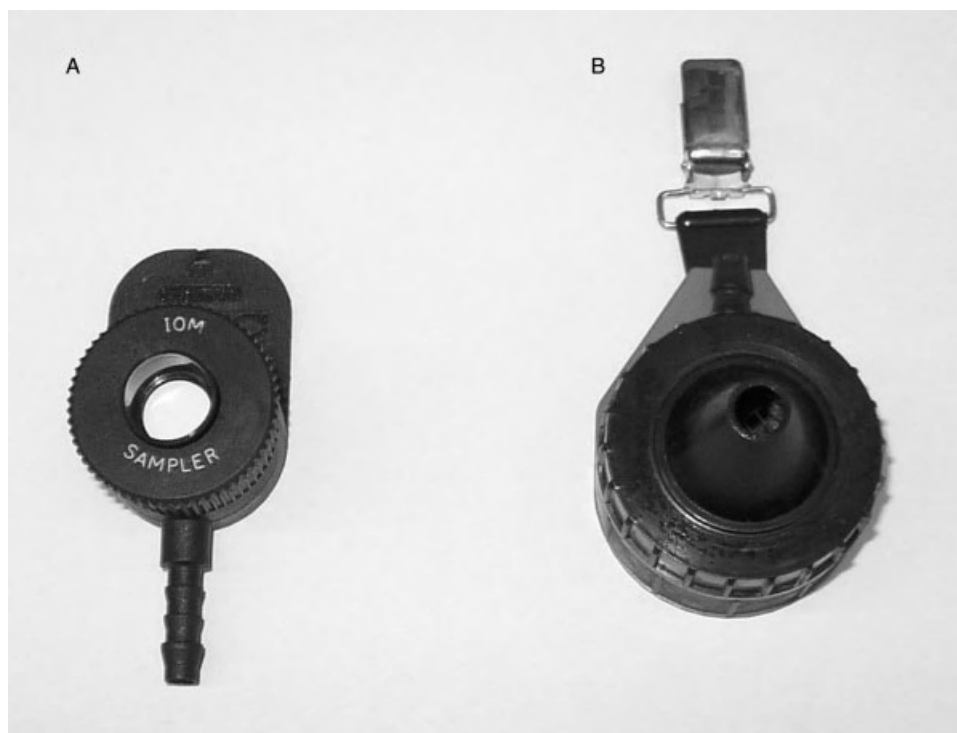


Figure 15.3 Two personal samplers for inhalable aerosols: (a) IOM sampler; (b) GSP conical sampler.



Figure 15.4 IOM static inhalable aerosol sampler.

relevance to health effects is tenuous. The only static sampler designed from the outset to match the inhalability criterion is the IOM 3 l min^{-1} static inhalable aerosol sampler (Mark *et al.*, 1985), shown in Fig. 15.4. It incorporates a number of novel features, but unfortunately is no longer commercially available. The sampler contains a single sampling orifice located in a head that is mounted on top of the cylindrical housing containing the pump, drive and battery pack and rotates slowly about a vertical axis, thereby sampling omnidirectionally. Similar to the IOM personal sampler, the entry orifice forms an integral part of an aerosol-collecting cassette that is located mainly inside the head. Use of this cassette ensures that the overall aspirated aerosol is always assessed.

Many hygienists use personal inhalable samplers as static samplers, and although this is probably reasonable in low wind speeds and fine particles, it is not recommended in high wind speeds and for coarse aerosols.

Practical samplers for respirable aerosol

The history of sampling fine aerosols in workplaces began with the respirable fraction, in par-

ticular with the emergence in the 1950s of the BMRC respirable aerosol criterion. A number of types of sampling device have since been developed. Most have in common the fact that they first aspirate a particle fraction that is assumed to be representative of the total workplace aerosol, from which the desired fine fraction is then aerodynamically separated inside the instrument, using an arrangement whose particle size-dependent penetration characteristics match the desired criterion. It is the fraction that remains uncollected inside the selector and passes through to collect on to a filter (or some other collecting medium) that is the fine fraction of interest. Only those that have been shown to conform to the respirable convention defined in EN 481 will be considered here.

Personal samplers

The most common particle size selectors for personal respirable aerosol samplers are cyclones, which are ideally suited for such purposes, and they have found wide application. In the UK, several conducting plastic versions of the well-known SIMPEDS cyclone (based on the original Higgins and Dewell cyclone) are available. To conform to the respirable convention as defined in EN 481,



Figure 15.5 Conducting plastic personal respirable aerosol sampler.

they must be operated at a flow rate of 2.2 l min^{-1} . There are many other designs available, but care should be taken when purchasing the cyclones that they sample according to the latest sampling conventions (EN 481 or IS 7708). A photograph of a conducting plastic personal cyclone is given in Fig. 15.5.

The French CIP10 (Fig. 15.6) has some interesting and unusual features and so deserves special mention. It incorporates its own built-in pumping unit, consisting of a battery-driven, rapidly rotating polyester foam plug. The aerosol is aspirated through a downwards-facing annular entry and is progressively selected by a combination of mainly gravitational and inertial forces in two static, coarse-grade foam plugs located inside the entry as well as on the finer grade rotating one. As a result of the low-pressure drop characteristics of such foam filtration media, a very high flow rate (by personal sampler standards) can be achieved: up to 10 l min^{-1} . In addition, it has the major benefit over conventional personal samplers of not requiring a separate sampling pump. However, recent tests with the sampler have shown that it is not suitable in wind speeds above 2 m s^{-1} and can

suffer from transfer of large particles to the rotating foam in workplaces where the sampler is likely to be subjected to vibration.

Static samplers

For the respirable aerosol fraction, the personal samplers mentioned above can be used successfully as static samplers. However, flow rates for personal samplers are necessarily low and if aerosol concentrations are expected to be low, higher flow rate static samplers are required. A variety of static samplers for respirable aerosol have been built and successfully used in practical occupational hygiene. There are two main methods of achieving particle size selection: by horizontal gravitational elutriation and centrifugal force. An example of an elutriator is the British 100 l min^{-1} Hexhlet, whereas examples of cyclone samplers include the

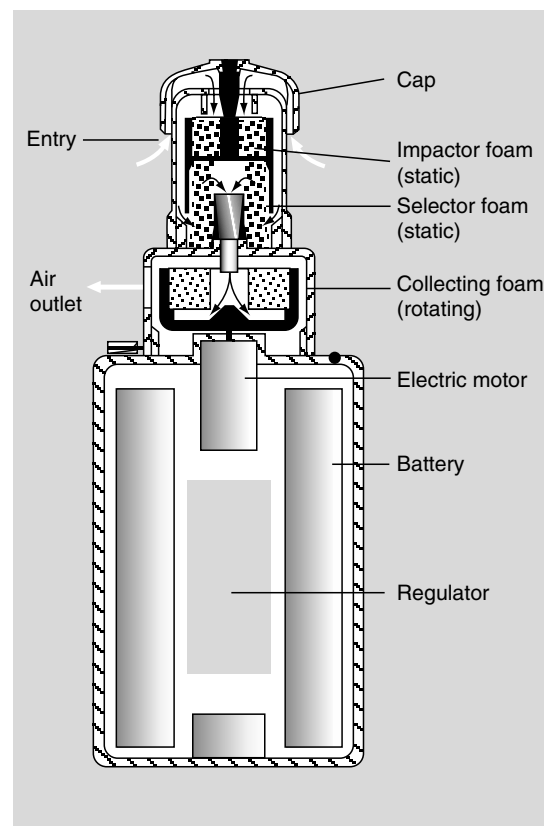


Figure 15.6 CIP10 personal respirable aerosol sampler.

German 50 l min^{-1} TBF50 sampler and the French 50 l min^{-1} CPM3. Although such cyclones can be designed having well-defined penetration characteristics, prediction of performance from theory is more complicated than for horizontal elutriators.

Practical samplers for thoracic aerosol

Personal samplers

Methodology for the sampling of thoracic aerosol in the occupational context was not widely considered prior to the establishment of the CEN/ISO/ACGIH sampling conventions. The MSP personal environmental monitor uses a single-stage impactor designed to select the thoracic fraction according to the earlier American PM₁₀ definition. With a flow rate of 4 l min^{-1} , it has a very sharp sampling curve that excludes some of the large particles allowed in the thoracic convention. This sampler has been widely used in the USA for monitoring personal exposures in non-occupational situations, such as homes and public places, and has formed part of a major study to investigate aerosol exposures both indoors and outdoors.

There are a number of new samplers available for the thoracic fraction. One is a modification of the CIP10 respirable aerosol sampler (described above) in which the foam size selector has been replaced by an inertial particle selection device. The other is a modification to the IOM personal inhalable aerosol sampler, in which a porous foam size selector is positioned in the front half of the cassette behind the inhalable aerosol entry. A similar arrangement has been designed for the GSP inhalable aerosol sampler.

Static samplers

Again, for thoracic aerosols, the personal samplers mentioned above may be operated successfully as static samplers. However, there is a wide range of static samplers used for sampling the PM₁₀ fraction in outdoor and non-industrial indoor environments that can be deployed in the workplace. Flow rates are generally between 5 and 30 l min^{-1} , and both battery- and mains-powered samplers are available.

Multifraction samplers

In recent years, a number of samplers have been developed in which exposure to more than one health-related fraction is measured. These include very simple modifications to the IOM and GSP personal samplers, in which two porous plastic foam plugs inserted into the sampler entry are used to select first the thoracic and then the respirable fractions of the inhalable aerosol sampled. The inhalable fraction is obtained by summing the particle masses collected on the two foams and the filter; the thoracic fraction by summing the particle masses collected on the second foam and the filter; and the respirable is the mass collected on the filter alone.

The RESPICON is an ingenious three-stage device that selects the sampled aerosol using a series of virtual impactors (Fig. 15.7). The aerosol enters through an annular slot and is then selected by the virtual impactors such that the respirable particles are collected on the top filter, the tracheobronchial on the second filter and the extrathoracic on the final filter. The inhalable fraction is obtained by summing the particle masses collected on all three filters, the thoracic fraction by summing the particle masses collected on the first two filters, and the respirable is the mass collected on the top filter.

Investigational instruments

In all of the instruments described above, the sampled aerosol is collected on a filter or some other substrate that may be assessed separately after sampling has been completed. Such instrumentation is suitable when time-averaged measurement can be justified. However, there are occasions when short-term (or even real-time) measurement is required, for example, when investigating the major sources of aerosol emission from an industrial process and the subsequent efficiency of control procedures introduced to minimize that emission. They may also form the basis of an alarm monitoring system of the type described above.

Optical techniques provide an effective means by which aerosols can be assessed in real time, based on the principles of light extinction and scattering.

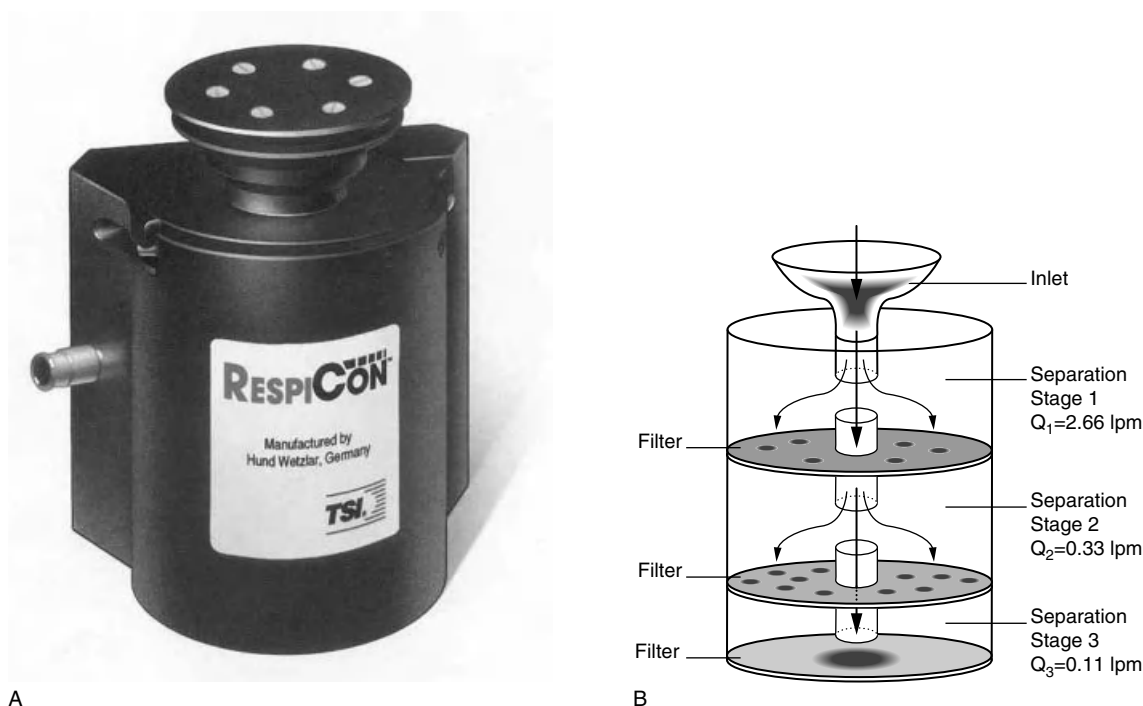


Figure 15.7 RESPICON three-fraction sampler. (a) Complete sampler; (b) schematic diagram.

They provide the great advantage that measurement can be made without disturbing the aerosol – provided of course that the particles can be introduced into the sensing zone of the instrument without loss or change. Their disadvantage is that interactions between light and airborne particles are strongly dependent on particle size and type, and so results are frequently difficult to interpret.

For workplaces, optical instruments generally operate on the basis of the detection of the light scattered by all of the particles passing through the sensitive volume of the instrument. There are many possibilities (e.g. optical geometry, scattering angle) on which to base an instrument, and a correspondingly wide range of instruments has appeared on the market. The most successful have been those designed for the monitoring of aerosol fractions within specific particle size ranges – in particular, the respirable fraction. They mainly involve devices that detect light (laser or infrared) scattered in the near-forward direction, using cyclones or porous foam size selectors to select the respirable particles from aspir-

ated samples. Examples include DataRAM, SKC Hazdust and the R&P DustScan Scout.

There is also a family of light-scattering instruments in which the aerosol enters the sensing region by direct convection without the aid of a pump. These ‘passive’ devices rely on the optical response curve (which is more sensitive for the finer particles) to simulate the respirable size selection. The lack of a pump means that the weight of the instruments can be minimized, allowing them to be carried from location to location. This is also of benefit to the hand-held wand-type instruments such as the Microdust Pro produced by Casella (London). These are widely used in walk-through surveys of workplaces to determine which processes are the major dust sources, and for this purpose the variation in response of the instrument with particle size and refractive index must not be forgotten.

Recently, a number of truly personal direct-reading aerosol monitors have been produced and their performances are currently being evaluated. Early indications show that although they all give



Figure 15.8 A selection of personal direct-reading aerosol monitors. (a) TSI Sidepak; (b) Sibata PDS1; (c) MIE *personalDataRAM*; (d) SKC Split 2.

linear relationships with concentration for the respirable fraction, their responses for the coarser particles in the thoracic and inhalable fractions are very low and variable. The claims of some

manufacturers that their devices will measure the inhalable fraction of workplace aerosols should be treated with caution. They are shown in the photograph of Fig. 15.8.

All light-scattering instruments report results in terms of the *mass* concentration. This is usually derived from the application of a factory-produced calibration factor using a standard test dust such as Arizona Road Dust. This calibration can also be carried out in the workplace using the real aerosol as the calibrant, by comparing the instrument's response either with its own built-in filter or with a collocated gravimetric respirable sampler.

The second type of instrument is based on the interaction between a focused light beam and each individual single particle. Such instruments are referred to as 'optical particle counters'. From light-scattering principles, if an individual particle can be detected and registered electronically, it can be not only counted but also sized (i.e. placed into a given size band or 'channel' based on the magnitude of signal arising from the scattered light). By such means, instruments can be designed which are capable either of counting particles within specified size ranges or of providing an overall particle size distribution. As with aerosol photometers, many practical instruments have evolved within this category, and have been widely used in research both in laboratories and in workplaces. A typical example of this is the GRIMM 1.105. This small, battery-powered monitor provides number- or mass-based size distributions in the particle size range 0.5–15 μm . The use of a backup filter allows calibration of the mass concentrations for the specific aerosol particles sampled. The fibrous aerosol monitor (FAM) is a version that sets out to provide counts of fibrous particles conforming to the 'respirable' fibre definition (as discussed earlier), even in the presence of non-fibrous particles.

A wide range of other types of direct-reading instruments is available (Table 15.1). One is based on the beta-attenuation concept, where the mass of particulate material deposited on a filter, or some other surface, is determined from the reduction in intensity of beta-particles passing through the accumulated layer. In such instruments, the change in attenuation reflects the rate at which particles are collecting on the filter and hence the concentration of the sampled aerosol. One advantage of this approach over optical instruments is that the attenuation of beta-particles

is directly dependent on particulate mass, and is almost independent of aerosol type or particle size distribution.

Another class of devices is what might be referred to as 'vibrational mass balances'. The tapered element oscillator microbalance (TEOM) involves the use of a tapered glass tube that is fixed at the large end and supports a filter at the narrow end. The tube and filter are oscillated and the deposition of particles on the filter causes a change in the resonant frequency of the tube (Fig 15.9). This device is widely used in ambient monitoring stations for providing time series information on PM10 and PM2.5 concentrations out of doors. They are rather bulky and expensive for normal workplaces, but they provide the only true continuous measurement of the mass concentration of airborne particles. A personal version of the TEOM is under development at present by Rupprecht and Patashnick (Albany, USA) and, if successful, it will be the only personal direct-reading mass monitor whose performance will be independent of particle size, shape and refractive index.

An interesting, and potentially very useful, development has been the combination of direct-reading instruments and videotaping of the operation. In this procedure, the operator wears a direct-reading personal monitor, and the contaminant levels of the operational process are superimposed on the video film. Although mostly used for the sampling of gases and vapours, some success has been obtained using various light-scattering instruments as the monitor. It has proved to be a very useful tool in demonstrating to workers ways in which they can reduce their personal exposures.

Errors involved in taking an aerosol sample

As with any measurement process, there are errors associated with that measurement. Aerosol sampling comprises a number of different stages, each one with its own error, which contributes to the overall inaccuracy of the measurement of aerosol concentration.

Table 15.1 Examples of direct-reading instruments for workplace aerosol.

Name	Measurement technique	Flow rate ($l\ min^{-1}$)	Particle fraction	Concentration range* ($\mu g\ m^{-3}$)	Static, portable or personal	Comments
R&P TEOM Series 1100 Particle Mass Monitor	Tapered element oscillating microbalance	3	Total mass, but can be PM10, PM2.5 with inlets	0.006–150	Static†	Particles collected on filter
Casella Microdust Pro hand-held dust monitor	Light-scattering photometer	Passive	Nominally fine	0.001–100 indicated	Portable, possibly personal	Output directly related to mass Hand-held passive device suitable for walk-through surveys
R&P DustScan Scout aerosol monitor	Light-scattering photometer	2	TSP inlet with thoracic and respirable from foam inserts	0.001–100 indicated		Response dependent upon refractive index and size of particles
SKC Hazdust III	Light-scattering photometer	Up to 3	Inhalable, or respirable from different inlets	0.001–200 indicated	Personal	Calibrated with Arizona Road Dust
SKC SplitZ	Light-scattering photometer	2 or passive	Uses IOM inhalable entry, thoracic and respirable from foam inserts	0.001–200 indicated	Personal	Response dependent upon refractive index and size of particles, will not detect larger particles within inhalable convention Design based on inserting small photometer between IOM personal inhalable aerosol entry and filter
DataRAM portable real-time aerosol monitor	Light-scattering photometer	2	'Total' and respirable from different inlets	0.001–100 indicated	Portable, personal	Methodology specifies calibration on site and manufacturer is building up a library of calibrations Optical device calibrated with AC fine test dust May need on-site calibration to give reliable mass measurements as response dependent upon refractive index and size of particles

(Continued)

Table 15.1 Examples of direct-reading instruments for workplace aerosol (*Continued*).

Name	Measurement technique	Flow rate ($l\ min^{-1}$)	Particle fraction	Concentration range* ($\mu g\ m^{-3}$)	Static, portable or personal	Comments
TSI Model 8520 DUSTTRAK aerosol monitor	Light-scattering photometer	1.4–2.4	Particle size range 0.1–10 μm	0.001–100 indicated	Portable	Calibrated with A1 test dust
TSI Side Pak personal aerosol monitor	Light-scattering photometer	0.7–1.8	Particle size range 0.1–10 μm , respirable with cyclone	0.001–20 indicated	Personal	Response dependent upon refractive index and size of particles Compact design with photometer, pump and control circuitry in unit on belt
GRIMM Model 1.105 dust monitor	Optical particle counter	1.2	'Total' and respirable from different inlets and size distribution in eight channels	0.001–100 indicated	Portable	Sampling head positioned on breathing zone and use of tubing minimizes particle loss Optical particle counter with built-in filter for on-site calibration, as mass response may be dependent upon refractive index and size of particles Gives number concentrations also

*Manufacturer's figures.

†Personal version is under field trials.

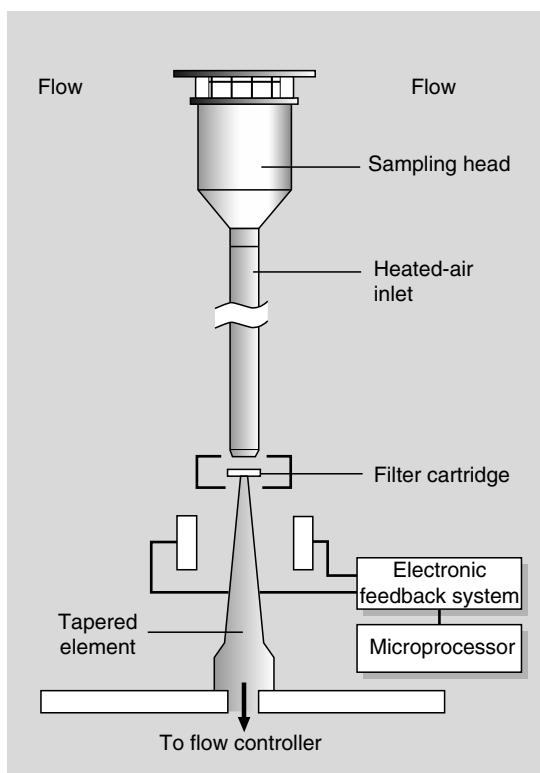


Figure 15.9 Principle of operation of the tapered element oscillating microbalance (TEOM).

There are systematic errors that we can control, such as:

- bias in concentration due to inadequate performance of the sampling head;
- errors in setting and maintaining the required flow rate;
- analytical errors, gravimetric and chemical.

There are also random errors over which we have little control, except in the design of the sampling strategy. These errors are associated with:

- choice of who to sample and where;
- variation in day-to-day aerosol concentration.

These errors all combine to give an overall error or uncertainty in the measurement of aerosol concentration.

Performance of the sampling head

As mentioned above, it is essential that the sampling head is chosen specifically for the aerosol

fraction of interest (i.e. inhalable, thoracic or respirable). Recent experiments have shown that there are wide variations in the sampling efficiencies of sampling heads previously used to collect so-called 'total' aerosol. Studies in a range of different industries have shown that aerosol concentration comparisons of two types of personal sampler for coarse particles can have ratios ranging from 1:1 to 3:1. Similar problems can be found for samplers for the respirable fraction.

It is therefore essential (and now required in Europe) that if sampling is to be undertaken for health-related purposes, suitable samplers should be chosen that have been shown to meet the requirements of the relevant sampling convention specified in European Standard EN 481.

Flow rate setting and control

Errors in the flow rate through samplers can have significant effects on the mass concentration measured. Both the entry efficiency and the selection efficiency of the size selector (if present) can be affected by incorrect flow rate setting and control.

The flow rate through most modern sampling pumps is easily adjusted by means of a single screw and, once set, is maintained by built-in flow control systems that compensate for build-up of particles on the filter. However, despite the fact that some pumps have built-in rotameter flowmeters, it is still necessary to check the flow rate entering the sampler using a primary standard, such as a bubble flowmeter. This is because the rotameters are either fixed in series in the flow line or on the pump exhaust, where flow rate measurement is unreliable. The flowmeter is fixed to the sampler entry so that the true flow rate through the sampler is measured. This process has been speeded up dramatically in recent years by the introduction of automatic bubble flowmeters, such as the Gilibrator and, more recently, automatic dry piston flowmeters.

Analytical errors: gravimetric

The majority of aerosol samples taken in the workplace are analysed gravimetrically. Although this

may seem to be a simple process, a number of precautions are necessary in order to obtain reliable results.

Ideally, all weighing should be carried out in a temperature- and humidity-controlled room set aside specially for the purpose. If full environmental control is not available, then care should be taken to ensure that the room is not subject to large changes in temperature due to, for example, solar gain and time-controlled central heating. The room should be big enough to contain a solid bench upon which the balance is sited and a set of shelves for storing filters and cassettes for conditioning prior to weighing. This conditioning should be allowed before each weighing (i.e. before and after aerosol exposure) and preferably for a period of at least 12 h – overnight is a useful time. Conditioning in a desiccator, which is sometimes used, is not recommended as the filter rapidly gains weight when transferred to the balance and is therefore very difficult to weigh. For some membrane filters made from cellulose esters, PVC, PTFE and polycarbonate, excess surface electrical charge must be neutralized, either with a radioactive source or a high electric field, before accurate weighing can be achieved. Finally, even with all of these precautions, it is essential that a small number of filters (about 10% of the batch) are kept unexposed to aerosol to serve as blank controls to allow for residual changes in filter weight due to moisture uptake and changes in balance performance.

Provided that the above precautions are taken and a suitable balance is used, a weighing accuracy of ± 0.03 mg is easily achievable.

Analytical errors: chemical

Many aerosol samples taken in workplaces contain a variety of different compounds. If measurement of a specific compound within the mixture, such as lead, nickel, cadmium, etc. is required, then the collected sample must be analysed by atomic absorption spectrometry or other methods to determine the concentration of that compound. For solid aerosol particles the main difficulty in the analysis process is ensuring that all of the collected particles are removed from the filter. Digestion of the filters by acid washing or low-temperature

washing are methods generally employed for this purpose. In samplers with integral cassettes, it is important to ensure that the material of the cassette is not leached by the washing process (e.g. metal cassettes should not be used when analysing the particles for metals).

Random errors: variability of exposure

Most operations in workplaces lead to short-term fluctuations in pollutant emissions. If the substance is acutely toxic it is important to measure the peak concentrations, to set short-term occupational exposure limits (STELs) and to instigate control procedures to protect the worker. For most substances, we are interested in the longer term integrated exposure, and full-shift sampling is required. However, even when sampling for a longer period, significant errors in assessing the actual exposure of a worker to aerosols can arise. Day-to-day exposure levels for a given worker can be very variable, as can exposure levels between workers doing the same job and between workers doing different jobs. This variability is random in nature and there is some evidence to suggest that exposure measurements follow a log-normal distribution pattern. Variations in full-shift exposure levels of up to 1000:1 have been observed – large enough to dwarf the instrumental errors described above!

To ensure that a sampling campaign gives realistic estimates of individual exposure, it is essential that the above variability is taken into account when deciding the duration and frequency of sampling and who and where to sample. This topic is the subject of much guidance from the regulatory authorities such as the UK Health and Safety Executive (Health and Safety Executive, 2000) and the US National Institute of Occupational Safety and Health, and a draft CEN standard is soon to be approved by the European Union. It is considered in more detail in Chapter 17.

Special problems

Fibrous aerosol particles

Fibrous aerosols can pose extreme risks to health and so are of special interest. Because of their

unusual morphological properties, and the role of these properties in the aetiology of lung disease, fibres are specifically excluded from the ISO, CEN and ACGIH conventions. Instead, there is a separate rationale for particle size-selective measurement, based on an appreciation of both the nature of particle motion that governs fibre deposition in the deep lung, and the biological effects that influence the fate of the particles after deposition. Thus, it has been a widespread convention since the 1960s to assess 'respirable' fibres in terms of the airborne number concentration, rather than mass concentration used for all other particles. Furthermore, these respirable fibres are not selected by aerodynamic means, but are defined when examined by optical microscopy under phase-contrast conditions as those particles having a length-to-diameter ratio of greater than 3, length greater than $5\ \mu\text{m}$ and diameter less than $3\ \mu\text{m}$.

The practical criteria that have emerged for fibres are based not only on the properties of the particle which can bear directly on possible health effects, but also on the technical means readily available for assessing them. Optical microscopy is relatively cost-effective and straightforward. However, to set an upper limit for fibre diameter of smaller than $3\ \mu\text{m}$ (which might be justified in the light of some of the biological evidence) could result in counting problems as a higher proportion would lie beyond the physical limits of detection by optical means. Therefore, the criteria currently in use for routine analysis are based on pragmatic as well as scientific considerations.

As already indicated, the definition of a 'respirable' fibre is based on purely geometric criteria so that selection is best carried out not aerodynamically but visually under the microscope. This means that, in practical sampling, the main priority is to achieve deposition onto a suitable surface (e.g. a membrane filter), which can then be 'cleared' and mounted for subsequent visual analysis. It follows that actual physical sampling can be very simple, and usually involves the collection of particles directly on to an open filter contained within a downwards-pointing filter holder. The filter holder is generally fitted with a cowl or some other baffle to protect the filter from large airborne material as well as from curious fingers. An example of such a

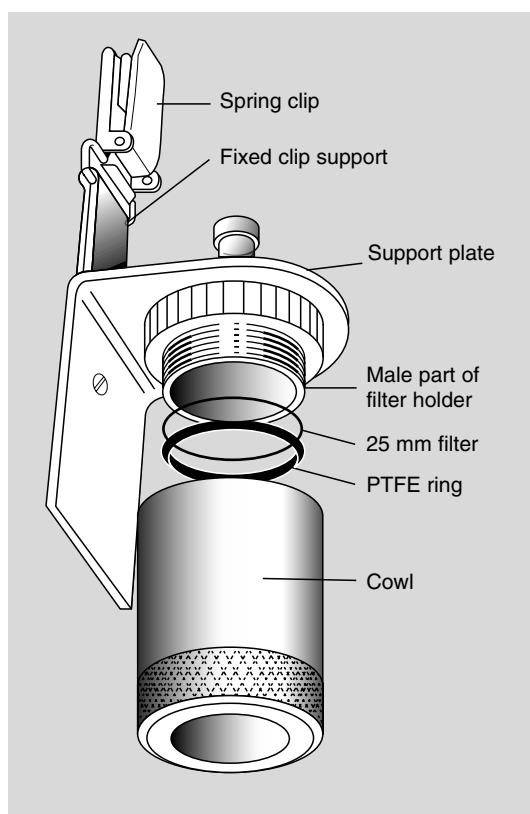


Figure 15.10 Cowed sampling head for asbestos and other fibres.

sampler, as recommended for use in the UK (MDHS39), is given in Fig. 15.10. The filters used are normally membranes made from mixed cellulose esters, which are easily made transparent (cleared) with acetone vapour from commercially available acetone boilers. The samplers are used routinely in both the static mode to monitor asbestos clearance sites, and the personal mode to estimate individual exposure.

Bioaerosols

The term 'bioaerosols' includes a wide range of airborne particles that are derived from living matter. They include micro-organisms ranging in size from submicron viruses to fungal spores which may exceed $200\ \mu\text{m}$. Between these extremes there is a wide variety of bacteria, fungi (including both yeasts and moulds) and spores, and

non-viable fragments of micro-organisms. Some viruses and bacteria are pathogenic (e.g. anthrax) and must be alive to cause harm, whereas allergic responses such as asthma and hay fever, etc. may be caused by fragments of cells as well as the live organism. Examples of all these types of particles can occur at the workplace, which includes farming situations as well as the indoor workplaces usually considered.

The principles of sampling are similar to those described above, but, as well as meeting the sampling criteria for inorganic aerosols, samplers for some bioaerosols must also collect the particles with minimal shear forces and static electricity forces, and must retain the particles in a moist atmosphere. This is because some micro-organisms (such as viruses, etc.) are fragile and are easily killed, thereby rendering them harmless to humans. Despite the wide-ranging occurrence of bioaerosols and the increased understanding of their role in many diseases, guidelines describing standard methods for the sampling of bioaerosols in workplace environments have only recently been published (EN 13098, 2001). In this standard, strategies for measurement are described dependent upon the bioaerosol to be sampled and generic information is given about possible approaches and methods. A large number of reviews concerning sampling methods used to assess bioaerosols have been published. These include reviews covering all applications such as those by Burge and Solomon (1987), Chatigny *et al.* (1989) and Griffiths and DeCosemo (1994). There are many others.

The most commonly used samplers for bioaerosols in workplaces are the Andersen microbial sampler (AMS) shown in Fig. 15.11 and the all-glass impinger (AGI). These are both static instruments that are designed to keep the collected particles alive. The AMS achieves this by collection on to a nutrient agar medium, whereas the AGI relies on particle collection into a liquid reservoir. Other devices designed specifically for collecting bioaerosols that have been used in workplaces include: the Casella slit sampler, the surface air sampler and the Biotest RCS sampler, all of which use agar for particle collection and retention; and cyclones with inlet spray-wetters such as

the Aerojet general liquid scrubber air sampler (Decker *et al.*, 1969). In these devices, collection fluid is continuously injected at the sampler inlet via a hypodermic syringe so that bioaerosol particles deposited on internal walls of the sampler are continuously swept into the collection reservoir.

Although all of the samplers described above have been designed to keep the bioaerosol particles alive (and we do not yet know how well this has been achieved), no attempt has been made to control the physical sampling efficiencies of the samplers. Some recent work on the AMS and the AGI has shown that their physical sampling efficiencies are dependent on both wind speed and particle size in a manner that does not conform to the ISO/CEN health-related sampling conventions.

For personal sampling in the UK, widespread use is made of the IOM personal inhalable sampler to collect the inhalable fraction of bioaerosols, recognizing the fact that fragile micro-organisms will desiccate on the filter. In addition, the development of size-selective porous foam plugs, inserted into the entry of the IOM sampler, has enabled either the thoracic or respirable fractions to be selected in accordance with EN 481. These foams have been shown to provide some form of protection from the desiccation process.

Two main methods of assessment are carried out for bioaerosols. Viable particles are assessed by counting the number of colony-forming units (CFUs) visible after culturing on a suitable growth medium such as agar. The total number of all bioaerosol particles sampled (both viable and non-viable) is determined by microscopy methods, such as epifluorescence microscopy, and, more recently, immunoassay methods, such as ELISA (enzyme-linked immunosorbent assay).

Two-phase compounds

Some compounds such as arsenic, aromatic carboxylic acid anhydrides and isocyanates, etc. can occur both as aerosols and vapours under certain conditions in workplaces. For the first two compounds, the particles are sampled with a normal sampling head (seven-hole sampler in the UK) and are collected on a filter. The vapour penetrating is then collected either on a second, treated filter or

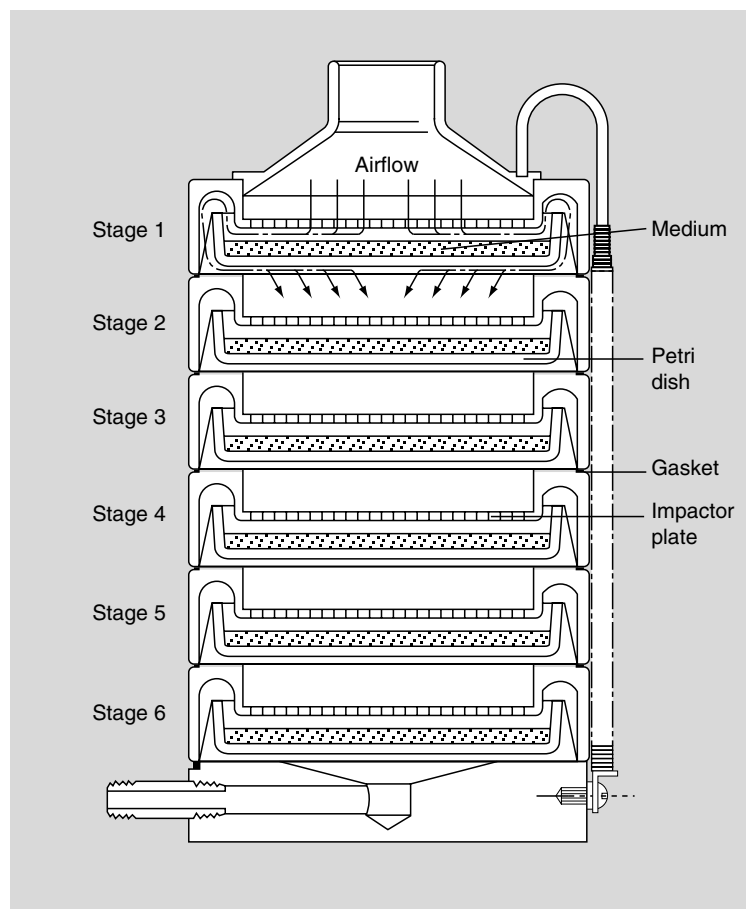


Figure 15.11 Andersen microbial sampler (AMS).

on an absorbent tube. For isocyanates, however, midjet impingers are used, with both aerosol and vapour being collected in the liquid.

These methods are applicable when the two phases are analysed together. However, it is likely that the aerosol and vapour phases will deposit in different regions of the respiratory tract, and so it may be important to determine the concentrations of the two phases separately. For this purpose it is essential that they remain in their original airborne state once collected to ensure that there is no cross-contamination between phases. This situation is very difficult to achieve, as air passing through collected particles may cause off-gassing, and there may also be back-diffusion from the absorbent stage once the sampling pump has been switched off.

A number of different approaches have been proposed. One approach involves the simultaneous sampling of the aerosol and the vapour phase separately. This is achieved by using a diffusive sampler to collect the vapour phase, and an inhalable aerosol sampler fitted with a filter for the particles followed by an adsorbent stage to collect the airborne vapour and the vapour released from the collected particles.

Future developments

The sampling of aerosols in all environments, including occupational, is going through a period of great change, with the emphasis in all

measurements being on international standardization. For aerosols, international agreement has recently been reached for a set of health-related sampling conventions. The next step, which has just been achieved, is the establishment of an agreed test protocol by which the performance of aerosol samplers can be tested for compliance with the health-related conventions. This work involves the classification of samplers according to the range of conditions in which they will give reliable results. Once this has been achieved, and provided that an agreed sampling strategy is followed, there is no reason why reliable, valid measurements of occupational exposures cannot be achieved. These will have the added benefit of being internationally similar.

At present, continuously recording instruments are used as investigational tools to determine major aerosol sources and as educational tools to reduce aerosol exposure by altering work practices. It is expected that this will continue, but with an additional use as alarm monitors in an overall sampling strategy employing both periodic personal sampling supported by continuous monitoring of all work shifts.

With an increasing incidence of allergenic diseases, such as asthma, which are thought to be caused by inhaling bioaerosols, and the increasing production of process micro-organisms, there is a desperate need for standardized guidance on the sampling of bioaerosols. Research work is being carried out at present in laboratories world-wide to develop practical, reliable and standard methods. It is expected that there will not be one universally applicable method but possibly two: one for micro-organisms that must be kept viable, using samplers that both treat the micro-organisms gently and keep them wet; and one where viability is not important and existing methodology for inorganic aerosols can be used.

References

- American Conference of Governmental Industrial Hygienists (1968). *Threshold Limit Values of Airborne Contaminants*. ACGIH, Cincinnati, OH.
- Asbestos International Association (1979). *Recommended Technical Method No. 1: Reference Method for the Determination of Airborne Asbestos Fibre Concentrations at Workplaces by Light Microscopy (Membrane Filter Method)*. AIA Health and Safety Publication, London.
- Burge, H.A. and Solomon, W.R. (1987). Sampling and analysis of biological aerosols. *Atmospheric Environment*, **21**, 451–4.
- Chatigny, M.A., Macher, J.M., Burge, H.A. and Solomon, W.A. (1989). Sampling airborne microorganisms and aeroallergens. In *Air Sampling Instruments for Evaluation of Atmospheric Contaminants*, 7th edn (ed. S.V. Hering). American Conference of Governmental Industrial Hygienists, Cincinnati, OH.
- Comité Européen de Normalisation (1993a). *Workplace Atmospheres. Size Fraction Definitions for Measurement of Airborne Particles*. CEN Standard EN 481.
- Comité Européen de Normalisation (1993b). *Workplace Atmospheres. Requirements and Test Methods for Pumps used for Personal Sampling of Chemical Agents in the Workplace*. CEN 1232.
- Comité Européen de Normalisation (2000). *Workplace Atmospheres. Assessment of Performance of Instruments for Measurement of Airborne Particles*. CEN Standard EN 13205.
- Griffiths, W.D. and DeCosemo, G.A.L. (1994). The assessment of bioaerosols: a critical review. *Journal of Aerosol Science*, **25**, 1425–58.
- Health and Safety Executive (2000). *General Methods for the Gravimetric Determination of Respirable and Total Inhalable Dust*. MDHS 14/3. HSE, London.
- Hering, S.V. (ed.) (1989). *Air Sampling Instruments for Evaluation of Atmospheric Contaminants*, 7th edn. American Conference of Governmental Industrial Hygienists, Cincinnati, OH.
- Lippmann, M. and Harris, W.B. (1962). Size-selective samplers for estimating 'respirable' dust concentrations. *Health Physics*, **8**, 155–63.
- Mark, D. and Vincent, J.H. (1986). A new personal sampler for airborne total dust in workplaces. *Annals of Occupational Hygiene*, **30**, 89–102.
- Mark, D., Vincent, J.H. and Gibson, H. (1985). A new static sampler for airborne total dust in workplaces. *American Industrial Hygiene Association Journal*, **46**, 127–33.
- Orenstein, A.J. (1960). Recommendations adopted by the Pneumoconiosis Conference. In *Proceedings of the Pneumoconiosis Conference*, pp. 619–21. Churchill, London.
- Vincent, J.H. and Mark, D. (1990). Entry characteristics of practical workplace aerosol samplers in relation to the ISO recommendations. *Annals of Occupational Hygiene*, **34**, 249–62.

Further reading

Cohen, B.S. and Hering, S.V. (eds) (2001). *American Conference of Governmental Hygienists: Air Sampling Instruments*, 8th edn. ACGIH, Cincinnati, OH.

Hinds, W.C. (1999). *Aerosol Technology: Properties, Behaviour and Measurement of Airborne Particles*. Wiley-Interscience, New York.

Vincent, J.H. (1989). *Aerosol Sampling: Science and Practice*. Wiley and Sons, Chichester.

Vincent, J.H. (1995). *Aerosol Science for Industrial Hygienists*. Elsevier Science, New York.

Chapter 16

The sampling of gases and vapours: principles and methods

Richard H. Brown

Introduction	Volume fraction
Selection of sampling devices and measurement methods	Diffusive samplers
Grab samplers (short-term samplers – seconds or minutes)	Overview
Evacuated flasks	Principles of diffusive sampling
Passivated canisters	Dimensions of diffusive uptake rate
Flexible plastic containers	Bias due to the selection of a non-ideal sorbent
Continuous active samplers (long-term samplers – hours or days)	Environmental factors affecting sampler performance
Sorbents	Temperature and pressure
Cold traps	Humidity
Sampling bags	Transients
Solid sorbents	The influence of air velocity
Activated charcoal	Calculations
Sample collection	Practical applications
Silica gel	Canisters
Thermal desorption	Charcoal tubes
Coated sorbents	Silica gel
Wet chemistry and spectrophotometric methods	Thermal desorption
Sampling train	Coated sorbents
Calculations	Wet chemistry and spectrophotometric methods
Impinger	Diffusive samplers
Sorbent tube	Quality systems and quality control
	References

Introduction

This chapter discusses the collection and analysis of gases and vapours commonly found in the ambient indoor or workplace environment. It is limited to descriptions of sampling methods for subsequent laboratory analysis. It does not, therefore, include any discussions of automatic analysers, direct-reading instruments, colorimetric indicators, tape samplers or other ‘on-the-spot’ testing devices. However, many of the principles involved are the same.

Selection of sampling devices and measurement methods

There are two basic methods for collecting gas and vapour samples. In one, grab sampling, a sample of contaminated air is collected in a flask, bottle, bag or other suitable container; in the other, called continuous or integrated sampling, gases or vapours are removed from the air and concentrated by passage through a sorbing medium.

The first method usually involves the collection of instantaneous or short-term samples, usually

within a few seconds or a minute, but similar methods can be used for sampling over longer periods.

Grab sampling is of questionable value when:

- the contaminant or contaminant concentration varies with time (unless several samples are taken to establish a concentration profile);
- the concentration of atmospheric contaminants is low (unless a highly sensitive detector is used so that the mass of analyte collected is above the limit of detection); or
- a time-weighted average exposure is desired.

In such circumstances, continuous or integrated sampling is used instead. The gas or vapour in these cases is extracted from air and concentrated by:

- solution in a sorbing liquid;
- reaction with a sorbing liquid (or reagent therein);
- collection on to a solid sorbent.

Collection efficiency must be determined for each case.

Integrated samples are frequently taken with a sampling pump and air metering device. However, in many cases, diffusive sampling may also be used, as discussed later in this chapter.

Of the many alternative approaches to gas and vapour sampling, the best one will depend on the circumstances. Factors that will need to be taken into account include:

- 1 the measurement task;
- 2 the concentration to be determined;
- 3 the time resolution required;
- 4 selectivity to the target gas or vapour and sensitivity to interfering gases and vapours;
- 5 bias, precision (i.e. measurement uncertainty) required;
- 6 susceptibility of the sampler to environmental factors;
- 7 fitness for purpose, e.g. weight, size, durability;
- 8 training requirements for the reliable operation, maintenance and calibration;
- 9 the total cost of purchase and operation, including calibration and maintenance;
- 10 compliance with the performance requirements of appropriate national or local governmental regulations;

11 conformity to the user's quality system.

Having established the requirements, the next step in the selection of a sampling method or device and analytical procedure is to search the available literature. Primary sources are the compendia of methods recommended by the regulatory authorities or governmental agencies in the USA, i.e. the *NIOSH Manual of Analytical Methods* (National Institute for Occupational Safety and Health, 1975) and the *OSHA Analytical Methods Manual* (Occupational Health and Safety Administration, 1979). Recommended methods from other countries, such as the UK (Health and Safety Executive's *Methods for the Determination of Hazardous Substances*), Germany or Sweden, might also be consulted. Other standards are available from the American Society for Testing of Materials (ASTM). Secondary sources are published literature references in, for example, the *Journal of Environmental Monitoring*, *Annals of Occupational Hygiene*, *Analytical Chemistry*, *American Industrial Hygiene Association Journal*, *Applied Occupational and Environmental Hygiene* or books such as the Intersociety Committee's *Methods for Air Sampling and Analysis*.

If a published procedure is not available, one can be devised from theoretical considerations. However, its suitability must be established experimentally before application.

The final stage is to review the performance characteristics of the available methods against the selection criteria that are already established.

Important information on the performance characteristics (points 5 and 10 in the list above) of devices or procedures can be obtained from various sources. These include:

- the manufacturer's instructions for use;
- published commercial technical information;
- technical and research publications;
- national and international standards;
- user groups, e.g. HSE/CAR/WG5, which issues *The Diffusive Monitor*, a newsletter produced since 1988 (obtainable from the Health and Safety Laboratory, Broad Lane, Sheffield, UK).

Grab samplers (short-term samplers – seconds or minutes)

Evacuated flasks

These are containers of varying capacity and configurations. In each case, the internal pressure of the container is reduced, either to near zero (< 1 mb) or to a known absolute pressure. These containers are generally removed to a laboratory for analysis, although it is possible to achieve field readability if the proper equipment and direct-reading instrument are available. Some examples of evacuated flasks are heavy-walled containers, separation flasks and various commercial devices.

Passivated canisters

Stainless steel containers that have been specially treated to reduce sorption effects have been used for collecting trace organic gases, especially the less reactive hydrocarbons and halocarbons.

Flexible plastic containers

Bags are used to collect air samples and prepare known concentrations that can range from parts per billion to more than 10% by volume in air. The bags are commercially available in sizes of up to 250 l. However, 5- to 15-l bags are the most useful to industrial hygienists.

These bags are constructed from a number of materials, including polyester, polyvinylidene chloride, Teflon[®], aluminized Mylar[®] or other fluorocarbons. Bags have the advantages of being light, non-breakable, inexpensive to ship and simple to use. However, they should be used with caution because storage stabilities for gases, memory effects from previous samples, permeability, precision and accuracy of sampling systems vary considerably.

Plastic bags should be tested before they are used. Such testing should be carried out under ambient conditions that approximate those of the sampling environment. Some general recommendations are available in the published literature for the use of such bags for air sampling. A good review of specific applications is Schuette (1967). More recently, Posner and Woodfin (1986)

made a useful systematic study of five bag types for the sampling of six organic vapours; they concluded that Tedlar[®] bags are best for short-term sampling, whereas aluminized Mylar[®] bags are better for long-term storage prior to analysis. Storage properties, decay curves and other factors, however, will vary considerably from those reported for a given gas or vapour because sampling conditions are rarely identical. Each bag, therefore, should be evaluated for the specific gas or gas mixture for which it will be used.

Continuous active samplers (long-term samplers – hours or days)

Sorbers

The sorption theory of gases and vapours from air by solution, as developed by Elkins *et al.* (1937), assumes that gases and vapours behave like perfect gases and dissolve to give a perfect solution. The concentration of the vapour in solution is increased during air sampling until equilibrium is established with the concentration of vapour in the air. Sorption is never complete, however, because the vapour pressure of the material is not reduced to zero but is only lowered by the solvent effect of the sorbing liquid. Some vapour will escape with continued sampling, but it is replaced. Continued sampling will not increase the concentration of vapour in solution once equilibrium is established.

According to formulas developed by Elkins *et al.* (1937) and verified by Gage (1960) in his experiments with ethylene oxide, the efficiency of vapour collection depends on:

- the volume of air sampled;
- the volume of the sorbing liquid; and
- the volatility of the contaminant being collected.

Efficiency of collection, therefore, can be increased by cooling the sampling solution (reducing the volatility of the contaminant), increasing the solution volume by adding two or more bubblers in series or altering the design of the sampling device. Sampling rate and concentration of the vapour in air are not primary factors that determine collection efficiency.

Sorption of gases and vapours by chemical reaction depends on the size of the air bubbles produced in the bubbler, the interaction of contaminant with reagent molecules, the rate of the reaction and a sufficient excess of reagent solution. If the reaction is rapid and a sufficient excess of reagent is maintained in the liquid, complete retention of the contaminant is achieved regardless of the volume of air sampled. If the reaction is slow and the sampling rate is not low enough, collection efficiency will decrease.

Four basic sorbers used for the collection of gases and vapours are:

- simple gas washing bottles;
- spiral and helical sorbers;
- fritted bubblers; and
- glass bead columns.

Sampling and sorbent capacities of these sorbers are found in the work of the Intersociety Committee (1988). The function of the sorbers is to provide sufficient contact between the contaminant in the air and the sorbing liquid.

Petri, Dreschel and midget impingers are examples of simple gas washing bottles. They function by applying a suction to an outlet tube, which causes sample air to be drawn through an inlet tube into the lower portion of the liquids contained in these sorbers. They are suitable for collecting non-reactive gases and vapours that are highly soluble in the sorbing liquid; the sorption of methanol and butanol in water, esters in alcohol and organic chlorides in butyl alcohol are examples. They are also used for collecting gases and vapours that react rapidly with a reagent in the sampling medium. High collection efficiency is achieved, for example, when toluene di-isocyanate is hydrolysed to toluene diamine in Marcali solution; hydrogen sulphide reaction with cadmium sulphate and ammonia neutralization by dilute sulphuric acid are other examples.

Several methods for testing the efficiency of a sorbing device are available:

- series testing when enough samplers are arranged in series so that the last sampler does not recover any of the sampled gas or vapour;
- sampling from a dynamic standard atmosphere or from a gas-tight chamber or tank containing a known gas or vapour concentration;

- comparing results obtained with a device known to be accurate; and
- introducing a known amount of gas or vapour into a sampling train containing the sorber being tested.

Cold traps

Cold traps are used for collecting materials in liquid or solid form primarily for identification purposes. Vapour is separated from air by passing it through a coiled tube immersed in a cooling system, i.e. dry ice and acetone, liquid air or liquid nitrogen. These devices are used when it is difficult to collect samples efficiently by other techniques. Water is extracted along with organic materials and two-phase systems result.

Sampling bags

Bags (as used for grab sampling) can also be used for collecting integrated air samples. Samples can be collected for 8 h, at specific times during the day or over a period of several days. The bags may be mounted on workers for personal sampling or may be located in designated areas.

Solid sorbents

Activated charcoal

Charcoal is an amorphous form of carbon formed by partially burning wood, nutshells, animal bones and other carbonaceous materials. A wide variety of charcoals are available; some are more suitable for liquid purification, some for decolorization, and others for air purification and air sampling.

Ordinary charcoal becomes activated charcoal by heating it with steam to 800–900°C. During this treatment, a porous, submicroscopic internal structure is formed, which gives it an extensive internal surface area as large as 1000 m² per gram of charcoal. This greatly enhances its sorption capacity.

Activated charcoal is an excellent sorbent for most organic vapours. During the 1930s and 1940s it was used in the well-known activated charcoal apparatus for the collection and analysis

of solvent vapour. The quantity of vapour in the air sample was determined by a gain in weight of the charcoal tube. Lack of specificity, accuracy and sensitivity of the analysis and the difficult task of equilibrating the charcoal tube, however, discouraged further use.

Renewed interest in activated charcoal as a sorbent for sampling organic vapours appeared in the 1960s. The ease with which carbon disulphide extracts organic vapours from activated charcoal and the capability of microanalysis by gas chromatography are the reasons for its current popularity.

Sample collection

The sorption capacity of a sampler, i.e. the volume of air that can be collected without loss of contaminant, depends on the sampling rate, the quantity of sorbent, the sorbent surface area, the density of active sites and bulk density, the volatility of the contaminant, the ambient humidity and the concentration of contaminant in the workroom air. For many organic vapours, a sample volume of 10 l can be collected without significant loss in NIOSH-recommended tubes. A breakthrough of more than 20% in the back-up section indicates that some of the sample was lost. Optimum sample volumes are found in NIOSH procedures.

It is always best to refer to an established procedure for proper sampling rates and air sample volumes. In the absence of such information, breakthrough experiments must be performed before field sampling is attempted. Normally, these experiments are conducted using dynamic standard atmospheres prepared at twice the exposure limit (normally the threshold limit value, TLV) and 80% relative humidity to give a suitable margin of safety to the measured breakthrough volume. See MDHS 3 and MDHS 4 (Health and Safety Executive, 1981–2001) for the preparation of known concentrations.

After the procedure has been validated, field sampling may be performed. Immediately before sampling, the ends of the charcoal tube are broken, rubber or Tygon[®] tubing is connected to the back-up end of the charcoal tube, and air is drawn through the sampling train with a calibrated battery or electrically driven suction pump. A per-

sonal or area sample may be collected. The duration of the sampling is normally 8 h but may be as short as 15 min or up to 24 h, depending on the information required. When sampling is completed, plastic (but not rubber) caps are placed on the ends of the tube.

For each new batch of charcoal tubes, test samples must be prepared to determine the analytical blank, collection efficiency, storage and recovery characteristics for a given contaminant. This may be achieved by introducing a known amount of the contaminant into a freshly opened charcoal tube, passing clean air through it to simulate sampling conditions and carrying through its analysis with the field samples. Another charcoal tube, not used to sample, is opened in the field and used as a field blank.

The first step in the analytical procedure is to desorb the contaminant from the charcoal. An early drawback to using charcoal for air sampling was the difficulty in recovering samples for analysis. Steam distillation was only partially effective. Extraction with carbon disulphide has been found quite satisfactory in many instances, although for the more volatile vapours, thermal desorption may also be used (see below).

The most frequently used liquid desorbant is carbon disulphide. Unfortunately, carbon disulphide does not always completely desorb the sample from charcoal. Recovery varies for each contaminant and batch of charcoal used. The extent of individual recovery must be determined experimentally and correction for desorption efficiency applied to the analytical result. Over a narrow range of analyte concentrations, as used in the NIOSH validations (NIOSH, 1977), this desorption efficiency is essentially constant, but it may vary widely over larger concentration ranges, particularly for polar compounds. Desorption efficiency can also be affected by the presence of water vapour and other contaminants. NIOSH recommends that methods be used only where the desorption efficiency is greater than 75%; ideally, it should be greater than 90%.

The practical desorption step in charcoal analysis is also critical because, upon the addition of carbon disulphide to charcoal, the initial heat of reaction may drive off the more volatile com-

ponents of the sample. This can be minimized by adding charcoal slowly to precooled carbon disulphide. Another technique is to transfer the charcoal sample to vials lined with Teflon[®] septum caps and to introduce the carbon disulphide with an injection needle. The sealed vial will prevent the loss of any volatilized sample. Headspace analysis is also possible.

It should be emphasized that carbon disulphide is a highly toxic solvent that produces severe health effects on the cardiovascular and nervous systems. An appropriate risk assessment must be undertaken before using carbon disulphide, and any indicated control measures implemented, e.g. containment and protective gloves.

Silica gel

Silica gel is an amorphous form of silica derived from the interaction of sodium silicate and sulphuric acid. It has several advantages over activated charcoal for sampling gases and vapours: polar contaminants are more easily desorbed by a variety of common solvents; the extractant does not usually interfere with wet chemical or instrumental analyses; amines and some inorganic substances for which charcoal is unsuitable can be collected; and the use of highly toxic carbon disulphide is avoided.

One disadvantage of silica gel is that it will sorb water. Silica gel is electrically polar and polar substances are preferentially attracted to active sites on its surface. Water is highly polar and is tenaciously held. If enough moisture is present in the air or if sampling is continued long enough, water will displace organic solvents (which are relatively non-polar in comparison) from the silica gel surface. With water vapour at the head of the list, compounds in descending order of polarizability are alcohols, aldehydes, ketones, esters, aromatic hydrocarbons, olefins and paraffins. It is obvious, therefore, that the volume of moisturized air that can be effectively passed over silica gel is limited.

Despite this limitation, silica gel has proven to be an effective sorbent for collecting many gases and vapours. Even under conditions of 90% humidity, relatively high concentrations of benzene, toluene and trichloroethylene are quantitatively

sorbed on 10 g of silica gel from air samples collected at the rate of 2.5 l min^{-1} for periods of at least 20 min or longer. Under normal conditions, hydrocarbon mixtures of two- to five-carbon paraffins, low-molecular-weight sulphur compounds (H_2S , SO_2 , mercaptans) and olefins may be collected without breakthrough on silica gel at dry ice-acetone temperature if the sample volume does not exceed 10 l. Significant losses of ethylene, methane, ethane and other light hydrocarbons occur if the sampling volume is extended to 30 l.

Many of the same considerations apply to silica gel tubes as to the charcoal tubes; the sampling capacity and desorption efficiency for the compound of interest should be determined before use, or a reliable, officially established method should be used. A variety of desorption solvents will be needed for desorbing specific compounds with high efficiency; polar desorption solvents, such as water or methanol, are commonly applied.

Thermal desorption

Because of the high toxicity and flammability of carbon disulphide and the labour-intensive nature of the solvent desorption procedure, a useful alternative is to desorb the collected analyte thermally. Except in a few cases, this is not practical with charcoal as sorbent because the temperature needed for desorption (e.g. 300°C) would result in some decomposition of the analytes. Graphitized carbon (e.g. Carbo-graph) or porous polymer sorbents (e.g. Tenax, Chromosorb 106) are used instead. Of these, Tenax has the lowest thermal desorption blank (typically less than a few nanograms per gram of sorbent when properly conditioned), but only modest sorption capacity compared with carbon. The advantages of thermal desorption over solvent extraction have been recognized for ambient (Zlatkis *et al.*, 1973; Pellizzari, *et al.* 1975), workplace (Brown and Purnell, 1979) and indoor air applications (Wolkoff, 1995).

The thermal desorption procedure typically uses larger tubes than the NIOSH method; usually 200–500 mg of sorbent are used, depending on type. Desorption can be made fully automatic and analysis is usually carried out by gas

chromatography. Some desorbers also allow automatic selection of sample tubes from a multiple-sample carousel. The whole sample can be transferred to the gas chromatograph, resulting in greatly increased sensitivity compared with the solvent desorption method. Alternatively, some desorbers allow the desorbed sample to be held in a reservoir from which aliquots are withdrawn for analysis, but then the concentrating advantage is reduced.

The main disadvantage of thermal desorption directly with an analyser is that it is essentially a 'one-shot' technique; normally, the whole sample is analysed. This is why many such methods are linked to mass spectrometry. However, with capillary chromatography, and instrumentation now available, it is possible to split the desorbed sample before analysis and, if desired, the vented split can be collected and reanalysed. Alternatively, the desorbate can be split between two capillary columns of differing polarity.

Coated sorbents

Many highly reactive compounds, for example isocyanates and lower molecular weight aldehydes, are unsuitable for sampling directly onto sorbents, because they either are unstable or cannot be recovered efficiently. In addition, some compounds may be analysed more easily, or with greater sensitivity, by deriving them first, which can sometimes be achieved during the sampling stage.

Wet chemistry and spectrophotometric methods

Several gases and vapours may be analysed by wet chemical methods or by ultraviolet spectrophotometry. Spectrophotometric methods have now been replaced largely by direct-reading instruments or detector tubes, or by high-performance liquid chromatography (HPLC) or other instrumental techniques.

Sampling train

Except for grab samplers and diffusive samplers, sampling devices are used in conjunction with a

sampling pump and air-metering device. To avoid contaminating the metering device and pump, these are usually placed downstream of the sampler during the sampling period. However, because many samplers introduce back-pressure, the sampling train should be precalibrated using an external flowmeter upstream of the sampling head. The sampling train should also be calibrated after sampling, and preferably should be calibrated periodically during sampling.

Calculations

The collected sample is analysed, either directly if it is a gas phase or impinger sample, or after desorption if it is collected on a solid sorbent, using appropriate gas or liquid standard solutions to calibrate the analytical instrument. Gas phase samples give a result directly in ppm (v/v), but other types of samples will give a mass of analyte per collected sample, or a concentration, which can be converted to a mass by multiplying by the sample volume.

The mass concentration of the analyte in the air sample is then calculated using the following equations:

Impinger

$$C = (m - m_{\text{blank}})/(E_s V) \quad (16.1)$$

where C = mass concentration of analyte in air (mg m^{-3}), m = mass of analyte in sample (μg), E_s = sampling efficiency, m_{blank} = mass of analyte in blank (μg) and V = volume of air sample (l).

Sorbent tube

$$C = (m_1 + m_2 - m_{\text{blank}})/(E_d V) \quad (16.2)$$

where m_1 = mass of analyte on first tube section (μg), m_2 = mass of analyte on back-up tube section (if used) (μg) and E_d = desorption efficiency corresponding to m_1 .

Note: If it is desired to express concentrations reduced to specified conditions, e.g. 25°C and 101 kPa, then,

$$C_{\text{corr}} = C(101/P) \times (T/298) \quad (16.3)$$

where P = actual pressure of air sampled (kPa) and T = absolute temperature of air sampled (K).

Volume fraction

The volume fraction of the analyte in air, in ppm (v/v), is:

$$C' = C_{\text{corr}}(24.5/M) \quad (16.4)$$

where M = molecular mass of the analyte of interest (g mol^{-1}).

Diffusive samplers

Overview

A diffusive sampler is a device that is capable of taking samples of gas or vapour pollutants from the atmosphere at a rate controlled by a physical process, such as diffusion through a static air layer or permeation through a membrane, but does not involve the active movement of the air through the sampler. The adjective 'passive' is sometimes used in describing these samplers and should be regarded as synonymous with 'diffusive'.

This type of diffusive sampler should not be confused with the annular or aerosol denuders, which not only rely on diffusion to collect the gas or vapours, but also upon the air in question being simultaneously drawn through the annular inlet into the sampler. Aerosol particles have diffusion coefficients too low to be collected on the annular inlet and are trapped on a back-up filter.

Principles of diffusive sampling

A general overview is given in Berlin *et al.* (1987). A specific review with environmental applications is given in Brown (1992).

The mass of the analyte that can diffuse to a suitable sorbent within a certain time is determined by the equation that is derived from Fick's first law of diffusion:

$$m_s = AD(\rho_1 - \rho_2)t/1 \quad (16.5)$$

where A = cross-sectional area of diffusion path (cm^2), D = coefficient of diffusion ($\text{cm}^2 \text{s}^{-1}$),

l = length of diffusion path (cm), m_s = mass of analyte sorbed by diffusion (ng), t = sampling time (seconds), ρ_1 = actual mass concentration at the beginning of the diffusion layer ($l = 0$) (mg m^{-3}) and ρ_2 = actual mass concentration at the end of the diffusion layer (mg m^{-3}).

Ideally, ρ_1 is equal to the concentration of the given analyte in the air outside the diffusive sampler (ρ) and ρ_2 equals zero ('zero sink' condition). In that case, the magnitude of the diffusive uptake rate, AD/l , is dependent only on the diffusion coefficient of the given analyte and on the geometry of the diffusive sampler used.

In practice, there are a number of factors that can give rise to non-ideal behaviour, so that:

$$m_s = AD\rho tk/1 \quad (16.6)$$

where k = correction factor for non-ideal behaviour and ρ = actual mass concentration of analyte in air (mg m^{-3}).

Dimensions of diffusive uptake rate

For a given concentration ρ in milligrams per cubic metre of gas or vapour, the diffusive uptake rate is given by:

$$U = m_s/\rho t' \quad (16.7a)$$

where U = sampling rate ($\text{cm}^3 \text{min}^{-1}$) and t' = sampling time (min).

Although the uptake rate, U , has dimensions of cubic centimetre per minute, this is really a reduction of nanograms per milligrams per cubic metre per minute [$\text{ng}(\text{mg m}^{-3})^{-1} \text{min}^{-1}$] and does not indicate a real volumetric flow of (analyte in) air.

Diffusive uptake rates are very often quoted in units of $\text{ng ppm}^{-1} \text{min}^{-1}$. These are practical units, as most environmental analysts use ppm for concentrations of gases and vapours. The dependency of uptake rates on temperature and pressure is explained later. Thus, for a given concentration (ppm) of gas or vapour, the sampling rate is given by:

$$U' = m_s/\phi t' \quad (16.7b)$$

where U' = sampling rate ($\text{ng ppm}^{-1} \text{min}^{-1}$) and ϕ = actual mass concentration of analyte in air (ppm, v/v).

Ideal and practical diffusive uptake rates are related by:

$$U' = (U \times M \times 293 \times P) / (24.0 \times T \times 101) \quad (16.8)$$

Bias due to the selection of a non-ideal sorbent

The performance of a diffusive sampler depends critically on the selection and use of a sorbent or collection medium that has high sorption efficiency. The residual vapour pressure of the sampled compound at the sorbent surface (ρ_2) will then be very small in comparison to the ambient concentration, and the observed uptake rate will be close to its ideal steady-state value, which can usually be calculated from the geometry of the sampler and the diffusion coefficient of the analyte in air.

In the case when a weak sorbent is used, then ρ_2 in Equation 16.5 is non-zero and m_s/t will decrease with the time of sampling. Hence, U in Equation 16.6 will also decrease with the time of sampling. The magnitude of this effect is dependent on the sorption isotherm of the analyte and sorbent concerned, and may be calculated with the aid of computer models.

Another manifestation of the same effect is back-diffusion, sometimes called reverse diffusion. This can happen some time after sampling has started, when the vapour pressure of the analyte at the sorbent surface, ρ_2 , is greater than the external concentration, ρ_1 , for example if a sampler is first exposed to a high concentration and then to a much lower or even zero concentration. This type of exposure profile can occur in certain applications, and the magnitude of any error introduced will depend on whether the period of high concentration occurs at the beginning, middle or end of the sampling period. The phenomenon has been discussed in detail by Bartley and co-workers and a simple test proposed (Bartley *et al.*, 1987) to give an estimate of the maximum bias to be expected between a pulsed exposure and an exposure to a constant concentration. This normally provides the basis for the sampler calibration. The extent of back-diffusion can also be modelled theoretically.

It is therefore desirable to choose a sorbent with high sorption capacity and low vapour pressure of the sorbed material or of the reaction product formed by a reactive sorbent.

Environmental factors affecting sampler performance

Temperature and pressure

For an ideal diffusive sampler, the dependence of U on absolute temperature and pressure is governed by that of the diffusion coefficient of the analyte. The latter dependence is given by:

$$D = f(T^{n+1}, P^{-1}) \quad (16.9)$$

with $0.5 < n < 1.0$.

Hence, the dependence of U , expressed in units of $\text{cm}^3 \text{min}^{-1}$ or equivalent is:

$$U = f(T^{n+1}, P^{-1}) \quad (16.10)$$

When U' is expressed in units of $\text{ng ppm}^{-1} \text{min}^{-1}$ or equivalent by application of Equation 16.8 then the dependence is given by:

$$U' = f(T^n) \quad (16.11)$$

In this case, the dependence will be of the order of 0.2–0.4% K^{-1} . In the case of a non-ideal sampler, the temperature dependence of U' may be compensated by the temperature dependence of the sorption coefficient of the analyte. In any case, accurate knowledge of the average temperature and pressure during the sampling period is important for a correct application of Equations 16.7a and 16.7b.

Humidity

High humidity can affect the sorption capacity of hydrophilic sorbents, such as charcoal and Molecular Sieve. This will normally reduce the sampling time (at a given concentration) before saturation of the sorbent occurs, when sampling becomes non-linear because of a significant ρ_2 term in Equation 16.5. High humidity can also alter the sorption behaviour of the exposed inner wall of tube-type samplers or draught screen, particularly if condensation occurs.

Transients

Simple derivations of Fick's law assume steady-state conditions, but in the practical use of diffusive samplers the ambient level of pollutants is likely to vary widely. The question then arises whether a sampler will give a truly integrated response (ignoring sorbent effects) or will 'miss' short-lived transients before they have had a chance to be trapped by the sorbent. The issue has been discussed theoretically and practically and shown not to be a problem, provided that the total sampling time is well in excess of (say 10 times) the time constant of the diffusive sampler, i.e. the time a molecule takes to diffuse into the sampler under steady-state conditions. The time constant, τ , for most commercial samplers is between about 1 and 10 s. τ is given by:

$$\tau = l^2/D \quad (16.12)$$

where τ = time constant of diffusive sampler (s), l = length of diffusion path (cm) and D = coefficient of diffusion ($\text{cm}^2 \text{s}^{-1}$)

The influence of air velocity

Effect of low and high wind speeds

Ambient air face velocity and orientation can affect the performance of a diffusive sampler because they may influence the effective diffusion path length. The diffusive mass uptake of a sampler (Equation 16.6) is a function of the length, l , and the cross-sectional area, A , of the diffusion gap within the sampler. The nominal diffusion path length is defined by the geometry of the sampler and is the distance between the sorbent surface and the external face of the sampler. The cross-sectional area is also defined by the geometry of the sampler and, if the cross-section of the diffusion gap is not constant along its length, is defined by the narrowest portion. The effective length, l , is not necessarily the same as the nominal length, and may be greater or less, depending on circumstances.

Under conditions of low external wind speeds, the effective diffusion path length may be increased. This is because a 'boundary layer' exists

between the stagnant air within the sampler and the turbulent air outside and contributes to the effective diffusion path length, l . In reality, there is an area outside the sampler where there is a transition between static air and turbulent air, but this is equivalent to an extra length (δl) of static air that must be included in the value of l . The magnitude of δl depends on the external geometry of the sampler, being roughly proportional to the linear cross-section of the sampler collection surface, where this surface is flat. It also decreases with increasing air velocity. Its significance depends on the value of the nominal path length of the diffusive sampler. Thus, a sampler with a small cross-section and long internal air gap will be relatively unaffected by air velocity, whereas a short, fat sampler will be significantly affected. This is borne out in practice, as has been demonstrated with samplers of varying length. Low sampling rates are observed at low air velocities, but increase to a plateau value as the boundary layer effect becomes insignificant.

Under conditions of high external wind speeds, the effective diffusion path length may be decreased. This is because turbulent air disturbs the static air layer within the sampler, which reduces the effective air gap by a factor δl . The magnitude of δl is small, provided the length-diameter ratio of the sampler air gap is greater than 2.5–3, or it can be avoided, or greatly reduced, by incorporating a draught shield, e.g. a stainless steel screen or plastic membrane.

The overall effect is therefore sinusoidal.

Consequence for different sampler geometries

Tube-type samplers are typically unaffected by low air velocities but those without a draught shield may be affected by high speeds.

Badge-type samplers generally have a large surface area and small air gap, so that they may be more affected by air velocity than tube designs and typically require a minimum face velocity of between 0.5 and 0.2 m s^{-1} . Some badges with an inadequate draught shield are also affected at high air velocities.

Radial diffusive samplers require a minimum face velocity of about 0.25 m s^{-1} .

Calculations

The method of calculation of atmospheric concentrations is essentially the same as for pumped samplers, i.e. the collected sample is analysed and the total mass of analyte on the sampler is determined. Then, as before:

$$C = (m_1 + m_2 - m_{\text{blank}})/(E_d V) \quad (16.13)$$

Note: m_2 is relevant only to samplers with a back-up section, and an additional multiplication factor may be needed to account for differing diffusion path lengths to primary and back-up sections. m_2 and E_d are ignored for liquid sorbent badges.

V , the total sample volume, is calculated from the effective sampling rate (l min^{-1}) and the time of exposure (min).

This calculation gives C in mg m^{-3} ; strictly speaking, an appropriate sampling rate for the ambient temperature and pressure should be made.

Alternatively, sampling rates can be expressed in units such as $\text{ng ppm}^{-1} \text{min}^{-1}$ (dimensionally equivalent to $\text{cm}^3 \text{min}^{-1}$), when C' is calculated directly in ppm:

$$C' = [(m_1 + m_2 - m_{\text{blank}}) \times 1000]/(E_d U t') \quad (16.14)$$

Practical applications

Canisters

The US Environmental Protection Agency (US EPA) has used passivated canisters for ambient air analysis alongside sorbent tubes (Varns *et al.*, 1990). A more recent study (Ballesta *et al.*, 1998) has shown significant under- and over-estimations of some compounds by the canister method compared with a continuously cycling gas chromatograph.

Charcoal tubes

Air sampling procedures using activated charcoal are widely used by industrial hygienists and form the basis of the majority of the official analytical methods for volatile organic compounds recom-

mended by NIOSH and OSHA. There is also an ISO standard in preparation (ISO/FDIS 16017-1, 2001, VOCs by pumped tube/solvent desorption) and a method in the HSE/MDHS series (MDHS 96, 2000, VOCs by pumped tube/solvent desorption) (HSE, 1981-2001).

Analytical information on selected NIOSH procedures is given in Table 16.1. In general, the NIOSH procedures use a 100-mg charcoal tube (with 50-mg back-up), but very volatile analytes may require a larger tube. The NIOSH study showed that the charcoal tube method is generally adequate for hydrocarbons, halogenated hydrocarbons, esters, ethers, alcohols, ketones and glycol ethers that are commonly used as industrial solvents. Compounds with low vapour pressure and reactive compounds (e.g. amines, phenols, nitro-compounds, aldehydes and anhydrides) generally have lower desorption efficiencies than charcoal and require alternative sorbents such as silica gel or porous polymers for collection, or alternative reagent systems for recovery.

Inorganic compounds, such as ozone, nitrogen dioxide, chlorine, hydrogen sulphide and sulphur dioxide, react chemically with activated charcoal and cannot be collected for analysis by this method.

Carbon disulphide is usually the extraction solvent of choice for charcoal tubes, but this solvent may not always be ideal. Reference to Table 16.1 will indicate that carbon disulphide is the recommended desorption solvent for non-polar compounds, whereas a variety of desorption cocktails are required for the more polar compounds. Difficulties arise, therefore, when sampling mixtures of polar and non-polar compounds because each will give poor recoveries with the other's desorption solvent. Several more universal solvents have been investigated, but none of these has achieved wide recognition. In such circumstances, it may be necessary to take two or more samples at the same time and desorb each one with a different solvent.

Silica gel

In some cases, silica gel tubes (in similar sizes to the NIOSH range of charcoal tubes) are used instead

Table 16.1 Collection and analysis of gases and vapours (solvent desorption).

Method name	Test compounds	Sorbent*	Desorption solvent	NIOSH method no.
Alcohols I	<i>t</i> -Butyl alcohol, isopropyl alcohol, ethanol	C	99:1 CS ₂ :2-butanol	1400
Alcohols II	<i>n</i> -Butyl alcohol, isobutyl alcohol, <i>s</i> -butyl alcohol, <i>n</i> -propyl alcohol	C	99:1 CS ₂ :2-propanol	1401
Alcohols III	Allyl alcohol, isoamyl alcohol, methyl isobutyl carbinol, cyclohexanol, diacetone alcohol	C	99:5 CS ₂ :2-propanol	1402
Alcohols IV	2-Butoxyethanol, 2-ethoxyethanol, 2-methoxyethanol	C	99:5 CH ₂ Cl ₂ :methanol	1403
Amines: aromatic	Aniline, <i>o</i> -toluidine, 2,4-xylidine, <i>N,N</i> -dimethyl- <i>p</i> -toluidine, <i>N,N</i> -dimethylaniline	S	95% ethanol	2002
Aminoethanol compounds	2-Aminoethanol, 2-dibutylaminoethanol, 2-diethylaminoethanol	S	80% ethanol	2007
Esters I	<i>n</i> -Amyl acetate, <i>n</i> -butyl acetate, 2-ethoxyethyl acetate, ethyl acrylate, methyl isoamyl acetate, <i>n</i> -propyl acetate, etc.	C	CS ₂	1450
Hydrocarbons: BP 36–126°C	Benzene, toluene, pentane through to octane, cyclohexane, cyclohexene	C	CS ₂	1500
Hydrocarbons: aromatic	Benzene, cumene, naphthalene, etc.	C	CS ₂	1501
Hydrocarbons: halogenated	Chloroform, tetrachloroethylene, <i>p</i> -dichlorobenzene, bromoform, etc.	C	CS ₂	1003
Ketones I	Acetone, cyclohexanone, di-isobutyl ketone, 2-hexanone, methyl isobutyl ketone, 2-pentanone	C	CS ₂	1300
Ketones II	Camphor, ethyl butyl ketone, mesityl oxide, 5-methyl-3-heptanone, methyl <i>n</i> -amyl ketone	C	99:1 CS ₂ :methanol	1301
Naphthas	Kerosine, petroleum ether, rubber solvent, Stoddard solvent, etc.	C	CS ₂	1550
Nitro-benzenes	Nitrobenzene, nitrotoluene, 4-chloronitrotoluene	S	Methanol	2005
Nitroglycerin and ethylene glycol dinitrate		T	Ethanol	2507
Pentachloroethane		R	Hexane	2517
Tetrabromoethane		S	Tetrahydrofuran	2003
Vinyl chloride		C	CS ₂	1007

*C, charcoal; S, silica gel; T, Tenax; R, Porapak R.

of charcoal tubes. NIOSH recommends such tubes for a variety of more polar chemicals such as amines, phenols, amides and inorganic acids (Table 16.1).

Thermal desorption

Thermal desorption has been adopted as a (non-exclusive) recommended method for the determin-

ation of volatile organic compounds in the UK (HSE/MDHS 72, 1993) (HSE, 1981–2001) Germany and the Netherlands, but it is less widely accepted in the USA. NIOSH has relatively few methods based on thermal desorption (compared with those which use solvent desorption). US EPA (1984) has a number of methods based on thermal desorption and mass spectrometry, particularly method TO-17. There is also an International

Organization for Standardization (ISO) standard (ISO 16017-1, 2001, VOCs by pumped tube/thermal desorption).

Desorption efficiency is usually 100% for the majority of common solvents and similar compounds in a boiling range of approximately 50–250°C. Thus, the analysis of complex mixtures is easier than for charcoal or silica gel solvent desorption methods. However, if a wide boiling range is to be covered, more than one sorbent may be required. Thus, gasoline may be monitored by a Chromosorb 106 tube and carbon tube in series. Extensive lists of recommended sampling volumes and minimum desorption temperatures for Tenax and other sorbents are given in Brown and Purnell (1979) and the HSE Method MDHS 72 (HSE, 1981–2001).

Coated sorbents

Methods have been developed that use coated sorbents, either sorbent tubes or coated filters. Table 16.2 lists a number of such methods. An ISO standard (ISO 16000-3, 2001, Formaldehyde and Other Carbonyl Compounds – Active Sampling Method) uses the DNPH reagent and HPLC.

Wet chemistry and spectrophotometric methods

There are two primary compendia of methods: the AIHA Analytical Chemistry Committee (1965) and the Intersociety Committee (1988).

Diffusive samplers

A variety of diffusive samplers have been described and only a selection of the major types manufactured can be described here. Diffusive equivalents to the more familiar pumped methods exist for nearly all types; the main exception being the direct collection of gas samples when the nearest equivalent is an evacuated canister. Thus, the diffusive equivalent of the charcoal tube is the charcoal badge such as the 3M OVM or the SKC 575 Passive Sampler; and a diffusive equivalent of the thermal desorption method is the Perkin–Elmer tube. There are also diffusive devices based on reagent-impregnated solid supports, usually for specific analytes.

In general, the regulatory authorities have been reluctant to accept diffusive monitoring methods, except in the UK and the Netherlands where several such methods have been adopted as non-exclusive recommended methods. Extensive lists of recommended sampling rates for solvent desorption methods are given in the HSE Method MDHS 88 (HSE, 1981–2001) and in an ISO standard (ISO 16200-2, 2000, VOCs by Diffusive Sampler/Solvent Desorption). Extensive lists of recommended sampling rates for thermal desorption methods are given in the HSE Method MDHS 80 (HSE, 1981–2001) and in a draft ISO standard (ISO/FDIS 16017-2, 2001, VOCs by Diffusive Tube/Thermal Desorption).

An ISO draft standard for aldehydes is in preparation (ISO/DIS 16000-4, 2000, Formaldehyde – Passive/Diffusive Sampling Method).

Table 16.2 Collection and analysis of gases and vapours (coated sorbents).

<i>Test compounds</i>	<i>Sorbent</i>	<i>Matrix*</i>	<i>Method no.</i>
Acetaldehyde	2 - (Hydroxymethyl) piperidine on Supelpak 20N	T	NIOSH 2538
Acrolein	2 - (Hydroxymethyl) piperidine on Supelpak 20N	T	NIOSH 2501
Arsenic trioxide	Sodium carbonate	F	NIOSH 7901
Butylamine	Sulphuric acid	T	NIOSH S138
Di-isocyanates/isocyanate group	1 - (2-Methoxypyridyl) piperazine	F + bubbler	NIOSH 5521, MDHS25/3
Di-isocyanates	1 - (2-Pyridyl) piperazine	F	OSHA 42, 47
Formaldehyde	<i>N</i> -benzylethanolamine on Supelpak 20F	T	NIOSH 2502
Methylene dianiline	Sulphuric acid	F	NIOSH 5029

*T, sorbent tube; F, filter.

Quality systems and quality control

Canister sampling should be checked by using certified reference gas standards when available.

Several inter-laboratory quality assurance schemes that apply to the charcoal tube method have been developed. One of these is the Proficiency Analytical Testing (PAT) Program and the Laboratory Accreditation Program of the American Industrial Hygiene Association (AIHA). Another is the Health and Safety Executive (HSE) Workplace Analysis Scheme for Proficiency (WASP). Details of these programmes may be obtained from The Laboratory Accreditation Coordinator, AIHA, 2700 Prosperity Ave., Suite 250, Fairfax, Virginia 22031, USA, and the WASP Coordinator, Health and Safety Laboratory, Broad Lane, Sheffield S3 7HQ, UK.

The WASP scheme also includes test samples appropriate to the thermal desorption technique, at both occupational and ambient concentration levels. Certified Reference Materials are available from the EC BCR (Community Bureau of Reference) for aromatic hydrocarbons (CRM 112) and chlorinated hydrocarbons (CRM 555).

References

- Analytical Chemistry Committee (1965). *Analytical Abstracts*. American Industrial Hygiene Association, Akron, OH.
- Ballesta, P.P., Field, R.A. and De Saeger, E. (1998). *Field Intercomparison of VOC Measurements*. EC Report EUR 18085 EN. ERLAP, JRC, Ispra, Italy.
- Bartley, D.L., Deye, G.J. and Woebkenberg, M.L. (1987). Diffusive monitor test: performance under transient conditions. *Applications of Industrial Hygiene*, **2**, 119–22.
- Berlin, A., Brown, R.H. and Saunders, K.J. (eds) (1987). *Diffusive Sampling: An Alternative Approach to Workplace Air Monitoring*. CEC Pub. No. 10555EN. Commission of the European Communities, Brussels.
- Brown, R.H. (1992). Diffusive sampling. In *Clean Air at Work* (eds R.H. Brown, M. Curtis, K.J. Saunders and S. Vandendriessche), pp 141–8. EC Publ. no. EUR 14214, Brussels.
- Brown, R.H. and Purnell, C.J. (1979) Collection and analysis of trace organic vapour pollutants in ambient atmospheres. The performance of a Tenax-GC adsorbent tube. *Journal of Chromatography*, **178**, 79–90.
- Elkins, H.B., Hobby, A.K. and Fuller, J.E. (1937). The determination of atmospheric contaminants; I: Organic halogen compounds. *Journal of Industrial Hygiene*, **19**, 474–85.
- Gage, J.C. (1960) The efficiency of absorbers in industrial hygiene air analysis. *Analyst*, **5**, 196–203.
- Health and Safety Executive (1981–2001, in series). *Methods for the Determination of Hazardous Substances*. Health and Safety Laboratory, Sheffield, UK.
- Intersociety Committee (1988). *Methods of Air Sampling and Analysis*, 3rd edn. Lewis Publishers, Chelsea, MI.
- National Institute for Occupational Safety and Health (1975). *NIOSH Manual of Analytical Methods*, 2nd edn. DHEW (NIOSH) Pub. no. 75–121 (1975); 3rd edn. DHEW (NIOSH) Pub. no. 84–100 (1984, revised 1990); 4th edn. DHHS (NIOSH) Pub. no. 94–113 (1994).
- National Institute for Occupational Safety and Health (1977). *Documentation of the NIOSH Validation Tests*. DHEW (NIOSH) Publ. no. 77–185.
- Occupational Safety and Health Administration (1979). *OSHA Analytical Methods Manual*. OSHA Analytical Laboratories, Salt Lake City, UT. Available from ACGIH, Cincinnati, OH.
- Pellizzari, E.D., Bunch, J.E., Carpenter, B.H. and Sawicki, E. (1975). Collection and analysis of trace organic vapour pollutants in ambient atmospheres. *Environmental Science and Technology*, **9**, 552–60.
- Posner, J.C. and Woodfin, W.J. (1986). Sampling with gas bags. 1. Losses of analyte with time. *Applied Industrial Hygiene*, **1**, 163–8.
- Schuette, F.J. (1967). Plastic bags for collection of gas samples. *Atmosphere and Environment*, **1**, 515–19.
- US Environmental Protection Agency (1984). *EPA Compendium of Methods for the Determination of Toxic Organic Compounds in Ambient Air*. US EPA, Washington, DC.
- Varns, J.L., Mulik, J.D. and Williams, D. (1990). Passive sampling devices and canisters: their comparison in measuring air toxics during a field study. 219–23.
- Wolkoff, P. (1995). Volatile organic compounds – sources, measurements, emissions, and the impact on indoor air quality. *Indoor Air, Suppl.* **3**.
- Zlatkis, A., Lichtenstein, H.A. and Tishbee, A. (1973). Concentration and analysis of volatile organics in gases and biological fluids with a new solid adsorbent. *Chromatographia*, **6**, 67–70.

Chapter 17

Noise

Kerry Gardiner

Introduction	Composite partitions
Noise	Sound in enclosed spaces
Basic acoustics	Reverberation time
Sound quantities	Measurement of hearing
Bel scales	Standard hearing
Properties of the decibel scale	The pure tone audiometer
Equal source levels	Test conditions
Unequal source levels	Limitations of audiometry
Background noise	Further tests
Loudness	Noise exposure and health
Frequency analysis	Noise immission level
Instrumentation	Hearing conservation
Estimation of dB(A) level from octave band levels	Control of noise exposure levels
Measurement of fluctuating levels	Noise reduction at source
Equivalent continuous sound level (L_{eq})	Control of the transmission path
Single event noise exposure level (L_{AX})	Control of noise exposure for the receiver
Statistical levels (L_n)	Ear protection
Noise dosimeter (dosimeter)	Earmuffs
Aural comfort	Earplugs
Noise criteria curves	Attenuation of earmuffs and earplugs
Noise rating curves	Calculation of received noise level when ear protection
Noise and materials	is worn
Absorption coefficient	Other tests on ear protection
Transmission coefficient	Survey report
Sound insulation	References

Introduction

This chapter aims to describe the basic physical properties of sound, the parameters by which it is measured and the means by which it can be controlled. Vibration is dealt with specifically in Chapter 18.

Noise

Basic acoustics

Sound is the form of energy that is detected by the hearing mechanism. This sensation is produced

when the eardrum is vibrated by a minute, fluctuating pressure change in the air inside the ear canal. This fluctuation has, in turn, been caused by a disturbance such as the vibrating cone of a loudspeaker or turbulent jet, by a vibrating machine panel or the vocal cords. Sound is propagated through materials by the longitudinal oscillation of individual molecules and interaction with adjacent molecules (hence it cannot pass through a vacuum).

The simplest form of vibration that a source can exhibit is known as simple harmonic motion (SHM). This motion is exhibited by the molecules of the propagating materials at a rate determined

by the bulk modulus (κ) and density (ρ) of the material. It may be shown that the velocity of sound (c) in a material is given by:

$$c = \sqrt{\frac{\kappa}{\rho}} \quad (17.1)$$

When sound passes through air, the vibration of air particles causes a minute fluctuating pressure known as the acoustic pressure, which is superimposed on the existing atmospheric pressure (about $101\,325\text{ N m}^{-2}$); the smallest detectable acoustic pressure is in the order of $2 \times 10^{-5}\text{ N m}^{-2}$. This is much smaller than diurnal atmospheric pressure changes and the Eustachian tube ensures that the air pressures in the middle ear and

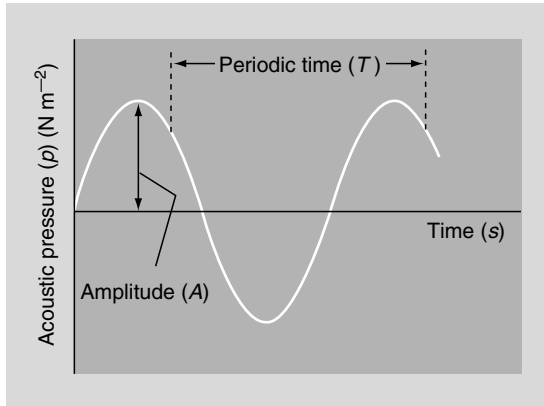


Figure 17.1 Change in acoustic pressure with time for a pure tone.

ear canal are equalized so that the eardrum is free to respond to small acoustic pressures.

For the simplest sounds, the acoustic pressure (p) (Fig. 17.1) may be described by a sinusoidal function.

$$p = A \frac{\sin}{\cos} \omega t \quad (17.2)$$

where A is the amplitude (i.e. the maximum value) (N m^{-2}), t is time (s) and ω is angular frequency (rad s^{-1}). This equation is cyclic over a periodic time of $T = 2\pi/\omega$, i.e. has the same value at times $t = t_0$ and $t = t_0 + 2\pi/\omega$. It is more convenient to express the periodicity in terms of its frequency (f): the number of cycles produced in 1 s, where $f = 1/T$ (s^{-1} or hertz, Hz). The audible frequency range is about 15 to 18 kHz, although acoustic pressures do exist at lower (infrasonic) and higher frequencies (ultrasonic).

Because of the finite value of sound velocity, points along the path of propagation exhibit phase differences. In Fig. 17.2, points 'a' and 'b' are separated by a distance X (m). The time taken for the sound to travel from 'a' to 'b' is X/c seconds. Hence the phase at 'b' is delayed by X/c on 'a' so that if:

$$p = A \frac{\sin}{\cos} \omega t \text{ at 'a'} \quad (17.3)$$

then

$$p = A \frac{\sin}{\cos} \omega \left(t - \frac{X}{c} \right) \text{ at 'b'}$$

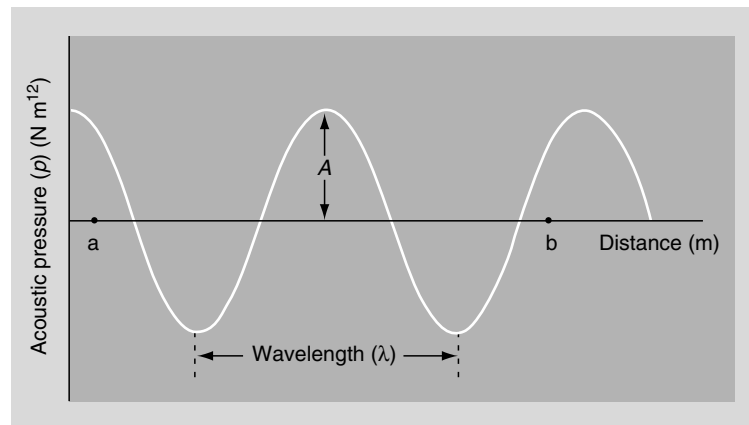


Figure 17.2 Change in acoustic pressure with distance for a pure tone.

and

$$p = A \frac{\sin(\omega t - kX)}{\cos(\omega t - kX)}$$

where k is ω/c (a wave constant). Where X has a value such that $X = 2\pi, 4\pi, 6\pi$, etc., 'a' and 'b' are said to be in phase.

Two adjacent points that are in phase are separated by a distance known as the wavelength (λ). Hence:

$$k\lambda = 2\pi$$

therefore

$$\lambda = Tc \quad (17.4)$$

(i.e. the time taken for the sound to travel through a distance equal to λ is T). It follows that $f\lambda = c$, so that the wavelength of sounds of different frequencies may be calculated.

As the velocity of sound in air is approximately 340 m s^{-1} at ground level, the wavelengths of audible sounds will range from 23 m (15 Hz) to 19 mm (18 kHz). The physical dimensions of objects encountered in buildings are also of this range and so the behaviour of sound in factories, for example, is greatly dependent on its frequency. In general, propagation of high-frequency sound is very directional: when high-frequency sounds meet a barrier, reflection occurs. Low-frequency sounds tend to diffract around barriers and to be generally non-directional.

Sound quantities

The rate at which sound energy leaves its source is known as 'sound power' (W) (measured in watts). A source of sound approximating to a point will produce a spherical sound field, so that the sound power is dissipated over an ever-increasing area. Thus the quantity of sound entering the ear depends on the sound power and its distance from the source (r). The average amount of energy passing through a unit area in unit time is known as the 'sound intensity' (I) (W m^{-2}). From Fig. 17.3:

$$I = \frac{\text{sound power}}{\text{surface area of sphere}} = \frac{W}{4\pi r^2}$$

hence

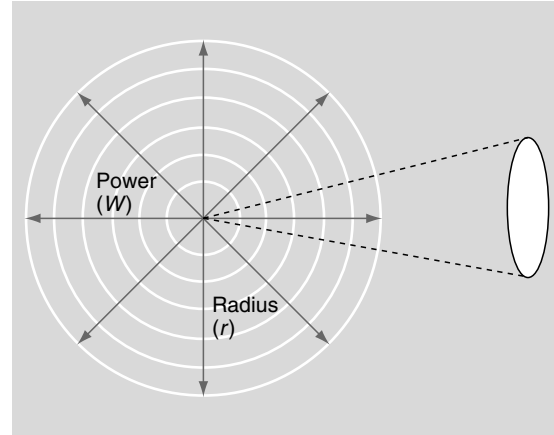


Figure 17.3 Distribution of sound propagated from a point source.

$$I \propto \frac{1}{r^2} \text{ (an inverse square law)} \quad (17.5)$$

If a plane-reflecting barrier is placed immediately behind the source, a hemispherical field is produced, so that the surface area is halved and the intensity doubled. If the field is further modified by reflection, the intensity will also change. In general:

$$I = \frac{QW}{4\pi r^2} \quad (17.6)$$

where Q is the directivity factor. ($Q = 2$ for a hemispherical field and 4 for a quarter-spherical field, etc.)

The value of acoustic pressure corresponding to the energy content of the sound wave is not the amplitude, as this value is never maintained by the wave, but is the average in time of the square of the pressure. This may be shown to be the root-mean-square (rms) value (p_{rms}), which is $A/\sqrt{2}$ for a sinusoidal function (Fig. 17.4). Furthermore:

$$I_{\text{av}} = \frac{p_{\text{rms}}^2}{\rho c} \quad (17.7)$$

where ρc is termed the characteristic impedance of the medium (for air $\rho c = 1.2 \times 340 \approx 400$ rayls). This relationship shows that an average intensity of $10^{-12} \text{ W m}^{-2}$ produces an rms acoustic pressure of $2 \times 10^{-5} \text{ N m}^{-2}$ in air.

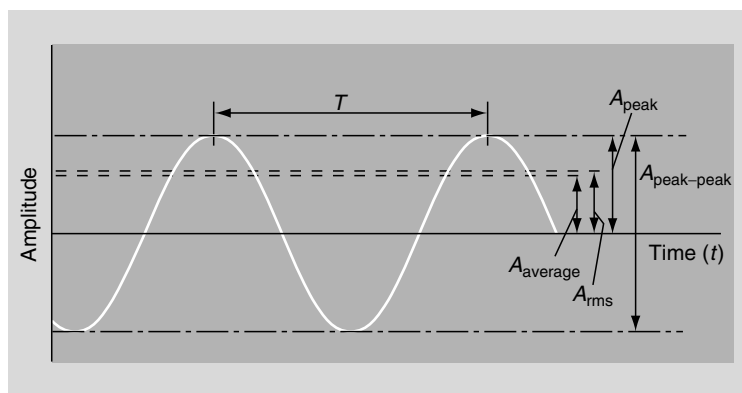


Figure 17.4 Relationship between root-mean-square (rms), peak and average values for a sinusoidal signal (from Waldron, 1989).

Bel scales

The Weber–Fechner law states that the change in a physiological response at a stimulus is proportional to the relative change in the stimulus, i.e.:

$$\delta R \propto \frac{\delta S}{S} \quad (17.8)$$

where δR is the change in response and S is the stimulus (e.g. sound intensity). The response to a stimulus change from S_1 to S_2 may be found by integrating the above expression:

$$R_2 - R_1 = K \ln \frac{S_2}{S_1} \quad (17.9)$$

where K is the constant of proportionality.

The unit of response is chosen so that the constant of proportionality becomes unity when logarithms to the base 10 are used. These units are bels:

$$\text{Response (in bels)} = \log_{10} \frac{S_2}{S_1} \quad (17.10)$$

This logarithmic scale has the property of ‘compressing’ the very wide stimulus range encountered in acoustics ($10^{-12} \text{ W m}^{-2}$ up to 1 W m^{-2}) into more manageable figures. However, the bel is a rather large unit and decibels (dB) are more practicable. Hence:

$$\text{Change in dB} = 10 \log_{10} \frac{S_2}{S_1} \quad (17.11)$$

It is important to realize that the decibel scale is a scale of comparison and can compare sound levels

on either side of a wall, each end of a ‘silencer’ or a worker’s hearing, with an accepted norm.

In noise measurement, S_1 is often given an agreed reference value so that noise level scales are produced. The most commonly encountered scales are listed below.

1 Sound intensity level scale:

$$\text{Reference intensity} = 10^{-12} \text{ W m}^{-2}$$

$$\text{Sound intensity level} = 10 \log_{10} \left(\frac{I}{10^{-12}} \right) \quad (17.12)$$

where I is the intensity of interest (W m^{-2}).

2 Sound power level scale:

$$\text{Reference intensity} = 10^{-12} \text{ W}$$

$$\text{Sound power level} = 10 \log_{10} \left(\frac{W}{10^{-12}} \right) \quad (17.13)$$

where W is the power of interest (W).

3 Sound pressure level scale. As it is the square of the sound pressure of a wave that is proportional to its intensity, the sound pressure level (SPL) is defined as:

$$\text{SPL} = 10 \log_{10} \frac{p_{\text{rms}}^2}{p_{\text{ref}}^2} \quad (17.14)$$

where p_{rms} is the sound pressure of interest. The reference pressure (p_{ref}) is chosen to correspond, in air, with the reference sound intensity (I_{ref}) of $10^{-12} \text{ W m}^{-2}$ and ρc of 400 rayls, i.e.:

$$I_{\text{ref}} = \frac{p_{\text{ref}}^2}{400} \quad (17.15)$$

therefore

$$p_{\text{ref}} = 2 \times 10^{-5} \text{ N m}^{-2}$$

and

$$\text{SPL (dB)} = 20 \log_{10} \left(\frac{p_{\text{rms}}}{2 \times 10^{-5}} \right)$$

For sounds in air, the intensity level and sound pressure level are numerically equal.

Properties of the decibel scale

When more than one source of sound is encountered, combined sound level may be calculated by finding the total amount of sound intensity occurring, and calculating the new sound intensity level from that.

Equal source levels

Example: find the combined sound intensity level when two similar sources of 40 dB each are heard together:

$$40 \text{ dB} = 4 \text{ bels} = \log_{10} \left(\frac{I}{10^{-12}} \right) \quad (17.16)$$

where I is the sound intensity of each source.

Total combined intensity = $2I$

therefore

$$\begin{aligned} \text{Combined intensity level} &= 10 \log_{10} \left(\frac{2I}{10^{-12}} \right) \\ &= 10 \log_{10} \left(\frac{2 \times 10^{-8}}{10^{-12}} \right) \\ &= 10 \log_{10} 2 + 10 \log_{10} 10^4 \\ &= 3 + 40 = 43 \text{ dB} \end{aligned}$$

This doubling of intensity will always give an extra 3 dB. If three similar sources are combined, the combined level will be almost 5 dB (i.e. $10 \log_{10} 3$) above the individual levels; four sources give an increase of 6 dB.

As a normal ear cannot distinguish a change of < 1 dB, even under ideal listening conditions, fractions of decibels are rarely used in practice.

Unequal source levels

The contribution of the lower level is quite small and so the increase will be < 3 dB.

Example: find the combined level when similar sounds of 60 dB and 70 dB are heard together:

$$60 \text{ dB} = 6 \text{ bels} = 10 \log_{10} \left(\frac{I_1}{10^{-12}} \right)$$

$$70 \text{ dB} = 7 \text{ bels} = 10 \log_{10} \left(\frac{I_2}{10^{-12}} \right)$$

therefore

$$\frac{I_1}{10^{-12}} = 10^6 \text{ and } \frac{I_2}{10^{-12}} = 10^7 \quad (17.17)$$

(NB: I_1 is only 10% of I_2)

$$\begin{aligned} \text{Combined intensity level} &= 10 \log_{10} \left(\frac{I_1 + I_2}{10^{-12}} \right) \\ &= 10 \log (10^6 + 10^7) \\ &= 10 \log_{10} (10^7 \times 1.1) \\ &= 70 + 0.4 \\ &= 70 \text{ dB} \end{aligned}$$

(i.e. no increase on the higher level)

(This approximation is justified as the normal variation of even supposedly steady noise is always > 1 dB).

As levels separated by 10 dB (or more) produce no significant increase on the higher level when combined, and identical levels give a 3-dB increase when combined, estimation of the combined levels of sounds separated by < 10 dB may be made mentally.

Example: find the combined level when similar sounds of 54 dB and 47 dB are heard together:

Overestimate: 54 dB and 54 dB combine \Rightarrow 57 dB

Underestimate: 54 dB and 44 dB combine \Rightarrow 54 dB

Combined level must be between 54 and 57 dB, say 55 dB (actually 54.8 dB).

Background noise

The total noise level existing in any location is made up of noise from many different sources. In a factory, for example, there will be a certain noise level when the plant is turned off. If this background noise level is > 10 dB below the plant noise level, the measured level will be that due to the plant. If the background noise level and plant

noise levels are equal, the total level measured will be 3 dB greater. As the background noise cannot be removed, the true level of the plant noise must be calculated from measurements of the background noise alone (i.e. with plant turned off) and the total level (i.e. with plant turned on).

Example: background noise alone = 75 dB and total level of plant and background = 80 dB:

$$75 \text{ dB} = 10 \log_{10} \left(\frac{I_b}{10^{-12}} \right)$$

$$\text{and } 80 \text{ dB} = 10 \log_{10} \left(\frac{I_b + I_p}{10^{-12}} \right) \quad (17.18)$$

therefore

$$\frac{I_b}{10^{-12}} = 10^{7.5} \text{ and } \frac{I_b}{10^{-12}} + \frac{I_p}{10^{-12}} = 10^8$$

$$\begin{aligned} \text{Level of plant alone} &= 10 \log_{10} \left(\frac{I_p}{10^{-12}} \right) \\ &= 10 \log_{10} (10^8 - 10^{7.5}) \\ &= 78 \text{ dB} \end{aligned}$$

where I_b is the intensity of background noise alone and I_p is the intensity of plant noise alone.

Loudness

The frequency response of the ear is not linear, the ear being the most sensitive to sounds in the 1- to 5-kHz frequency range and particularly insensitive at low frequencies. Loudness is the subjective assessment of sound quantity and has a complex relationship with the sound pressure level actually presented to the ear. When the sound pressure levels of pure tones of different frequencies that are judged to be equal in loudness are plotted, these 'equal loudness curves' (Fig. 17.5) exhibit 'dips' at around 4 kHz and 12 kHz which are due to one-quarter and three-quarter wave resonance in the ear canal. Furthermore, these equal loudness curves are not parallel, the ear's response becoming more linear as the level increases. As loudness level depends on frequency and sound pressure level, a new scale was introduced to facilitate comparisons of loudness level. Sounds that are equal in loudness level are assigned the same numerical value of phons. Hence, all points on any one curve bear the same value in phons taken as numerically equal to the number of decibels the curve possesses at 1 kHz. A doubling of loudness occurs when the loudness level increases by 10 phon.

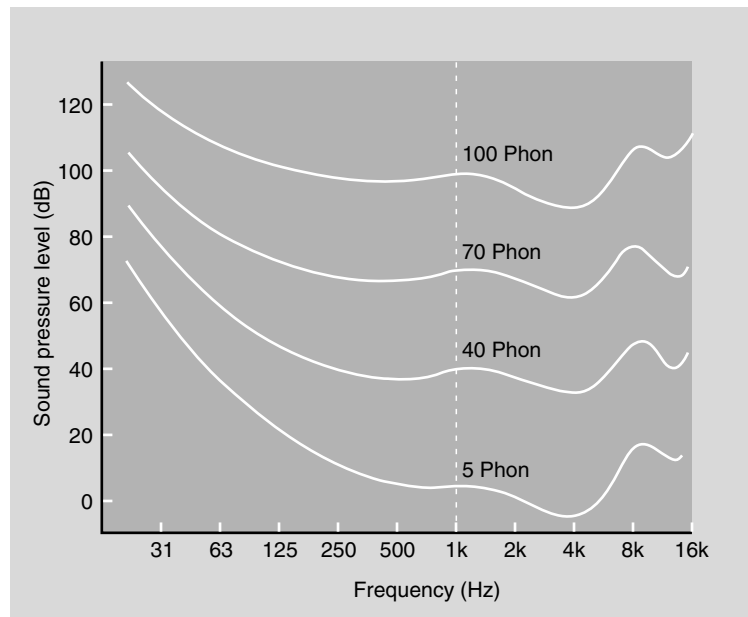


Figure 17.5 Equal loudness curves for pure tones.

Frequency analysis

As the perception of sounds depends on level and frequency, a full investigation of noise must include the measurement of the sound pressure level of each frequency present. This would be an arduous task if the total frequency range were to be covered. As the ear is fairly insensitive to low and very high frequencies, a reduced range is acceptable in nearly all cases (a range of 45 Hz to 11 200 Hz is most common).

Even in this reduced range, single frequency measurements would be time-consuming and so this range is divided into eight frequency groups (or bands), and the total number of decibels in each group is measured. Each band is one octave wide, the upper frequency of the octave band being twice the lower frequency, and the geometric mean frequency being taken as the octave band's 'label'. This can be shown as: $f_2 = 2f_1$, where f_1 and f_2 are the lower and upper band limits and $(f_0)^2 = (f_1 \times f_2)$ for the centre or mean frequency. International agreements have produced 'preferred octave bands' that have mean frequencies of 63, 125, 250, 500, 1k, 2k, 4k and 8k Hz.

Octave band analysis is useful in gaining a quick guide to the frequency distribution of the noise, but, as Fig. 17.6 demonstrates, much detail is lost, and the lower levels contribute little to the total level in each band (because of the properties of the logarithmic scale). Less detail is lost if $\frac{1}{3}$ -octave bands are used (upper frequency = $2\frac{1}{3} \times$

lower frequency), and $\frac{1}{3}$ -octave band levels combine to give the octave band level. Therefore $\frac{1}{3}$ -octave band levels are always less than the octave band level. Figure 17.7 shows the improved detail given by $\frac{1}{3}$ -octave band analysis.

Instrumentation

A wide variety of sound level meters are available, which have different facilities and levels of accuracy and precision. Clearly the equipment selected should be suitable and sufficient to enable the assessment of interest to be successfully completed.

Early attempts to give the sound level meter a similar frequency response to that of the ear resulted in the weighting networks A, B and C. These were based on the ear's response at 40, 70 and 100 phon. Their relative response is shown in Fig. 17.8. When the A, B and C networks are used, the meter readings are quoted as dB(A), dB(B) and dB(C) respectively. This extra complication was found unhelpful and the B and C weighting networks have fallen from general use. The dB(A) scale has been shown to have certain unexpected advantages when assessing the nuisance value of a noise, and remains in common use (see Table 17.1). Additional weighting networks D and E have been added: D is used for aircraft noise only and E is another attempt at a loudness level measurement.

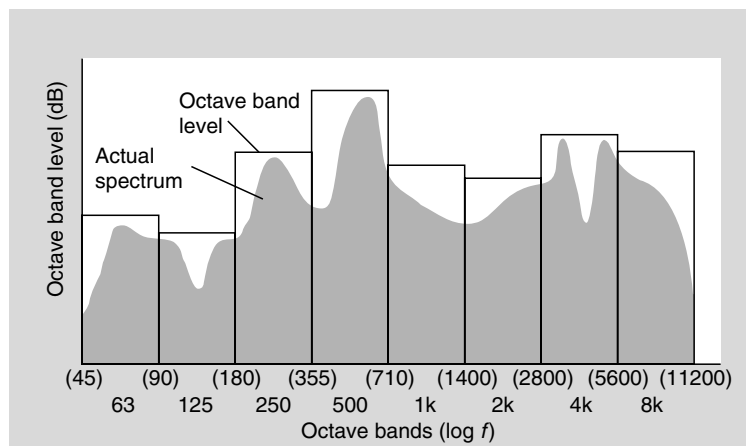


Figure 17.6 Frequency analysis of a noise, showing octave band levels.

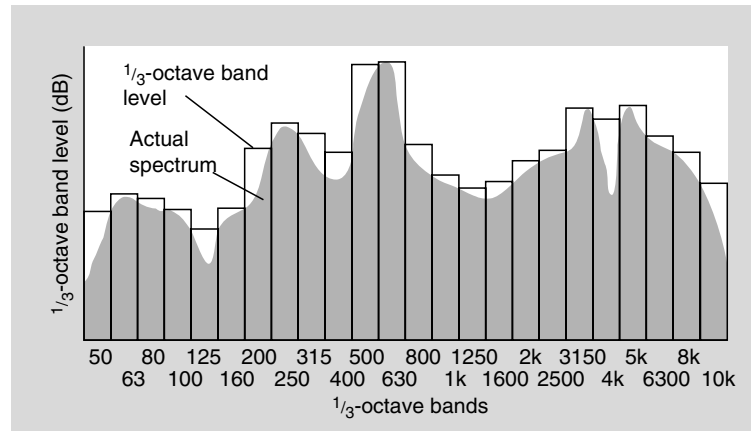


Figure 17.7 Comparison of octave and $\frac{1}{3}$ -octave band analysis.

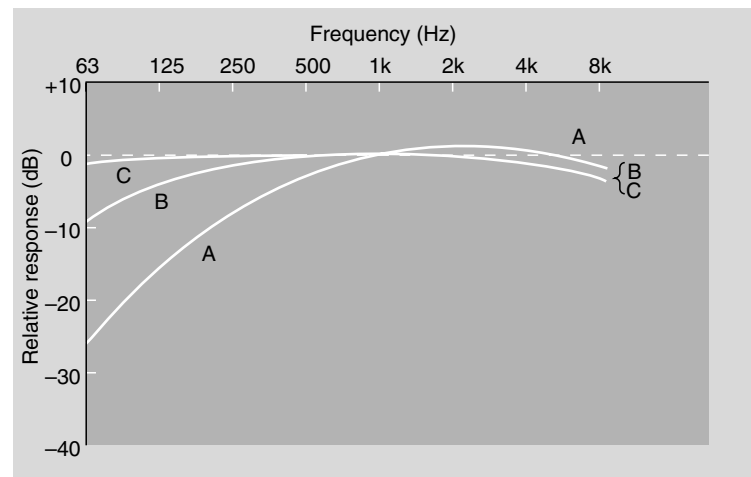


Figure 17.8 Relative responses of the weighting networks.

Table 17.1 Typical levels in dB(A) of a range of common noise.

Type of noise	Typical level in dB(A)
Jet at 30 m	140
Riveting, grinding metals at 1 m	120
Pop group	110
Machine shop with heavy plant	90
City traffic	80
Typing pool	55
Quiet office	40
Countryside at night	20

Estimation of dB(A) level from octave band levels

The corrections given in Table 17.2 are applied to the respective octave band levels, and the total level in dB(A) is found by combining the eight corrected band levels.

Measurement of fluctuating levels

Because noises are rarely steady or regularly fluctuating, methods of assigning numerical values to them have been devised and are being constantly reviewed.

Table 17.2 Octave band corrections for 'A' weighting.

Octave band mid-frequency	63	125	250	500	1k	2k	4k	8k	Hz
Correction	-26	-16	-9	-3	0	+1	+1	-1	dB

Equivalent continuous sound level (L_{eq})

This is the notional steady level that would have emitted the same 'A' weighted sound energy over the same time as the actual noise, i.e.:

$$L_{eq} = 10 \log_{10} \frac{1}{T} \int_0^T \left(\frac{p_A}{2 \times 10^{-5}} \right)^2 dt \quad (17.19)$$

where T is measurement time and p_A is the instantaneous 'A' weighted acoustic pressure in pascals in the undisturbed field in air at atmospheric pressure.

As a doubling of sound energy increases the level by 3 dB, and a 10-fold increase raises the level by 10 dB, noise which is nominally 90 dB(A) for 4 h and 70 dB(A) for 4 h will not produce a L_{eq} of 80 dB(A). In this case:

$$\begin{aligned} \text{Average energy} &= \left[\frac{(4 \times 100\text{-fold}) + (4 \times 1)}{8} \right] \\ &= \frac{404}{8} \end{aligned} \quad (17.20)$$

therefore giving a 50.5-fold increase on an energy content of 70 dB(A), which is equivalent to 87 dB(A). Thus the L_{eq} is more dependent on the higher levels occurring in the measurement period.

When the duration of work and measurement T (see Equation 17.19) are 8 h (28 800 s), in the UK the L_{eq} is called the 'daily personal noise exposure' ($L_{EP,d}$). It is this measure of noise exposure that is required for comparison with the first two action levels of 85 and 90 dB(A). Where more sophisticated equipment is not available, it is possible to calculate an $L_{EP,d}$ by use of the following formulae:

$$f = \frac{t}{8} \text{antilog}[0.1(L-90)] \quad (17.21)$$

and

$$L_{EP,d} = \frac{\log f_{tot}}{0.1} + 90 \text{ dB(A)}$$

where t is the exposure to sound level L (in hours) and f_{tot} is the total value of fractional exposure f over the working day. However, a more simple means is by the use of a nomogram (Fig. 17.9), as is highlighted by the two following examples. When there is only one significant level of noise during the day the value of $L_{EP,d}$ can be obtained from the nomogram in Fig. 17.9. By drawing a straight line connecting the measured level on the L scale with the exposure duration on the t scale, $L_{EP,d}$ can be read at the point of intersection with the centre scale.

Example: a person is exposed to a sound level of 102 dB(A) for $2\frac{1}{4}$ h per day. During the rest of the day the level is below 75 dB(A), which may be ignored. From Fig. 17.9, $L_{EP,d} = 96$ dB(A) (rounded to the next higher decibel).

When periods of exposure at more than one level are significant, the exposure at each level can be converted to a value of 'fractional exposure' (f) using the nomogram in Fig. 17.9. The values of ' f ' received during one day should be added together, and the total value of f converted to $L_{EP,d}$ using the centre scale of the nomogram.

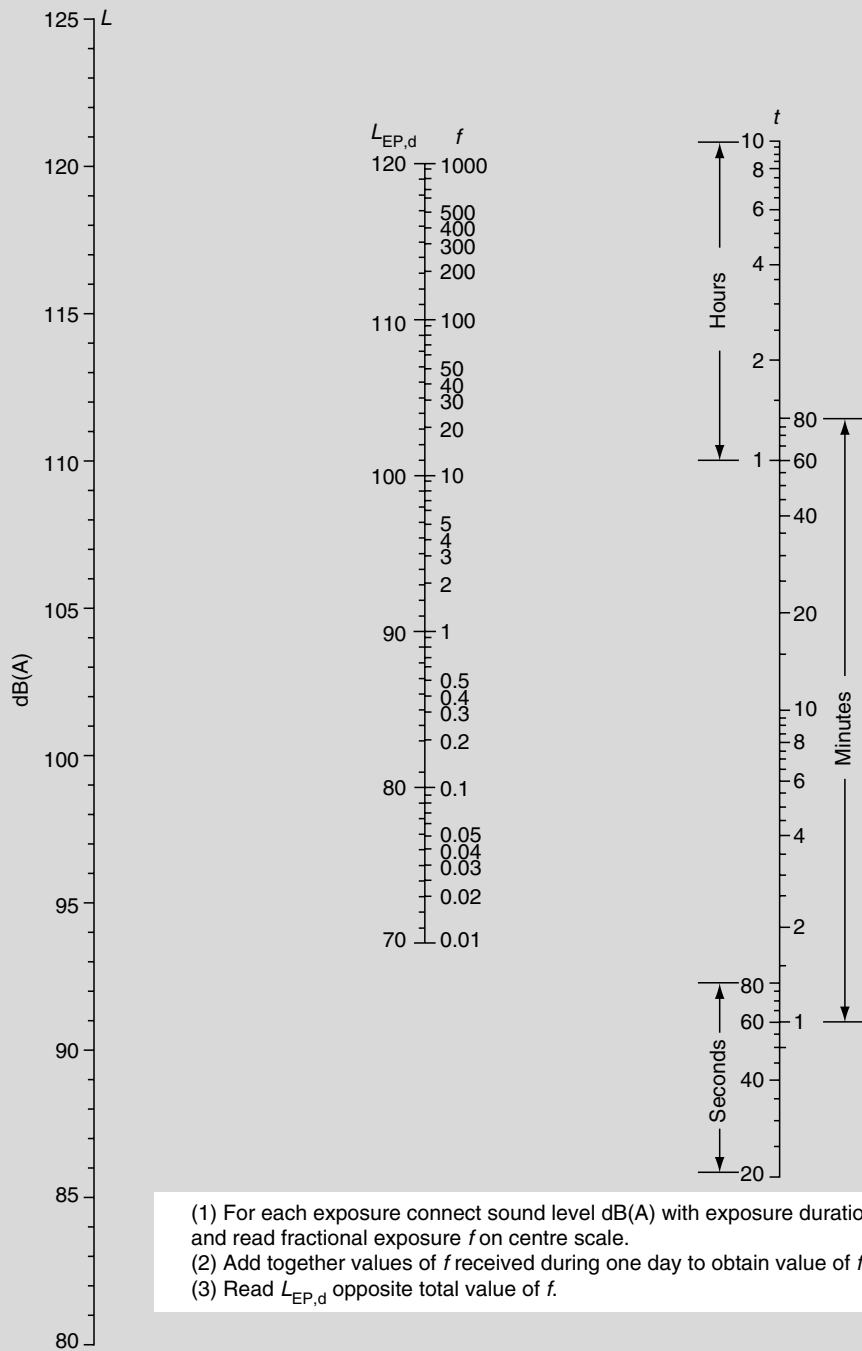
Example: a person is exposed to the pattern of sound in the first two columns of the table below. The third column shows the corresponding values of f which are added together and converted to $L_{EP,d}$.

Sound level dB(A)	Duration of exposure	f (from Fig. 17.9)
114	10 min	5.2
105	45 min	3.0
92	10 h	2.0
Total		10.0

From Fig. 17.9, $L_{EP,d} = 100$ dB(A) (to the nearest decibel).

Weekly average noise exposure

The weekly average of an employee's daily personal noise exposure ($L_{EP,w}$) can be calculated



- (1) For each exposure connect sound level dB(A) with exposure duration t and read fractional exposure f on centre scale.
- (2) Add together values of f received during one day to obtain value of f .
- (3) Read $L_{EP,d}$ opposite total value of f .

Figure 17.9 Nomogram for calculation of $L_{EP,d}$ (from Health and Safety Executive, 1990). Crown copyright is reproduced with the permission of the Controller of HMSO.

from the following formula and is expressed in dB(A):

$$L_{EP,w} = 10 \log_{10} \left[\frac{1}{5} \sum_{k=1}^{k=m} 10^{0.1(L_{EP,d})_k} \right] \quad (17.22)$$

where $(L_{EP,d})_k$ is the $L_{EP,d}$ value for each of m working days of the week.

Single event noise exposure level (L_{AX})

At present this is used for single, short events and is also based on energy. The L_{AX} is the level which, if it lasted for 1 s, would have emitted the same energy as the actual event. In practice, the time of the actual event is taken as the time for which the level is within 10 dB of its maximum.

$$L_{AX} = 10 \log_{10} \int_{t_1}^{t_2} \left(\frac{p(A)}{2 \times 10^{-5}} \right)^2 dt \quad (17.23)$$

where t_1 and t_2 define the time interval (in seconds) in which the level remains within 10 dB of its maximum.

As L_{eq} and L_{AX} use the same concept over different times, they may be related:

$$L_{eq} = 10 \log_{10} \frac{1}{T} \sum_{i=1}^N 10(L_{AXi}/10) \quad (17.24)$$

where N is the number of events in time T .

Statistical levels (L_n)

Statistical levels are used to assess the variation of level with time. Consider the noise 'history' shown in Fig. 17.10. It can be seen that 70 dB(A) was exceeded on two occasions during the 20 s of the measurement: first for $\frac{1}{2}$ s; second for $1\frac{1}{2}$ s, i.e. 70 dB(A) was exceeded for $(\frac{1}{2} + 1\frac{1}{2})/20$ of the time, i.e. 10%. For this noise history, 70 dB(A) is the 'ten per cent level' (L_{10}): the level exceeded for 10% of the time. The L_{10} is a useful measure as it provides a well-defined 'near peak' evaluation of a varying noise level. Other percentages can also be used, such as the L_{90} , which is a well-defined 'near background' level (45 dB(A) in Fig. 17.10).

In 1969, the USA became the first country in the Western world to introduce industrial noise regulations, setting a limit of 90 dB(A), but measured with a simple sound level meter (slow response), for an 8-h daily exposure. In 1971, this limit was incorporated with the Occupational Health and Safety Act, and defined as '100% Noise Dose'. In 1972, the Department of Employment in the UK settled on a limit of 90 dB(A) for an 8-h daily exposure and published the 'Code of Practice for Reducing the Exposure of Employed Persons to Noise'. This has subsequently been updated by the Noise at Work Regulations 1989, in which the three action levels of 85 and 90 dB(A) (as $L_{EP,d}$ values) and 200 pascals (140 dB or 20 μ Pa) (as a peak sound pressure) have been set. This is set

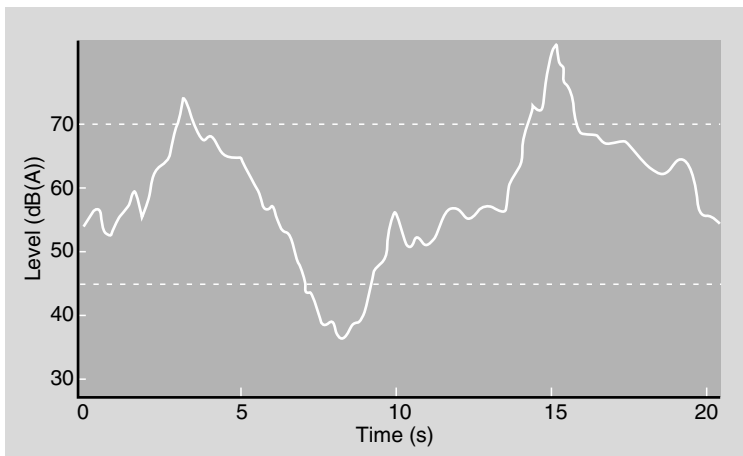


Figure 17.10 History of a fluctuating noise.

to change again wherein the two action levels are both reduced by 5 dB(A).

The British approach to the problem of noise fluctuation was to employ the equal energy concept of the L_{eq} of either 85 or 90 dB(A) for 8 h per day, subject to a maximum level of 140 dB (fast response) for unprotected ears. By contrast, the Occupational Safety and Health Act (OSHA) of the USA allows an increase of 5 dB for a halving exposure duration up to a maximum of 115 dB(A). This relationship makes an allowance for the recovery of temporary threshold shift (TTS) during the periods of less intense noise when exposure time has been reduced.

Noise dose may be defined for both countries as:

$$\text{Noise dose} = 100 \int_0^{T/8} \left(\frac{p(t)}{0.632} \right)^n dt\% \quad (17.25)$$

where $p(t)$ is the A-weighted varying sound pressure (N m^{-2}), 0.632 N m^{-2} corresponds to 90 dB(A), T is the measurement duration (in hours) and n takes the value 2 in the UK (and most of Europe) and 1.2 in the USA. Comparison of 100% dose for different exposures is given in Table 17.3.

Noise dosimeter (dosimeter)

If workers experience many different levels during their shifts, their noise dose calculations can only be achieved accurately by means of an instrument

Table 17.3 Comparison of duration and levels in the UK and USA for 100% dose at 90 dB(A) over 8 h.

Exposure permitted (h day ⁻¹)	UK L_{eq} (dB(A))	OSHA dB(A) (slow)
8	90	90
4	93	95
2	96	100
1	99	105
$\frac{1}{2}$	102	110
$\frac{1}{4}$	105	115
$\frac{1}{8}$	108	115
$\frac{1}{16}$	111	115 max
$\frac{1}{32}$	114	115
$\frac{1}{64}$	117	115

capable of measuring the L_{eq} over the whole shift. Personal noise dosimeters are convenient as they fit into the worker's pocket, and the read-out (depending on type) is directly in percentage dose, L_{eq} max peak, $\text{Pa}^2 \text{ h}$, etc.

Guidance for the UK Noise at Work Regulations recommends that measurement should be made in the 'undisturbed field', however, results are unlikely to be significantly affected by reflections if the microphone is kept at least 4 cm away from the operator and most dosimeter microphones are provided with a clip to hold them onto the brim of a safety helmet or overall lappel (Fig. 17.11). The microphone should also be placed on the side of the subject likely to receive most noise. Thus the microphone receives the same sound pressure as the worker's ear, which the dosimeter 'A weights'

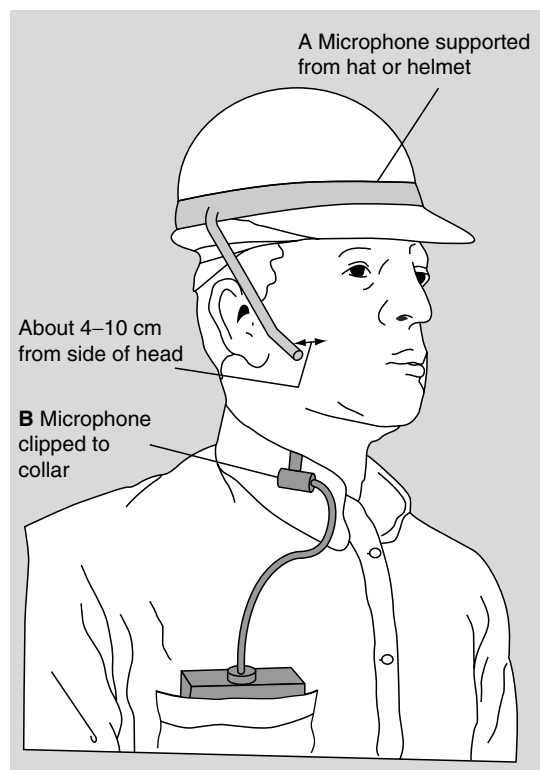


Figure 17.11 Location of microphone for dosimeters. A, head-mounted microphone; B, collar- or shoulder-mounted microphone (from Health and Safety Executive, 1990). Crown copyright is reproduced with the permission of the Controller of HMSO.

and then, after squaring (or in the USA raising to the power of 1.2), totals over the measurement period, and displays as the noise dose.

The advantage in using the concept of noise dose is that the 'usual rules' of arithmetic apply, the maximum permitted in one shift being always 100%. For example, if a worker receives 60% of the dose in the first 2 h, he/she may only receive 40% during the remaining 6 h. He/she must therefore be moved to a location with a lower L_{eq} . This may be calculated as follows:

$$\begin{aligned} 40\% \text{ in } 6 \text{ h} &= \frac{40}{6} \% \text{ per h} \\ &= \left(\frac{40}{6} \times 8\right) \% \text{ in } 8 \text{ h} \quad (17.26) \\ &= 53\frac{1}{3} \simeq 50\% \end{aligned}$$

Thus, the new L_{eq} (to give a noise dose of 50% in 8 h):

$$\begin{aligned} &= 87 \text{ dB(A) in the UK} \\ &\quad (85 \text{ dB(A) in the USA}) \end{aligned}$$

Such a relocation of a worker would call upon good industrial relations within the organization.

Aural comfort

Various criteria have evolved to provide guidance on acceptable maximum background noise levels in different situations. In general, the quieter the activity, the lower the acceptable background level. Low-frequency sounds are less well heard and so their acceptable levels are greater than higher frequency sounds, which the ear detects readily.

Noise criteria curves

In the 1950s, research in the USA determined the maximum levels in the eight octave bands (63 Hz–8 kHz) that caused minimal interference with two women conversing on the telephone. As a result, curves numbered from 15 to 70 (the octave band sound level at 1 kHz) were produced for use in the office environment. These were known as noise criteria (NC) curves (Fig. 17.12). The background noise in an environment is measured in decibels

(linear) and the intensities at each octave band plotted on the graph. The NC rating for that environment is taken as the number of the curve above the highest value.

A number of different environments have been assigned NC values and these are given below. For example, the NC for a typing pool is 45 and therefore background noise levels measured in the office of interest in excess of this value are too noisy.

Noise rating curves

This method uses the same concept as the NC system, but the curves are based on the results of a large-scale survey of the reaction of the community to noise and as a result have a much wider range (0–135) than the NC curves. The rating of rooms in dwellings involves the corrections shown in the table of Fig. 17.13. For example, a living room (noise rating (NR) 30) in a residential urban area (+5) suffering an impulsive noise (–5) for 6% of the time (+10) would be assessed at NR (30 + 5 – 5 + 10) = NR(40), i.e. the background noise may rise to NR 40 when the impulsive noise is present without conditions becoming unacceptable. Existing background noises may be assigned NR values as with NC curves.

The main advantage of using criteria which utilize data from octave band analysis, such as these two (NR and NC), is that by comparison with these curves the frequencies of most concern are easily identified. A dB(A) value, on the other hand, has incorporated this frequency.

Noise and materials

Materials reflect, absorb and transmit sound, the proportions depending on the material and the frequency of the sound (Fig. 17.14).

The *absorption coefficient* (α) is defined as:

$$\frac{\text{Intensity of sound reflected by material}}{\text{Intensity of sound incident on same area of material}}$$

The *reflection coefficient* (r) is defined as:

$$\frac{\text{Intensity of sound absorbed by material}}{\text{Intensity of sound incident on same area of material}}$$

The *transmission coefficient* (τ) is defined as:

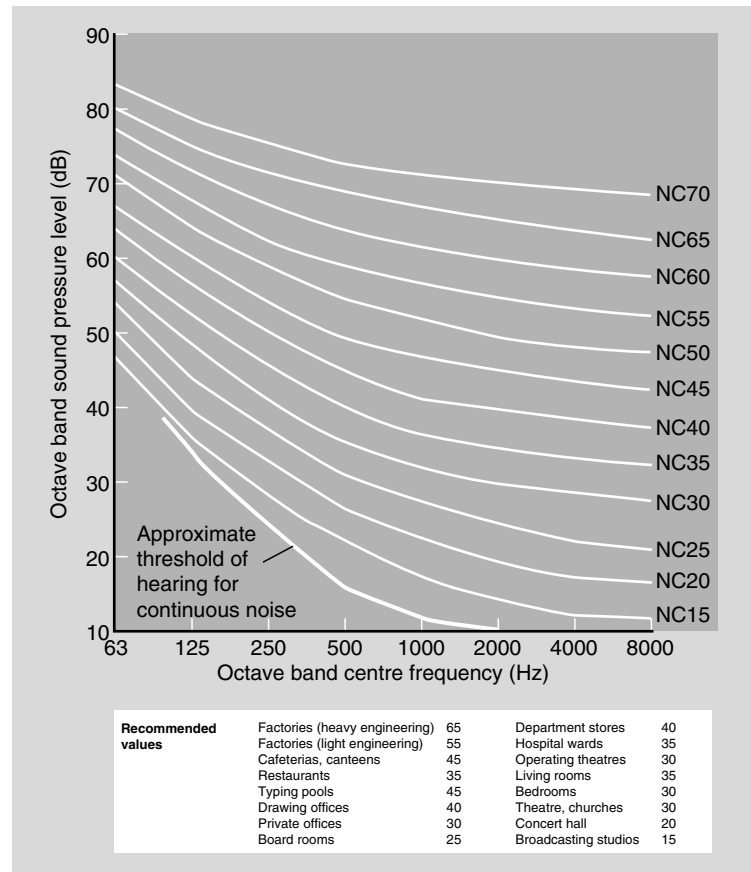


Figure 17.12 Noise criteria (NC) curves and recommended values showing NC 30 for a private office.

$$\frac{\text{Intensity of sound transmitted by material}}{\text{Intensity of sound incident on same area of material}}$$

It follows that $r + \alpha + \tau = 1$.

For many materials in practical use τ is very much smaller than α or r , and so in some situations it is convenient to take $\alpha + r = 1$. Such a situation is met in the consideration of the growth and decay of sound in an enclosure where the very small amount transmitted through the walls has an insignificant effect.

Absorption coefficient

Absorption is really the conversion of sound into other forms of energy. Materials that absorb sound may be classified into three categories.

1 *Porous materials*: sound enters the pores and the viscous forces so generated lead to heat production.

2 *Non-porous panels*: sound energy is converted into vibrational (i.e. mechanical) energy.

3 *Perforated materials*: the perforations act as cavity resonators, absorbing narrow bands of frequency.

An idealized summary of the behaviour of these materials is given in Fig. 17.15

Although in theory α can never exceed unity, standard measurement methods can arrive at values in excess of 1. These procedures calculated α on a projected surface area, and should the material be made up into a non-plane shape its active area will be greater than that used in the calculation, giving α a value greater than unity.

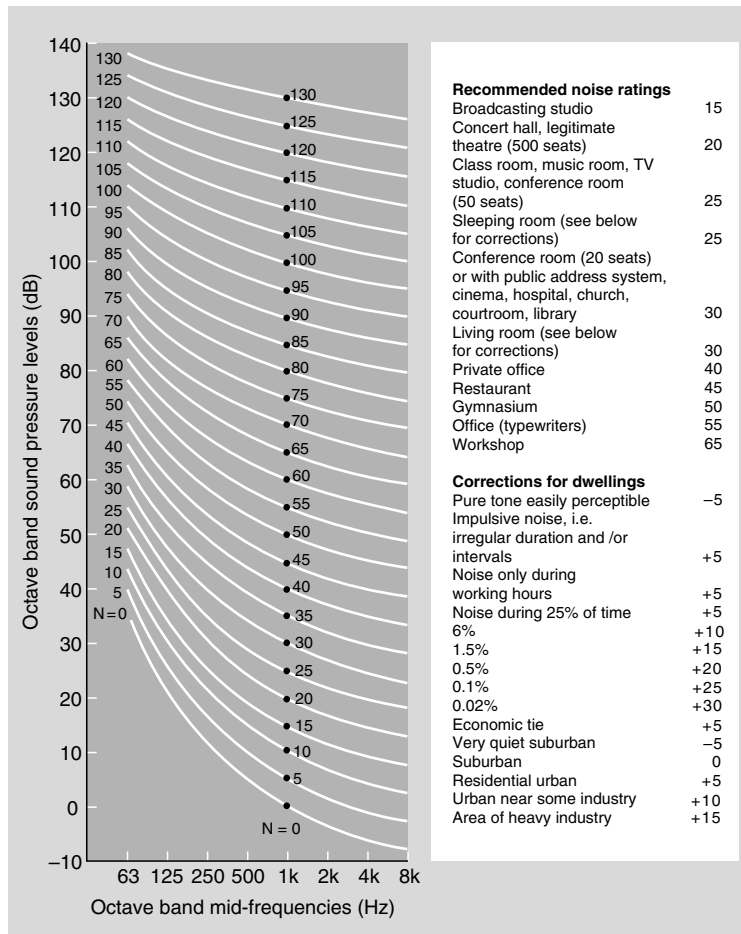


Figure 17.13 Noise rating curves and recommended values.

Transmission coefficient

The transmission coefficient depends on the material's density and thickness and the frequency of the sound being transmitted. Mention has already been made of the small value of τ usually encountered ($\tau \approx 0.0001$ at 500 Hz for a brick wall). It is more convenient to express transmission quantities in terms of a transmission loss (TL), defined as:

$$TL = 10 \log_{10} \frac{1}{\tau} \text{ dB} \quad (17.27)$$

So, for the brick wall:

$$TL = 10 \log_{10} \frac{1}{0.0001} = 40 \text{ dB}$$

Sound insulation

Figure 17.16 shows the general behaviour of partitions used in buildings. For the lowest frequencies, the behaviour of the partition depends mainly on its edge fixing and stiffness; the TL falls as the frequency increases. For the highest frequencies, the behaviour depends on wavelengths, there being certain frequencies where the wavelength in air corresponds exactly with the wavelength of the bending wave set up in the flexing wall. When such coincidences occur there is little energy lost and so the TL is low. Above and below these critical frequencies the TL remains high.

In the intermediate frequency range, the mass per unit area of the partition is the controlling

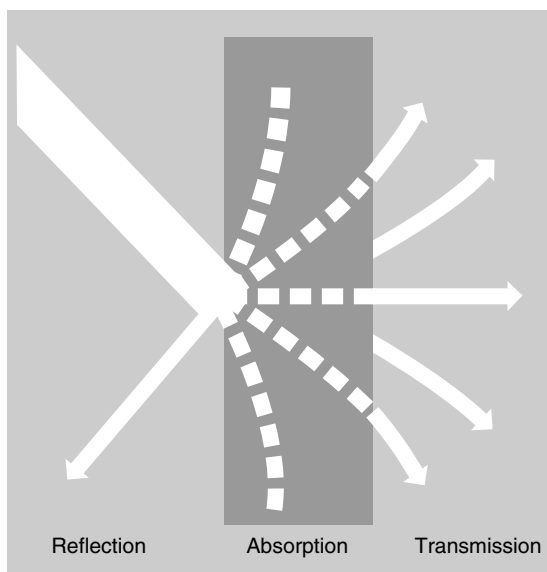


Figure 17.14 Reflection, absorption and transmission.

quantity, the behaviour of the partition obeying the Mass law as a first approximation. This states:

$$TL = 20 \log_{10} mf - 43 \text{ dB} \quad (17.28)$$

where m is the mass per unit area of wall (= density \times thickness) and f is the frequency of the sound. Actual partitions vary greatly from this general picture. It is imperative that measurements made on 'real' partitions should cover the required frequency range (usually 100–3150 Hz) in some detail (at least $\frac{1}{3}$ -octave bands). Trade literature often quotes single, averaged, figures for the TL of a product. This should be used only as a guide because of the large variations occurring over the frequency range.

Composite partitions

Where more than one material is used in a wall, the average transmission loss for the composite partition may be found by calculating the area-weighted average transmission coefficient (τ_{av}):

$$\tau_{av} = \frac{\sum_{i=1}^N \tau_i S_i}{\sum_{i=1}^N S_i} \quad (17.29)$$

where S_i is the area of the i th material in the partition.

Thence:

$$TL_{av} = 10 \log_{10} \frac{1}{\tau_{av}} \quad (17.30)$$

This method gives a good estimate when the sound fields on each side of the partition are diffuse. Air gaps are classed as a material making up the partition having a TL equal to 0 dB (i.e. $\tau = 1$). Their effect on TL_{av} is marked, a small air gap reducing the insulation value of a wall considerably.

Sound in enclosed spaces

The total sound field produced in an enclosed space has two components (Fig. 17.17).

- 1 The direct sound field: which travels from source to listener by the shortest route without encountering any room surface.
- 2 The reverberant sound field: which reaches the listener after at least one reflection from a room surface.

The size of the direct field depends on the acoustic power of the source, the distance between the source and the listener and the position of the source in the space (which affects the directionality of the source). The size of the reverberant compon-

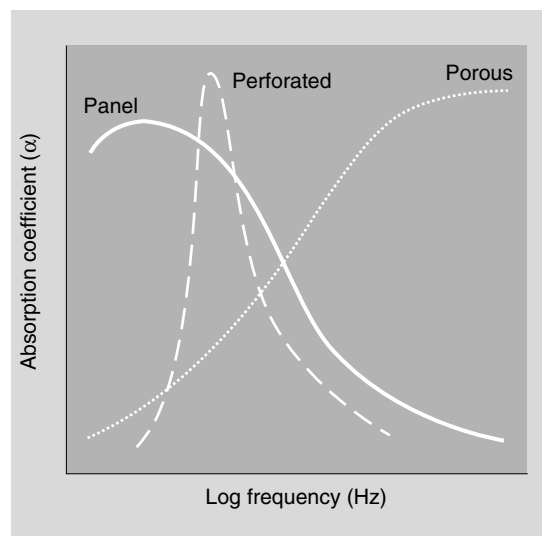


Figure 17.15 Absorption coefficient variation with frequency.

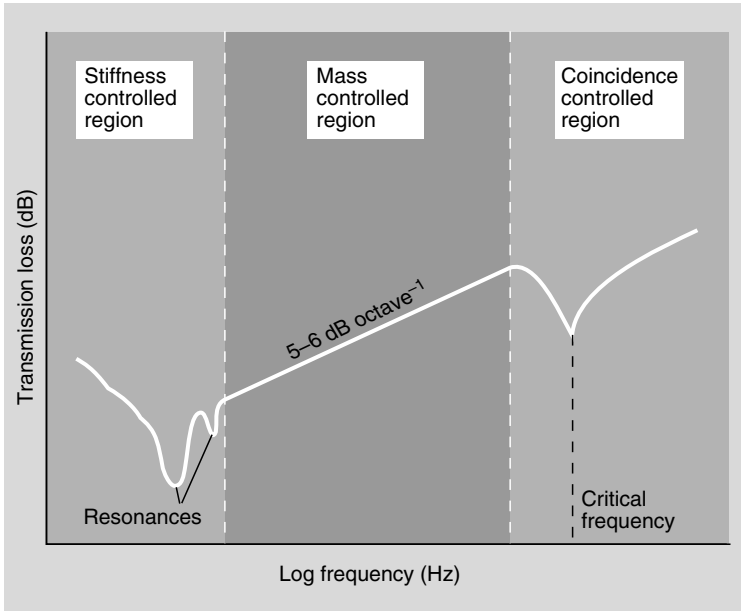


Figure 17.16 Generalized variation of transmission loss of a typical building partition with frequency.

ent depends on the amount of sound reflected at each reflecting surface and the number of reflections that each individual sound wave undergoes before reaching the listener. This is found to depend on the area-weighted average absorption coefficient ($\bar{\alpha}$) and the room's surface area (S):

$$\bar{\alpha} = \frac{\sum_{i=1}^N S_i \alpha_i}{\sum_{i=1}^N S_i} \quad (17.31)$$

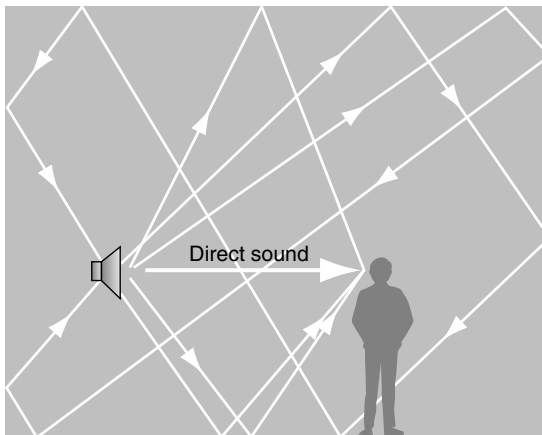


Figure 17.17 Distribution of sound in an enclosed space.

where S_i is the surface area of the i th material and α_i is the respective absorption coefficient.

The reverberant component does not vary greatly over a given room, whereas the direct sound component for a point source varies inversely with the square of the distance from the source. Therefore, near the source, where the direct component dominates, the total sound field falls rapidly as the distance increases, becoming constant in the far field where the field is predominantly reverberant (Fig. 17.18).

It may be shown that the sound pressure level (SPL) at a point in a room is related to the acoustic power of the source by the following expression:

$$\text{SPL} = L_w + 10 \log_{10} \left(\frac{Q}{4\pi d^2} + \frac{4(1-\bar{\alpha})}{S\bar{\alpha}} \right) \quad (17.32)$$

where L_w is acoustic power (W), Q is a directivity factor, d is the distance from the source (m), S is the total surface area of the room and $\bar{\alpha}$ is an area-weighted average absorption coefficient. Readings of sound pressure levels taken near the source greatly depend on the position of the microphone. It is good practice to make these measurements at least 1 m from the source, where the variation in level is not so large.

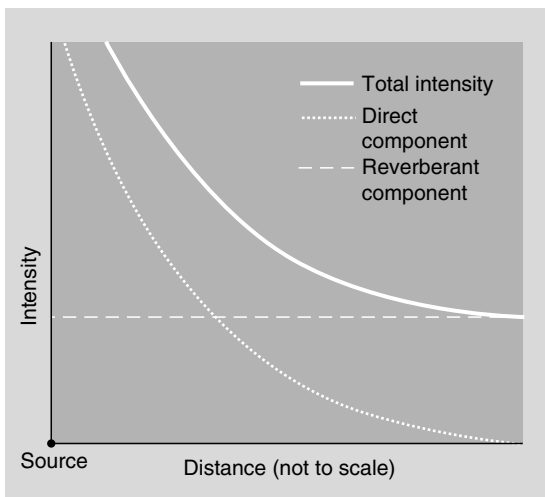


Figure 17.18 Contribution of direct and reverberant sound to total field.

Reverberation time

As the velocity of sound is finite, reverberation components arrive at different times, following the direct component. Thus, the total sound field does not start and stop but grows and decays. The time taken for a total sound decay of 60 dB is known as the reverberation time, and is a useful guide to the acoustic quality of a room (Fig. 17.19).

Work by Sabine has shown that the reverberation time (RT) of a room of moderate absorption is given by:

$$RT = \frac{0.16V}{A} \text{ seconds} \quad (17.33)$$

and

$$A = \sum_{i=1}^N S_i \alpha_i$$

where V is the room volume (m^3). As the absorption coefficients vary with frequency, the reverberation time also varies across the frequency range.

Sabine's formula may be used to estimate the reverberation times of a planned room and to calculate the change in absorption necessary to comply with the recommended optimal values. The optimum reverberation time depends on the volume of the room and the activity taking place. Such values are based on subjective judgements and are available in chart or tabular form.

Doubling the absorption in a room will halve its reverberation time, and the reverberant component of the total field. Therefore the noise level in a room will fall by about 3 dB in the far field if the room absorption is doubled.

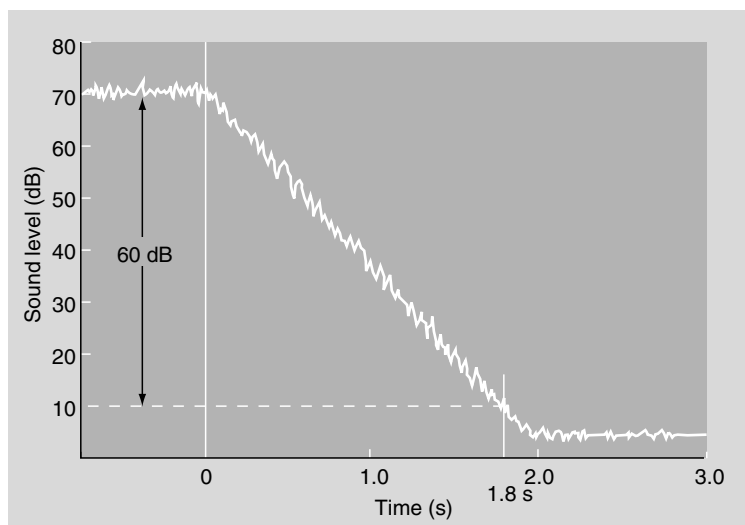


Figure 17.19 Reverberation time (from Waldron, 1989).

Measurement of hearing

The most commonly used assessment of hearing is the determination of the threshold of audibility. This is the level of sound required to be just audible. It is not absolutely fixed for the individual but seems to vary over a range of 2–6 dB from day to day, and from determination to determination. In practice it is necessary to take the threshold as the level which is just heard 50% of the times for which it is presented.

In addition, further variations occur if the test sounds are presented from above or below the threshold range. The ‘descending threshold’ is found by presenting the higher levels first; the ‘ascending threshold’ is reached by beginning with levels below the threshold range and increasing them until the sounds are just heard. Because of these variations, 5-dB steps of level are used in practical audiometry, the likely variation being smaller than the step.

Standard hearing

In 1952, two groups of workers (at the National Physical Laboratory and the Central Medical Establishment, RAF) tested a total of 1200 subjects, aged from 18 to 23 years. The subjects were screened for good general health and no otological history. Although the techniques employed in the threshold tests were slightly different, comparable results were obtained and became adopted in the 1954 British Standard. Two interesting facts emerged. First, the spread of hearing threshold for these young, carefully screened volunteers was quite large, the standard deviation being at least 6 dB at all audiometric frequencies. Second, comparison with the US Standard (based on small samples) showed that the US Standard under estimated hearing acuity by some 10 dB.

Further investigations in other countries confirmed the British findings and in 1964 the ISO published ‘Standard Reference Zeros for Pure Tone Audiometers’ (ISO 389). In this Standard, standard hearing is assigned 0 dB hearing level (HL), and is referred to as the level generated by a headphone of agreed construction into an artificial ear (a 6-cm³ cavity). This level has different

values across the frequency range, but standard hearing is 0 dBHL at every frequency.

The pure tone audiometer

Pure tone signals are generated (usually 125, 250, 500, 1k, 2k, 3k, 4k, 6k and 8k Hz) which, after amplification, are passed to the headphone. With the hearing level control at 0 dBHL, the level of the signal is automatically adjusted at each frequency to be that required by the ISO standard reference zero. When the hearing level control is set to 50 dBHL, for example, 50 dB more than the reference zero emerges from the headphones at each frequency. Thus, the audiometer makes no attempt to present signals of equal loudness to the subject, as equal loudness increments are not identical at different frequencies.

The subject to be tested is comfortably seated and ears checked for absence of wax. The headphones are fitted, ensuring that the earphone orifice coincides with the ear canal. The subject is asked to indicate, by raising a finger or operating a signal lamp, when sounds are heard (no matter how faintly). Verbal communication should be discouraged once testing has begun. A general assessment of hearing will have already been made from casual conversation before the test, and a sufficiently high signal at 1 kHz for about 1 s is presented so that a confident response is made. This level is reduced in steps until the subject no longer responds. The threshold is then crossed by raising the signal until a response is made. Threshold is taken when the same level is just heard on three occasions.

The other frequencies are then tested and 1 kHz repeated as a check. If there is significant difference, the subject should be retested. The other ear is then tested. Results are conveniently recorded on an audiogram (Fig. 17.20). An experienced audiometrician will complete the test in about 15 min.

The demand for large-scale screening and monitoring by industry has greatly increased the use of the automatic audiometer. In this instrument the frequency is changed, and the hearing level control motor driven, automatically. The hearing level value is traced by a pen onto an audiogram. The subject is instructed to press a switch when they

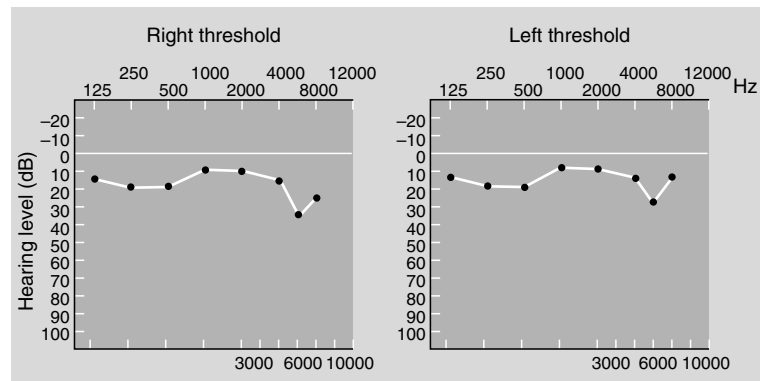


Figure 17.20 Pure tone audiogram.

can hear the signal (which steadily reduces the signal) and to release the switch when the signal is no longer audible (which steadily increases the signal). In this way the threshold crossings are drawn out (Fig. 17.21). Simultaneous supervision of up to three machines is possible.

Test conditions

Careful listening is essential for these tests and so the subject must be placed in a very quiet environment. The headphones give some attenuation at middle and high frequencies, which can be increased by the provision of hemispherical shells

fitting over each earphone. Better protection is given by purpose-built test booths (or carefully designed rooms). Such quiet conditions are essential as many of those screened will have normal hearing. In order to test to -10 dBHL with an error not exceeding 2 dB, the background level at the ear should not exceed the octave band levels given in Table 17.4.

Limitations of audiometry

The primary objective of legislation specific to noise is that it should prevent or reduce the incidence of noise-induced hearing loss (NIHL) rather

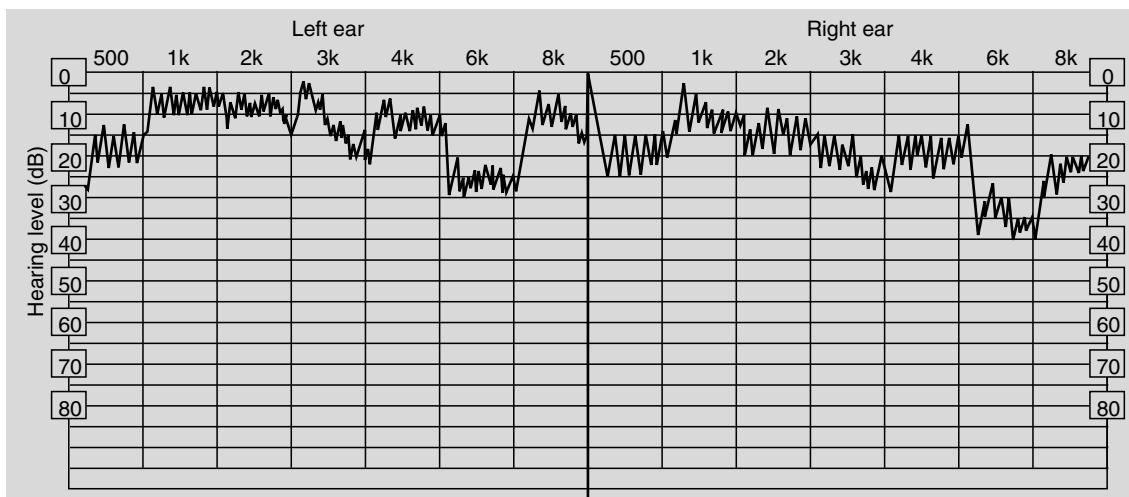


Figure 17.21 Audiogram from an automatic (self-recording) audiometer. The threshold is taken as the mean of the peaks' average and valleys' average.

Table 17.4 Maximum allowable noise levels for audiometry: (a) at ear, for measurements down to -10 dBHL with error not exceeding 2 dB (no headphones); (b) in booth if MX41 / AR cushions are used with audiometer headphones.

(a)									
Octave band mid- frequency	63	125	250	500	1k	2k	4k	8k	Hz
Maximum noise at ear	61	46	31	7	1	4	6	9	dB
(b)									
Octave band mid- frequency	63	125	250	500	1k	2k	4k	8k	Hz
Maximum noise in booth	62	48	36	14	16	29	37	32	dB

than the simple reduction of noise exposure *per se*. It is therefore implicit that employers should assess the hearing levels of their workforce in order to identify those that require additional protection. Audiometry should be able to accurately measure the hearing of an industrial population from 16 to at least 65 years old and to be able to reliably detect small changes over time.

The 'audiometric zero' is set at the mode hearing level of young, non-noise-exposed, otologically normal people at each frequency (see earlier section 'standard hearing'). It therefore follows that a number of the younger population will have hearing levels markedly better than audiometric zero. Individuals with the most sensitive hearing (in the top 5% of the population) will have hearing levels ranging between -8 dB at 500 Hz and 1 kHz to -14 dB at 8 kHz (manual audiometer), with the data about 3 dB better if conducted on an automatic self-recording audiometer. In addition, the audiogram will range by about 5 dB above and below the mean level. As it is desirable for the audiometer to present inaudible sounds to each individual, it should range down to -20 dB at 500 Hz and 1 kHz to -25 dB at 8 kHz.

In the main, audiometry only measures hearing levels down to -5 dB (as it is assumed that these are 'normal people'); however, if an individual with a true hearing level of -15 dB was tested it would simply be recorded as -5 dB. If that individual was exposed to noise, conventional audiometry would only facilitate the identification of hearing loss when it had degraded to worse than 0 dB. This would appear as a degradation of 5 dB but in reality would be ≥ 15 dB, thereby failing to identify a susceptible individual. Improvements in audiometers would also necessitate that the facil-

ities in which these tests are carried out would have to be at least 10 dB quieter than is currently recommended.

It has been reported that the 10% of the population most susceptible to NIHL, when exposed to $L_{EP,d}$ of 90 dB(A), could be expected to exhibit hearing level changes at 4 kHz of 13 dB over a 5-year period. It would therefore not be unreasonable if audiometry were to be able to detect a change in hearing of one-half of this value – a suggested objective is to be reliably able to detect a 5-dB change. However, to be able to detect a 5-dB change over time the resolution of each audiogram must be to less than 2.5 dB. Conventional audiometry is only able to resolve to '5 dB at best' (Department of Health and Social Security, 1982).

The accuracy of audiometry can be affected by three main factors: technical limitations; the 'learning' effect; and fit of the headphones. There are two characteristics of self-recording audiometers that could affect resolution, these being step width and the calculation of the mean hearing levels at each frequency. Step widths are usually 2.5 dB, but to achieve this with any resolution step widths should not exceed 1–1.5 dB and many audiometers have computers that indicate mean hearing level at each frequency but round-off to the nearest 2.5 dB!

The learning effect is when the examinee becomes more proficient over the period of the test, with the result that the ear first tested is worse than the second. Some audiometers retest the first ear and significant differences between the first and second test of the same ear are indicative of this effect. The magnitude of this effect could be as much as 2.5 dB.

Audiometry is known not to be very repeatable (the very criteria you need and want) (Department of Health and Social Security, 1982) and this may be partly due to headphone location. It has been recommended that at least two tests are conducted (without removal of the headphones) until the difference between tests is ≤ 2 dB.

Further tests

Headphone stimulation tests the complete auditory pathway, from canal to brain, and malfunction in any part of the pathway will give an elevated hearing level indicating a hearing loss. Such losses are classified into:

- 1 conductive losses, due to malfunction in the outer and middle ear; and
- 2 sensorineural losses, due to malfunction in the inner ear and auditory nerve.

To isolate the loss caused by sensorineural deafness, the cochlea is stimulated by vibrating the skull with a small electromagnetic vibrator applied to the mastoid area. The pure tone audiometer is calibrated so that the thresholds obtained for a normally hearing subject are numerically equal for both headphone and vibrator tests. A difference between hearing levels for headphone (air conduction) and vibrator (bone conduction) tests indicates a conductive loss.

One important complication is that the interaural attenuation by vibration stimulation is almost 9 dB – stimulating one ear also stimulates the other. The non-test ear must be sufficiently occupied with masking noise from a headphone so that the pure tone threshold determination of the test ear may be accomplished. Interaural attenuation for air-conducted sound is around 40 dB and so the masking noise does not interfere with the test ear unless it is too great. The correct amount to be used may be found by increasing the masking level in 10 dB steps until the test ear threshold is identical for three successive determinations. This need for masking has delayed the introduction of an automatic bone-conduction audiometer.

Pure tone manual and automatic audiometric tests are subjective in that they rely on the cooperation of the subject. If the subject is unable or

unwilling to cooperate, objective tests are now available in some of the larger hospitals. One such method detects the electrical activity of the cortex (using small surface electrodes) at levels very near threshold even when the subject is asleep. This method has been used where there have been disputes over the amount of hearing present in a subject.

Noise exposure and health

The effects of excessive exposure to noise are discussed in detail in Chapter 7.

Noise immission level

A UK government-sponsored investigation using 1000 subjects has suggested that the amounts of threshold shift are related to the total noise exposure (dependent on the noise level and its duration). The actual shift in threshold, corrected for natural loss of acuity with age, was found to correlate well with the noise immission level (NIL), defined as:

$$\text{NIL} = L_A + 10 \log_{10} t \quad (17.33)$$

where L_A is the level of noise in dB(A) (L_{eq} if noise fluctuates) and t is the number of years' duration.

Although industrial noises have many different spectra, the UK investigation found this to have no significant effect on the threshold shifts of those tested. The precise amount of shift varied considerably from subject to subject. The likely effect of various NILs on hearing is shown in Fig. 17.22. It will be noted that the maximum threshold shift occurs at 4 kHz, which is a characteristic of this NIHL (although some people exhibit their maximum shifts at 6 kHz). In the early stages of exposure the threshold shift diminishes after a few hours' rest. Increased exposure leads to increased shifts of which only part is recoverable with rest. The amount remaining after 40 h is termed a 'permanent threshold shift' (PTS), the recovered shift being termed a 'temporary threshold shift' (TTS). TTS recovery exhibits a 'bounce effect', shown in Fig. 17.23, being rapidly reduced in the first minute, and increasing in the second minute, of rest. Subsequent recovery tends to follow a logarithmic

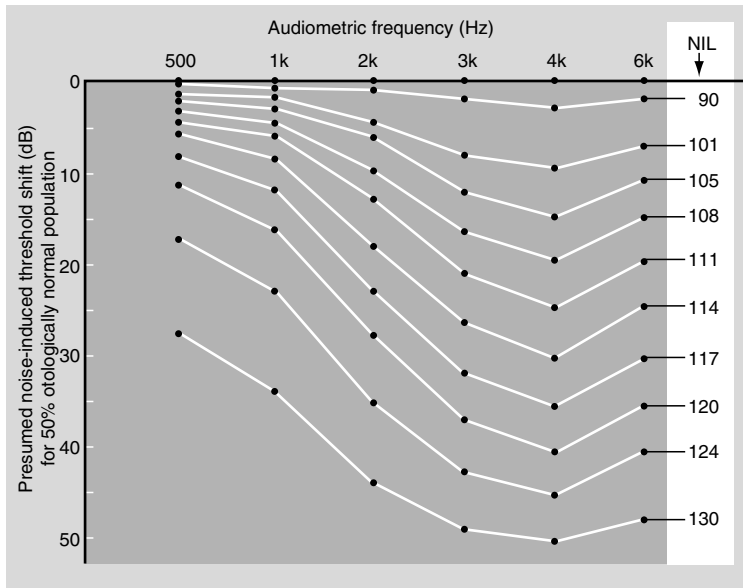


Figure 17.22 Likely effects of noise immersion levels (NILs) on hearing.

relationship with time. For consistency, TTSs are measured after 2 min of rest have elapsed (TTS_2).

Over many years of noise exposure the total loss increases and a greater proportion of it becomes permanent. Standard agreed amounts of presbycusis (see Chapter 7), based on large-scale tests, can be deducted from audiometric test results to estimate the amount of NIHL present.

Hearing conservation

The greatest noise levels received by many in the course of their occupation are the levels encountered while travelling to and from work, rather than at work. For many people, the amount of

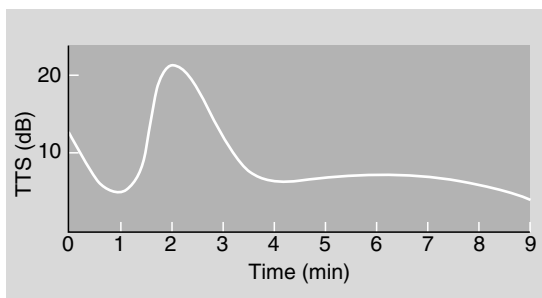


Figure 17.23 The 'bounce effect'.

NIHL accrued will depend on their leisure activities. However, very high noise levels are associated with many industries, and if the hearing of those employed there is to be conserved, exposure to such noise must be controlled. It is necessary to decide upon a level of exposure that affords sufficient reduction for hearing conservation and, at the same time, is realistic. Too low a level at work does not prevent loss due to leisure activities, etc., and would cause grave engineering and economic problems.

NIHL affects those frequencies that are required for good speech reception (the information content of speech depending on the consonants which are high frequency). Small losses at these frequencies affect speech intelligibility insignificantly, and so a limited amount of NIHL at the end of a working life may be tolerated. If these amounts are restricted to < 15 dB at 4000 Hz, between 10 and 15 dB at 3000 Hz and < 10 dB at 500, 1000 and 2000 Hz then it has been found that a NIL over 50 years of 105 dB will just meet this criterion for the majority of the working population. For a few individuals the losses will be greater with this exposure.

It is unlikely that the same conditions will occur for 50 years, and a NIL of 104 dB over 40 years

will more adequately meet this criterion for most of the working population. A 40-year NIL of 104 dB implies an L_{eq} of 88 dBA, at home and at work. If levels at work are controlled to be < 88 dBA, then the conversation criterion will not be exceeded there. If a duration of 30 years is taken, a NIL of 105 is obtained with an L_{eq} of 90 dB(A). Such an increase has an insignificant effect on hearing conservation but a dramatic effect on noise control problems (a 2 dB change implies a sound intensity increase of 60%).

Control of noise exposure levels

As with the control of any overexposure in the work environment, one must attempt to prevent rather than control exposure, and if prevention or elimination is not possible then to descend down the hierarchy of acceptable/effective control measures (see Chapter 29). Control of noise exposure provides the classic example of viewing the work environment in three distinct sections: source; transmission path; and receiver (Fig. 17.24).

Noise reduction at source

As movement causes vibration that is passed on to the air particles and perceived as sound, minimization of movement in any process will achieve a measure of noise control at source. A number of methods of preventing noise generation are given below.

1 Substitution of a quieter process, i.e. welding not riveting.

- 2 Avoiding or cushioning of impacts.
- 3 Introduce or increase the amount of damping.
- 4 Reduction of turbulence of air exhausts and jets by silencers, either of the ‘absorption’ type with which the attenuation (insertion loss) is achieved by a lining of absorbent material, or the ‘expansion chamber’ type with which the insertion loss is achieved by acoustic mismatch between the volume of the chamber and inlet/outlet pipe (a number are now a hybrid of these two types).
- 5 Introduction of low-noise air nozzles and pneumatic ejectors.
- 6 Matching the pressure of the supplied air to the needs of the air-powered equipment.
- 7 Avoiding ‘chapping’ airstreams by rotating components.
- 8 Improved design of fans, fan casings, compressors, etc.
- 9 Dynamic balancing of rotating parts.
- 10 The use of better quality control in design and manufacturing procedures to obviate the need for *post hoc* rectification.
- 11 Better machine maintenance.
- 12 Limit the duration for which a noisy machine or part of a machine is used.

Control of the transmission path

Having made every effort to control the noise exposure at the source, the next most appropriate course of action is to minimize the progress of the energy from the source to the receiver. A number of examples are given below.

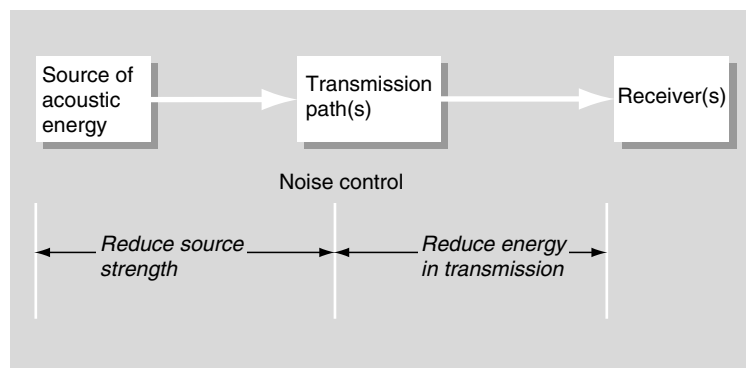


Figure 17.24 Energy flow diagram.

- 1 use of correctly chosen reflecting and absorbent barriers for the direct component;
- 2 use of correctly chosen absorbent material on surrounding surfaces to minimize the reflected component;
- 3 use of anti-vibration mountings under machines.
- 4 enclosure of the source;
- 5 provision of a noise refuge;
- 6 increasing the distance between source and receiver:
 - (a) segregation of noisy processes;
 - (b) use of remote control;
 - (c) use of flexible exhaust hoses to ensure the exhaust is discharged away from the operator(s).
- 7 active noise control: where the addition of a second source with the same amplitude but with reversed phase causes destructive superposition.

Control of noise exposure for the receiver

Reduction of the time for which a worker is exposed to high levels of noise will achieve a lowering of their noise dose. However, a short period in a high level will increase the dose markedly, and no amount of time spent at lower levels can reduce this dose already received. Job rotation within shifts will also allow reduction in time spent in the higher level.

Work study of a task may show that the presence of a worker at their machine is unnecessary throughout the shift – machines may be minded as long as visual contact is maintained. The worker may mind their machine from within an acoustic enclosure possessing a viewing window. When they are required at the machine they may leave this refuge, wearing personal protection for these comparatively short periods.

Ear protection

Ear protection is rarely comfortable when worn for long periods. The isolation, perspiration and enclosed feeling experienced encourage its removal. Once removed, even for the shortest period, the majority of protection it affords is lost as the dose received from the higher level will be large.

Ear protection is, therefore, to be regarded as a last resort measure, emphasis being placed on the reduction of noise at its source and its transmission. There are a number of different types of personal protectors, a brief description of which are given below.

Earmuffs

These usually consist of hard plastic cups which surround the ears. A seal is made with the head by cushions filled with soft foam or a viscous liquid. A headband is used to retain the two cups in the correct position with the appropriate pressure.

Some earmuffs are designed to: emphasize the attenuation of certain frequencies; passively attenuate loud noises more than quiet sounds ('amplitude sensitive'); actively attenuate at certain intensities or frequencies by the use of electronics incorporated within the cup; direct the noise outside the cup and then to generate (as far as possible) the same noise inside the cup but exactly out of phase, which thereby cancels the incident noise.

Earplugs

Earplugs are designed to fit into the ear canal and are generally of three types: permanent, disposable and reusable. The permanent types are available in a variety of sizes and therefore care is needed in selection – if these are required then 'custom moulded' are better than 'universal fitting'. Disposable plugs are made from various compressible materials and if correctly fitted will fit most people. Reusable plugs require regular cleaning and replacement due to the degradation of the material over time.

Attenuation of earmuffs and earplugs

As the working population differs in terms of head size and shape, ear size and shape, etc. manufacturers quote a mean attenuation and, to provide a degree of uncertainty, its standard deviation. These data are generated by determining the level of hearing in each octave band of a group of subjects both with and without the protection. As it is better to 'fail-safe', a level of

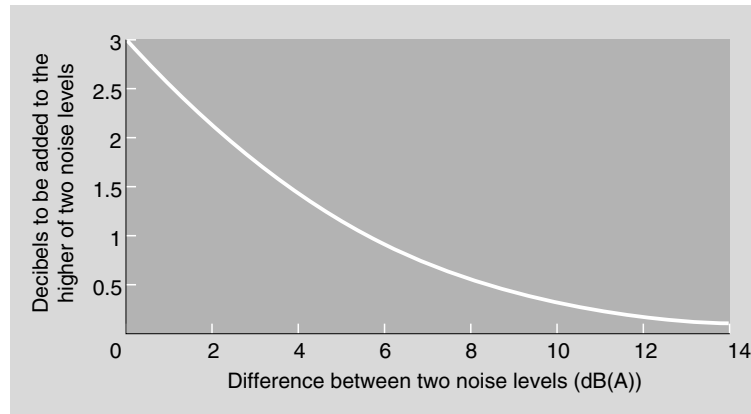


Figure 17.25 Noise level addition chart.

Mid-frequency	63	125	250	500	1k	2k	4k	8k	Hz
SPL	104	98	95	87	84	80	78	72	dB
Correction	-26	-16	-9	-3	0	+1	+1	-1	dB
Corrected level	78	82	86	84	84	81	79	71	dB

	83		88		85		80	
		89					86	
			90 dB(A)					

∴ Estimated level = 90 dB(A).

Tests on the earmuff to be worn give the following results:

Mid-frequency	63	125	250	500	1k	2k	4k	8k	Hz
Mean attenuation	-	15	19	25	28	39	46	43	dB
Standard deviation	-	1.5	2	2.1	1.7	1.7	1.5	2.6	dB
Assumed protection	0	13.5	17	22.9	26.3	37.3	44.5	40.4	dB

Therefore, received 'A' weighted levels with earmuff are:

Mid-frequency	63	125	250	500	1k	2k	4k	8k	Hz
Corrected level	78	82	86	84	84	81	79	71	dB
Assumed protection	0	13.5	17	22.9	26.3	37.3	44.5	40.4	dB
Received level	78	68.5	69	61.1	57.7	43.7	34.5	30.6	dB

	78		69		57.7		36	
		78					58	
			78 dB(A)					

∴ Estimated received level with earmuffs = 78 dB(A).

Noise Survey Report

Name and Address of Premises

Date of Survey Survey Conducted by

Equipment Used Date of Last Calibration

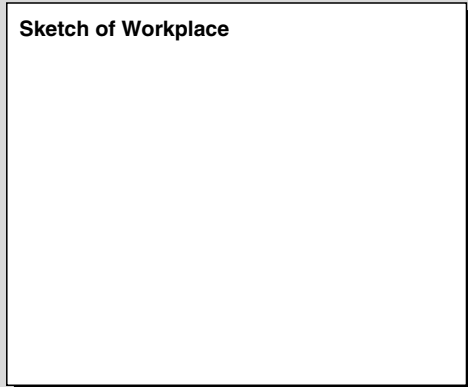
Description of Workplace

Location	Number of Persons Exposed	Noise level (L_{eq})	Duration of Exposure	$L_{EP,d}$	Peak Pressure	Comments

Octave Band Data dB

Location	63 Hz	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz

Sketch of Workplace



Signature

Date

Figure 17.26 Sample noise survey report.

'assumed protection' is calculated by subtraction of one standard deviation away from the mean in that octave band. However, this still may leave a proportion of the population underprotected and therefore it may be prudent to subtract two standard deviations.

Calculation of received noise level when ear protection is worn

The 'assumed protection' is subtracted from an octave band spectrum of the offending noise, 'A' weighted corrections are made in each band and the total 'A' weighted level calculated.

Example: a plant room containing a large compressor yields the following octave band analysis near the ear of the attendant:

Mid-frequency	63	125	250	500	Hz
SPL	104	98	95	87	dB
Mid-frequency	1k	2k	4k	8k	Hz
SPL	84	80	78	72	dB

The total level in dB(A) may be found by correcting this spectrum using Table 17.2 and summation of the different intensities by use of Fig. 17.25 (see also section 'Properties of the decibel scale').

It will be noted that, as the calculations have been performed approximately, the results can only be estimates for a given individual.

Other tests on ear protection

Some countries have developed standard methods of testing the physical properties of ear protection: their behaviour in extremes of temperature, reaction to prolonged vibration and consistency of springiness of the headband of an earmuff. Such tests try to simulate the conditions that the device will encounter in the workplace, and indicate the likelihood of the ear protection retaining its initial fit and attenuation.

Survey report

As with all occupational hygiene survey work, the accurate documentation of sampling strategy issues (see Chapter 11) such as when, where, what was happening, etc., along with the actual results, should always be made at the time of the survey. This can be facilitated by the use of a record sheet. Although no single means of recording data would be suitable and sufficient for all situations, an example is given in Fig. 17.26.

Usually of great importance in noise surveys is the location of the person or process of interest relative to walls, other machines, etc. and the number of machines operative and inoperative at the time. This information should all be recorded. It is also hoped that the 'sketch of the workplace' section of the record sheet will be used to undertake noise mapping. This is where sound measurements around a machine or process are taken and lines drawn between the points of equal intensity. These noise contours assist in the identification of areas likely to give rise to excessive exposure and where personal hearing protectors may be necessary. If a number of these are completed for different situations (i.e. a different number of machines operative or variations in production rate) in the same locality, the magnitude of the effect of each is very apparent.

References

- Department of Health and Social Security (1982). *Report of the Industrial Injuries Advisory Council*. HMSO, London.
- Health and Safety Executive (1990). *Noise at Work – Noise Assessment, Information and Control*. Noise Guides 3 to 8. HMSO, London.
- Waldron, H.A. (ed.) (1989). *Occupational Health Practice*, (3rd edn). Butterworths, Sevenoaks.

Chapter 18

Vibration

Michael J. Griffin

Introduction	
Characteristics of vibration	
Vibration magnitude	
Vibration frequency	
Vibration direction	
Vibration duration	
Hand-transmitted vibration	
Vascular effects (vibration-induced white finger)	
Signs and symptoms	
Diagnosis	
Neurological effects	
Musculoskeletal effects	
Other effects	
Tools and processes causing hand-transmitted vibration	
Preventative measures for hand-transmitted vibration	
National and international standards	
Vibration measurement	
Vibration evaluation	
	Vibration assessment according to ISO 5349 (2001)
	EU Machinery Safety Directive
	EU Physical Agents Directive (2002)
	Whole-body vibration
	Discomfort caused by whole-body vibration
	Effects of vibration magnitude
	Effects of vibration frequency and direction
	Effects of vibration duration
	Vibration in buildings
	Health effects of whole-body vibration
	Evaluation of whole-body vibration
	Assessment of whole-body vibration
	EU Machinery Safety Directive
	EU Physical Agents Directive (2002)
	Interference with activities by whole-body vibration
	Control of whole-body vibration
	Conclusions
	References

Introduction

Vibration is oscillatory motion. The human body is exposed to vibration in many occupations. The effects may be variously described as pleasant or unpleasant, insignificant or interesting, beneficial or harmful. This chapter introduces the various human responses to vibration, defines methods of evaluating occupational exposures to vibration, and lists possible preventative procedures. Guidance on the application of alternative evaluation procedures will be found in the relevant guides, standards, legislation and other texts (e.g. Griffin, 1990).

Hand-transmitted vibration is the vibration that enters the body through the hands. It is caused by various processes in industry, agriculture, mining and construction, where vibrating tools or work pieces are grasped or pushed by the hands or fingers. Exposure to hand-transmitted vibration can lead to the development of several disorders.

Whole-body vibration occurs when the body is supported on a surface that is vibrating (e.g. sitting on a seat that vibrates, standing on a vibrating floor or lying on a vibrating surface). Whole-body vibration occurs in all forms of transport and when working near some industrial machinery.

Characteristics of vibration

Vibration magnitude

During the oscillatory displacements of an object, it has alternately a velocity in one direction and then a velocity in the opposite direction. This change of velocity means that the object is constantly accelerating, first in one direction and then in the opposite direction. A vibration can be quantified by its displacement, its velocity or its acceleration. For practical convenience, the magnitude of vibration is now usually expressed in terms of the acceleration and measured using

accelerometers. The units of acceleration are metres per second per second (i.e. m/s^2 or m s^{-2}). The acceleration due to gravity on Earth is approximately 9.81 m s^{-2} .

The magnitude of an oscillation could be expressed as the distance between the extremities reached by the motion (i.e. the peak-to-peak acceleration) or the maximum deviation from some central point (i.e. the peak acceleration). The magnitude of vibration is now most commonly expressed in terms of an average measure of the acceleration of the oscillatory motion, usually the root-mean-square value (i.e. m s^{-2} r.m.s.). For a sinusoidal motion, the r.m.s. value is the peak value divided by $\sqrt{2}$ (i.e. approximately 1.4).

When observing vibration it is sometimes possible to estimate the displacement caused by the motion. For a sinusoidal motion, the acceleration, a , can be calculated from the frequency, f , in Hz, and the displacement, d :

$$a = (2\pi f)^2 d \quad (18.1)$$

For example, a sinusoidal motion with a frequency of 1 Hz and a peak-to-peak displacement of 0.1 m will have an acceleration of 3.95 m s^{-2} peak-peak, 1.97 m s^{-2} peak, and 1.40 m s^{-2} r.m.s. The above expression may be used to convert acceleration measurements to corresponding displacements. However, the conversion is accurate only when the motion occurs at a single frequency (i.e. it has a sinusoidal waveform).

Logarithmic scales for quantifying vibration magnitudes in decibels (dB) are sometimes used. When using the reference level in International Standard ISO 1683, the acceleration level, L_a , is expressed by $L_a = 20 \log_{10} (a/a_0)$, where a is the measured acceleration (in m s^{-2}) and a_0 is the reference level of 10^{-6} m s^{-2} . With this reference, an acceleration of 1.0 m s^{-2} corresponds to 120 dB; an acceleration of 10 m s^{-2} corresponds to 140 dB. Other reference levels are used in some countries.

Vibration frequency

The frequency of vibration is expressed in cycles per second using the SI unit, hertz (Hz). The frequency of vibration greatly influences the extent

to which vibration is transmitted to the surface of the body (e.g. through seating), the extent to which it is transmitted through the body (e.g. from fingers to the arm) and the response to vibration within the body. From the section on Vibration magnitude it will be seen that the relation between the displacement and the acceleration of a motion is also dependent on the frequency of oscillation: a displacement of 1 mm will correspond to a low acceleration at low frequencies but a very high acceleration at high frequencies. Consequently, the vibration visible to the human eye does not provide a good indication of vibration acceleration.

Oscillations of the whole body at frequencies below about 0.5 Hz can cause motion sickness. The frequencies of greatest significance to whole-body vibration are usually at the lower end of the range from 0.5 to 100 Hz. For hand-transmitted vibration, frequencies as high as 1000 Hz or more may have detrimental effects.

Vibration direction

The responses of the body differ according to the direction of the motion. Vibration is usually measured at the interfaces between the body and the vibrating surfaces in three orthogonal directions. Fig. 18.1 shows a coordinate system used when measuring vibration in contact with a hand holding a tool. The axes may differ at the second handle on the same tool: a diagram is often necessary to identify the axes of measurement.

The three principal directions for seated and standing persons are: fore-and-aft (x -axis), lateral (y -axis) and vertical (z -axis). The vibration is measured at the interface between the body and the surface supporting the body (i.e. between the seat and the ischial tuberosities, for a seated person, beneath the feet for a standing person). Figure 18.2 illustrates the translational and rotational axes for an origin at the ischial tuberosities of a seated person. A similar set of axes is used for describing the directions of vibration at the back and feet of seated persons. The direction of vibration of a control held in the hand or a display viewed by the eyes can also be important, with the axes defined using similar principles.

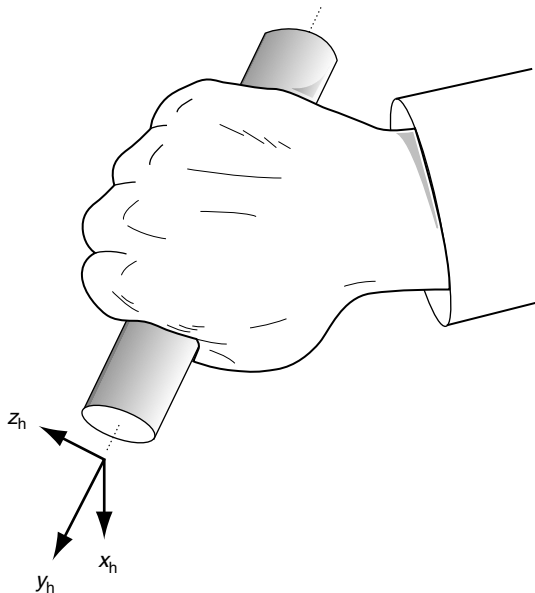


Figure 18.1 Axes of vibration used to measure hand-transmitted vibration.

Vibration duration

Some effects of vibration depend on the total duration of vibration exposure. Additionally, the duration of measurement may affect the measured magnitude of the vibration. The r.m.s. acceleration will not provide a useful indication of the relative vibration severity of vibrations that differ in duration. The r.m.s. value is also of limited usefulness if the vibration is intermittent, contains shocks or otherwise varies in magnitude from time to time (see Health effects of whole-body vibration).

Hand-transmitted vibration

Prolonged and regular exposure of the fingers or the hands to vibration or repeated shock can give rise to various signs and symptoms of disorder. The precise extent and inter-relation between the signs and symptoms are not fully understood but five types of disorder may be identified (see Table 18.1).

The various disorders may be inter-connected; more than one disorder can affect a person at the same time and it is possible that the presence of

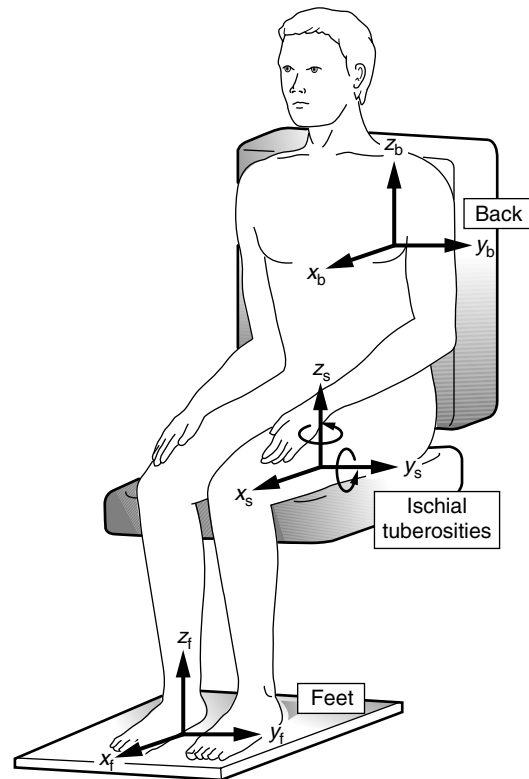


Figure 18.2 Axes of vibration used to measure whole-body vibration.

one disorder facilitates the appearance of another. The onset of each disorder is dependent on several variables, such as the vibration characteristics, the dynamic response of the fingers or hand, individual susceptibility to damage and other aspects of the environment. The terms ‘vibration

Table 18.1 Five types of disorder associated with hand-transmitted vibration exposures.

Type	Disorder
Type A	Circulatory disorders
Type B	Bone and joint disorders
Type C	Neurological disorders
Type D	Muscle disorders
Type E	Other general disorders (e.g. central nervous system)

Some combination of these disorders is sometimes referred to as the ‘hand–arm vibration syndrome’ (HAVS).

syndrome' or 'hand–arm vibration syndrome' (HAVS) are sometimes used to refer to one or more of the disorders listed in Table 18.1.

Vascular effects (vibration-induced white finger)

The first published cases of the condition now most commonly known as 'vibration-induced white finger' (VWF) are assumed to be those reported in Italy by Loriga in 1911 (Loriga, 1911). A few years later, cases were documented at limestone quarries in Indiana. Vibration-

induced white finger has subsequently been reported to occur in many other widely varied occupations in which there is exposure of the fingers to vibration (see Griffin, 1990). Table 18.2 summarizes the state of knowledge related to the symptoms and signs of vibration-induced white finger and related matters (see Griffin and Bovenzi, 2002).

Signs and symptoms

VWF is characterized by intermittent blanching of the fingers. The fingertips are usually the first to blanch but the affected area may extend to all of one or more fingers with continued vibration

Table 18.2 Vibration-induced white finger (see Griffin and Bovenzi, 2002).

The nature of the vibration-induced white finger

Vibration-induced white finger is a disorder characterized by complete episodic closure of digital blood vessels. Both central and local pathogenic mechanisms may be involved. The pathogenesis of vibration-induced white finger is not yet fully understood

Symptoms of vibration-induced white finger

A necessary symptom for the diagnosis of vibration-induced white finger is the occurrence of attacks of well-demarcated finger blanching (Raynaud's phenomenon)

- Attacks of blanching normally commence with blanching in the distal phalanges and may extend to other more proximal phalanges before receding to the distal phalanges and recovery
- Blotchiness (patches of blanching) may occur during onset or recovery from an attack
- Anaesthesia will occur during an attack of blanching but numbness may not be noticed
- There may be a sequence of colour changes in which blanching is followed by cyanosis and redness, sometimes accompanied by pain
- Attacks are mainly provoked by exposure to cold conditions (including dampness) but cold will not necessarily provoke an attack
- Persons with vibration-induced vascular disorders may feel their fingers to be abnormally cold, even without a blanching attack

Signs and objective tests of vascular disorder

A sufficient sign of vibration-induced white finger is the observation of an attack of well-demarcated finger blanching

- Finger systolic blood pressures measured following cooling of the digits to 15°C and 10°C will often be low in persons with vibration-induced white finger; a finger systolic blood pressure of approximately zero can verify an attack of Raynaud's phenomenon
- Finger rewarming times following cold exposure may be prolonged
- Standardization of cold provocation (rewarming) tests is desirable
- Tests are recommended on all potentially affected digits on both hands
- Current objective tests (finger systolic blood pressures and rewarming times following cold provocation) do not indicate the severity of vibration-induced white finger and are therefore not required if an attack of finger blanching has been witnessed

Other considerations

- Vascular damage caused by hand-transmitted vibration should be distinguished from primary Raynaud's phenomenon and other causes of secondary Raynaud's phenomenon
- Effects of age, smoking, medication and vasoactive agents should be taken into account

Minimal vibration exposure required for diagnosis

Regular exposure to vibration known to be capable of causing vibration-induced white finger

exposure. Attacks of blanching are precipitated by cold and therefore usually occur in cold conditions or when handling cold objects. The blanching lasts until the fingers are rewarmed and vasodilatation allows the return of the blood circulation.

Many years of vibration exposure often occur before the first attack of blanching is noticed. Affected persons often have other signs and symptoms, such as numbness and tingling. Cyanosis and, rarely, gangrene have also been reported. It is not yet clear to what extent these other signs and symptoms are causes of, caused by or unrelated to attacks of 'white finger'.

Diagnosis

There are other conditions that can cause similar signs and symptoms to those associated with VWF. Vibration-induced white finger cannot be assumed to be present merely because there are attacks of blanching. It will be necessary to exclude other known causes of similar symptoms (by medical examination) and also necessary to exclude so-called primary Raynaud's disease (also called 'constitutional white finger'). This exclusion cannot yet be achieved with complete confidence but if there is no family history of the symptoms, if the symptoms did not occur before the first significant exposure to vibration, and if the symptoms and signs are confined to areas in contact with the vibration (e.g. the fingers, not the ears, etc.), they will often be assumed to indicate vibration-induced white finger.

Diagnostic tests for vibration-induced white finger can be useful but, at present, they are not

infallible indicators of the disease. The measurement of finger systolic blood pressure during finger cooling and the measurement of finger rewarming times following cooling can be useful, but many others tests are in use.

The severity of the effects of hand-transmitted vibration is sometimes recorded by reference to the 'stage' of the disorder. The staging of vibration-induced white finger is based on verbal statements made by the affected person. In the 'Taylor-Pelmeur' system, the stage of vibration-induced white finger was determined by the presence of numbness and tingling, the areas affected by blanching, the frequency of blanching, the time of year when blanching occurred and the extent of interference with work and leisure activities (see Taylor *et al.*, 1974). A slightly more simple procedure, the Stockholm Workshop staging system, was subsequently evolved (see Table 18.3; Gemne *et al.*, 1987). However, this staging system compounds the frequency of attacks of blanching with the areas of the digits affected by blanching.

A numerical procedure for recording the areas of the digits affected by blanching is known as the 'scoring system' (Fig. 18.3). The blanching scores for the hands shown in Fig. 18.3 are 01300_{right} and 01366_{left}. The scores correspond to areas of blanching on the digits commencing with the thumb. On the fingers, a score of 1 is given for blanching on the distal phalanx, a score of 2 for blanching on the middle phalanx and a score of 3 for blanching on the proximal phalanx. On the thumbs, the scores are 4 for the distal phalanx and 5 for the proximal phalanx. The blanching

Table 18.3 Stockholm Workshop scale for the classification of vibration-induced white finger (Gemne *et al.*, 1987).

Stage	Grade	Description
0	-	No attacks
1	Mild	Occasional attacks affecting only the tips of one or more fingers
2	Moderate	Occasional attacks affecting distal and middle (rarely also proximal) phalanges of one or more fingers
3	Severe	Frequent attacks affecting all phalanges of most fingers
4	Very severe	As in stage 3, with trophic skin changes in the fingertips

If a person has stage 2 in two fingers of the left hand and stage 1 in a finger on the right hand then the condition may be reported as 2L(2)/1R(1). There is no defined means of reporting the condition of digits when this varies between digits on the same hand. The scoring system in Fig. 18.3 is more helpful when the extent of blanching is to be recorded.

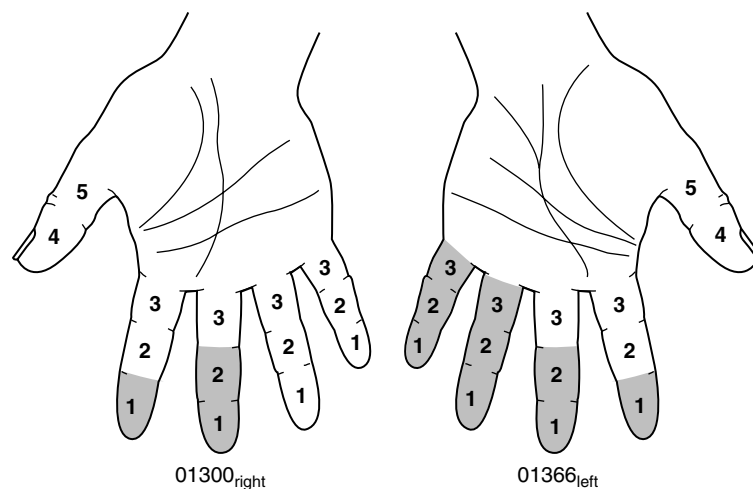


Figure 18.3 Method of scoring the areas of the digits affected by blanching (from Griffin, 1990).

score may be based on statements from the affected person or on the visual observations of a designated observer (e.g. a nurse).

Neurological effects

Neurological effects of hand-transmitted vibration (e.g. numbness, tingling, elevated sensory thresholds for touch, vibration, temperature and pain, and reduced nerve conduction velocity) are now recognized as separate effects of vibration and not merely symptoms of vibration-induced white finger. Carpal tunnel syndrome seems to be associated with the use of hand-held vibratory tools but there is not universal agreement on the extent to which carpal tunnel syndrome is caused by vibration as opposed to other aspects of the job, such as a tight grip and repetitive hand movements. Table 18.4 summarizes the state of knowledge related to neurological disorders caused by work with vibratory tools.

A method of reporting the extent of vibration-induced neurological effects of vibration has been proposed (see Table 18.5). The 'sensorineural stage' is a subjective impression of a physician based on the statements of the affected person and the results of any available clinical or scientific testing.

Musculoskeletal effects

The literature includes several reports of musculoskeletal disorders in users of vibratory tools. Workers exposed to hand-transmitted vibration sometimes report difficulty with their grip, including reduced dexterity, reduced grip strength and locked grip. Many of the reports are derived from the symptoms of exposed persons rather than signs detected by physicians and could be a reflection of neurological problems. Measurements of muscle function have rarely been obtained using repeatable tests. Table 18.6 summarizes the state of knowledge related to musculoskeletal disorders caused by work with vibratory tools.

Muscle activity may be of great importance to tool users since a secure grip can be essential to the performance of the job and safe control of the tool. The presence of vibration on a handle may encourage the adoption of a tighter grip than would otherwise occur and a tight grip may increase the transmission of vibration to the hand. If the chronic effects of vibration result in reduced grip this may help to protect operators from further effects of vibration, but interfere with both work and leisure activities.

Surveys of the users of hand-held tools have found evidence of bone and joint problems: most often among men operating percussive tools, such

Table 18.4 Neurological disorders caused by work with vibratory tools (see Griffin and Bovenzi, 2002).*The nature of the neurological disorders*

Neuropathy to peripheral, mainly the sensory but sometimes also the motor, nervous system, related to work with vibrating machines in which:

- There may be disorders of end organs
- There may be nerve fibre dysfunction resembling entrapment neuropathy
- There may be diffuse or multifocal neuropathy
- Any of the nerves of the upper limbs may be affected by hand-transmitted vibration
- The disorder is not necessarily confined to digits but extending to the palm and the arms
- The involvement of the autonomic nervous system has been considered, but this was not the subject of the workshop

Symptoms of neurological disorders

There are no minimal symptoms of neurological disorders caused by hand-transmitted vibration because manifestations of disorder can pass unnoticed by affected persons:

- Numbness and tingling are commonly reported
- It is desirable to unify the terminology for the description of symptoms in different languages

Signs and objective tests of neurological disorders

There are no minimal signs of neurological disorder – symptoms can exist without signs:

- A thorough neurological and musculoskeletal physical examination is a prerequisite for the diagnosis and interpretation of any objective tests
- Useful objective measures include sensory tests (e.g. thresholds for heat, cold and vibration and aesthesiometry) and electrodiagnostic testing
- Standardization of tests is desirable

Other considerations

- Endocrine, metabolic and immunological disorders, traumatic injuries, infections, polyneuropathies and idiopathic focal neuropathies should be excluded
- Effects of age, smoking, alcohol, medication and neurotoxic agents should be taken into account

Minimal vibration exposure required for diagnosis

Exposure to vibration known to be capable of causing neurological disorders:

- There is currently no established exposure–response relationship between the physical characteristics of occupational exposures to hand-transmitted vibration and the development of neurological disorders

as those used in metalworking jobs and mining and quarrying. It is speculated that a characteristic of such tools, possibly the low-frequency shocks, is responsible. Some of the reported injuries relate to specific bones and suggest the existence of cysts, vacuoles, decalcification, or other osteolysis

and degeneration or deformity of the carpal, metacarpal or phalangeal bones. Osteoarthritis and olecranon spurs at the elbow, and other problems at the wrist and shoulder, are also documented.

Notwithstanding the evidence of many research publications, there is not universal acceptance that vibration is the cause of articular problems and at present there is no dose–effect relation that predicts their occurrence. In the absence of specific information, it seems that adherence to current guidance for the prevention of vibration-induced white finger may provide reasonable protection.

Table 18.5 Proposed ‘sensorineural stages’ of the effects of hand-transmitted vibration (Brammer *et al.*, 1987).

Stage	Symptoms
0 _{SN}	Exposed to vibration but no symptoms
1 _{SN}	Intermittent numbness with or without tingling
2 _{SN}	Intermittent or persistent numbness, reduced sensory perception
3 _{SN}	Intermittent or persistent numbness, reduced tactile discrimination and/or manipulative dexterity

Other effects

Effects of hand-transmitted vibration may not be confined to the fingers, hands and arms: many studies have found a high incidence of problems

Table 18.6 Musculoskeletal disorders caused by work with vibratory tools (see Griffin and Bovenzi, 2002).*The nature of the musculoskeletal disorders*

The pathophysiological mechanisms underlying musculoskeletal disorders in workers using vibratory tools are often unclear

Symptoms of musculoskeletal disorders

- The most common symptom of musculoskeletal disorders is pain (type, onset and location of pain should be explored)
- It is desirable to unify the terminology for the description of symptoms of musculoskeletal disorders in different languages

Signs and objective tests of musculoskeletal disorders

There are no minimal signs of musculoskeletal disorders – symptoms can exist without signs:

- A thorough neuromuscular and skeletal physical examination is a prerequisite for the diagnosis and interpretation of any objective tests of musculoskeletal disorders
- Work with low-frequency percussive tools may be associated with an increased occurrence of abnormal radiological findings in the wrist and elbow joints (e.g. premature osteoarthritis, exostoses at the sites of tendon insertion)

Other considerations

- Systemic inflammatory disorders, neuromuscular diseases, endocrine disorders, traumatic injuries and infections or tumours should be excluded
- Confounding and effect-modifying variables (e.g. gender, age, smoking and other personal characteristics) should be taken into account

Minimal vibration exposure required for diagnosis

- The relative importance of hand-transmitted vibration and other ergonomic and psychosocial risk factors in the causation of musculoskeletal disorders is often unclear
- At present, there is no established exposure–response relationship between the physical characteristics of occupational exposures to hand-transmitted vibration and the development of any musculoskeletal disorder

such as headaches and sleeplessness among tool users and have concluded that these symptoms are caused by hand-transmitted vibration. Although these are real problems to those affected, they are ‘subjective’ effects that are not accepted as real by all researchers. Some research is seeking a physiological basis for such symptoms. At present it would appear that caution is appropriate, but it is reasonable to assume that the adoption of the modern guidance to prevent vibration-induced white finger will also provide some protection from any other effects of hand-transmitted vibration within, or distant from, the hand.

Tools and processes causing hand-transmitted vibration

The vibration on tools varies greatly depending on tool design and method of use, so it is not possible to categorize individual tool types as ‘safe’ or ‘dangerous’. However, Table 18.7 lists tools and processes that are common causes of vibration-induced injury according to the UK Health and Safety Executive (1994).

Preventative measures for hand-transmitted vibration

Protection from the effects of hand-transmitted vibration requires actions from management, tool manufacturers, technicians and physicians at the workplace and from tool users. Table 18.8 summarizes some of the actions that may be appropriate.

When there is reason to suspect that hand-transmitted vibration may cause injury, the vibration at tool–hand interfaces should be measured. It will then be possible to predict whether the tool or process is likely to cause injury and whether any other tool or process could give a lower vibration severity.

The duration of exposure to vibration should also be quantified. Reduction of exposure duration may include the provision of exposure breaks during the day and, if possible, prolonged periods away from vibration exposure. For any tool or process having a vibration magnitude sufficient to cause injury there should be a system to quantify and control the maximum daily duration of exposure of any individual.

Table 18.7 Tools and processes potentially associated with vibration injuries (from Health and Safety Executive, 1994).

<i>Category of tool</i>	<i>Examples of tool</i>
Percussive metalworking tools	Powered percussive metalworking tools, including powered hammers for riveting, caulking, hammering, clinching and flanging; hammer swaging
Percussive tools used in stoneworking, quarrying, construction, etc.	Percussive hammers, vibratory compactors, concrete breakers, pokers, sanders and drills used in mining, quarrying, demolition and road construction, etc.
Grinders and other rotary tools	Pedestal grinders, hand-held portable grinders, flex-driven grinders and polishers, and rotary burring tools
Timber and woodworking machining tools	Chain saws, brush cutters (clearing saws), hand-held or hand-fed circular saws, electrical screwdrivers, mowers and shears, hardwood cutting machines, barking machines and strimmers
Other processes and tools	Pounding machines used in shoe manufacture, drain suction machines, nut runners, concrete vibro-thickeners, and concrete levelling vibro-tables

The Health and Safety Executive suggests that health surveillance is likely to be appropriate for all workers using these vibratory tools.

Gloves are sometimes recommended as a means of reducing the adverse effects of vibration on the hands. When using the frequency weightings in current standards, commonly available gloves do *not* normally provide effective attenuation of the vibration on most tools. Gloves and 'cushioned' handles may reduce the transmission of high frequencies of vibration but current standards imply that these are not usually the primary cause of disorders. Gloves may protect the hand from other forms of mechanical injury (e.g. cuts and scratches) and protect the fingers from temperature extremes. Warm hands are less likely to suffer an attack of finger blanching and some consider that maintaining warm hands while exposed to vibration may also lessen the damage caused by the vibration.

Workers who are exposed to vibration magnitudes sufficient to cause injury should be warned of the possibility of vibration injuries and educated on the ways of reducing the severity of their vibration exposures. They should be advised of the symptoms to look out for and told to seek medical attention if the symptoms appear.

There should be pre-employment medical screening wherever a subsequent exposure to hand-transmitted vibration may reasonably be expected to cause vibration injury. Medical supervision of each exposed person should continue throughout employment at suitable intervals, pos-

sibly annually. There is no single test that will diagnose the existence or extent of all possible effects of hand-transmitted vibration. Tests in common use include direct and indirect measures of finger blood flow, the measurement of finger systolic blood pressure, the determination of various tactile thresholds and more extensive neurological investigations. Although these and other investigations may assist the diagnosis of specific disorders, their sensitivities and specificities are currently unknown (Faculty of Occupational Medicine of the Royal College of Physicians, Working Party on Hand-transmitted Vibration, 1993).

National and international standards

There are various standards for the measurement, evaluation and assessment of hand-transmitted vibration.

Vibration measurement

International Standards ISO 5349-1 (2001) and ISO 5349-2 (2002) give recommendations on methods of measuring the hand-transmitted vibration on tools and processes. Guidance on vibration measurements on specific tools is given elsewhere, such as in various parts of International Standard ISO 8662.

Table 18.8 Some preventative measures to consider when persons are exposed to hand-transmitted vibration (adapted from Griffin, 1990, Chapter 19).

<i>Group</i>	<i>Action</i>
Management	Seek technical advice Seek medical advice Warn exposed persons Train exposed persons Review exposure times Provide policy on removal from work
Tool manufacturers	Measure tool vibration Design tools to minimize vibration Have ergonomic design to reduce grip force, etc. Design to keep hands warm Provide guidance on tool maintenance Provide warning of dangerous vibration
Technical at workplace	Measure vibration exposure Provide appropriate tools Maintain tools Inform management
Medical	Provide pre-employment screening Provide routine medical checks Record all signs and reported symptoms Warn workers with predisposition Advise on consequences of exposure Inform management
Tool user	Use tool properly Avoid unnecessary vibration exposure Minimize grip and push forces Check condition of tool Inform supervisor of tool problems Keep warm Wear gloves when safe to do so Minimize smoking Seek medical advice if symptoms appear Inform employer of relevant disorders

Care is required to obtain representative measurements of tool vibration with appropriate operating conditions. There can be difficulties in obtaining valid measurements using some

commercial instrumentation (especially when there are high shock levels). It is wise to determine acceleration spectra and inspect the acceleration time histories before accepting the validity of any measurements.

Vibration evaluation

All national and international standards use the same frequency weighting (called W_h) to evaluate hand-transmitted vibration over the approximate frequency range 8–1000 Hz. This weighting is applied to measurements of vibration acceleration in each of the three axes of vibration at the point of entry of vibration to the hand.

The frequency-weighted acceleration on different tools may be compared. The standards imply that if two tools expose the hand to vibration for the same period of time, the tool having the lowest frequency-weighted acceleration will be least likely to cause injury or disease.

Occupational exposures to hand-transmitted vibration can have widely varying daily exposure durations – from a few seconds to many hours. Often, exposures are intermittent. To enable a daily exposure to be reported simply, the standards refer to an equivalent 8-h exposure:

$$a_{\text{hw(eq, 8h)}} = A(8) = a_{\text{hw}} \left[\frac{t}{T_{(8)}} \right]^{1/2} \quad (18.2)$$

where t is the exposure duration to an r.m.s. frequency-weighted acceleration, a_{hw} , and $T_{(8)}$ is 8 h (in the same units as t).

Vibration assessment according to ISO 5349 (2001)

ISO 5349-1 (International Organization for Standardization (ISO), 2001) uses the W_h frequency weighting with the assessment of exposure based on the root-sums-of-squares of the r.m.s. acceleration occurring in all three axes.

In an informative (i.e. not normative) annex of ISO 5349-1 (ISO, 2001), there is a suggested relation between the lifetime exposure to hand-transmitted vibration, D_y , (in years) and the 8-h energy-equivalent daily exposure $A(8)$ for the

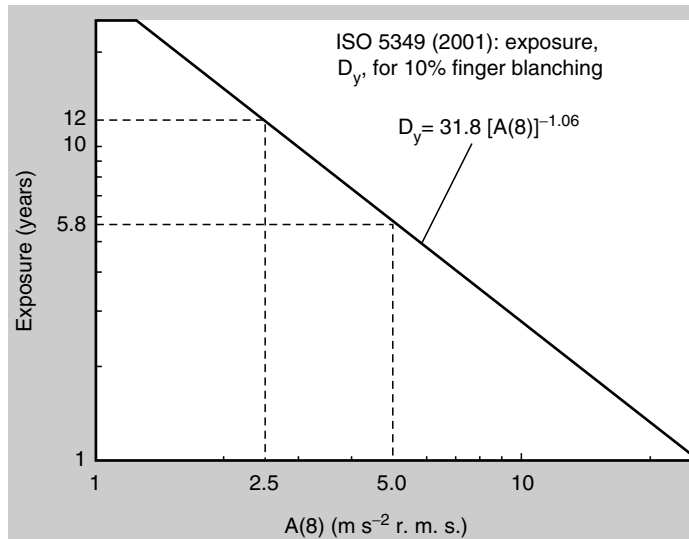


Figure 18.4 Daily equivalent 8-h exposures expected to produce a 10% prevalence of finger blanching according to ISO 5349-1 (ISO, 2001) for exposures of between 1 and 25 years.

conditions expected to cause 10% prevalence of finger blanching:

$$Dy = 31.8[A(8)]^{-1.06} \quad (18.3)$$

Figure 18.4 shows the daily equivalent 8-h exposures expected to produce a 10% prevalence of finger blanching according to ISO 5349-1 (ISO, 2001) for exposures of up to 25 years.

The percentage of affected persons in any group of exposed persons will not always closely match the values shown in Fig. 18.4. The frequency weighting, the time dependency and the dose-effect information are based on less than complete information and they have been simplified for practical convenience. Additionally, the number of persons affected by vibration will depend on the rate at which persons enter and leave the exposed group. The complexity of the above equation makes it unusable by some and it implies far greater precision than is possible. A more convenient estimate of the years of exposure (in the range 1 to 25 years) required for 10% incidence of finger blanching is:

$$Dy = 30.0/A(8) \quad (18.4)$$

This equation gives the same result as the equation in the standard (to within 14%).

The informative annex to ISO 5349 (ISO, 2001) states:

Studies suggest that symptoms of the hand-arm vibration syndrome are rare in persons exposed with an 8-h energy-equivalent vibration total value, $A(8)$, at a surface in contact with the hand, of less than 2 m/s^2 and unreported for $A(8)$ values less than 1 m/s^2 .

However, this sentence should not be interpreted too literally in view of the very considerable doubts over the frequency weighting and time dependency in the standard (see Griffin *et al.*, 2003).

EU Machinery Safety Directive

The Machinery Safety Directive of the European Community (89/392/EEC) states that machinery must be designed and constructed so that hazards resulting from vibration produced by the machinery are reduced to the lowest practicable level, taking into account technical progress and the availability of means of reducing vibration. The instruction handbooks for hand-held and hand-guided machinery must specify the equivalent acceleration to which the hands or arms are subjected where this exceeds some stated value (proposed at present as a frequency-weighted acceleration of 2.5 m s^{-2} r.m.s.). The relevance of any such value will depend on the test conditions to be specified in other standards. Very many hand-held vibrating tools can exceed this value.

Standards defining test conditions for the measurement of vibration on chipping and riveting

hammers, rotary hammers and rock drills, grinding machines, pavement breakers and various garden and forestry equipment (including chain saws) are in preparation (see International Standard ISO 8662.1) (International Organization for Standardization, 1988).

EU Physical Agents Directive (2002)

In 2002, the Parliament and Commission of the European Community agreed 'minimum health and safety requirements' for the exposure of workers to the risks arising from vibration. For hand-transmitted vibration, the Directive defines an 8-h equivalent 'exposure action value' of 2.5 m s^{-2} r.m.s. and an 8-h equivalent 'exposure limit value' of 5.0 m s^{-2} r.m.s. Member States of the European Union must bring into force laws to comply with the Directive by 6 July 2005.

The Directive says workers shall not be exposed above the 'exposure limit value'. If the 'exposure action values' are exceeded, the employer shall establish and implement a programme of technical and/or organizational measures intended to reduce to a minimum exposure to mechanical vibration and the attendant risks. The Directive makes it clear that workers exposed to mechanical vibration in excess of the exposure action values shall be entitled to appropriate health surveillance, but health surveillance is not restricted to situations in which the exposure action value is exceeded; health surveillance is required if there is any reason to suspect that workers may be injured by the vibration, even if the action value is not exceeded.

According to ISO 5349-1 (ISO, 2001), the onset of finger blanching would be expected in 10% of persons after 12 years at the EU 'exposure action value' and after 5.8 years at the 'exposure limit value' (Fig. 18.4). It is clear that the exposure action value and the exposure limit value in the Directive do not define 'safe exposures' to hand-transmitted vibration (Griffin, 2004).

Whole-body vibration

Vibration of the whole body is produced by various types of industrial machinery and by all forms

of transport. The vibration may affect health, comfort and the performance of activities. The comments of persons exposed to vibration mostly derive from the sensations produced by vibration rather than knowledge that the vibration is causing harm or interfering with their activities.

Discomfort caused by whole-body vibration

For very low magnitude motions it is possible to estimate the percentage of persons who will be able to feel vibration and the percentage who will not be able to feel the vibration (Griffin, 1990). For higher vibration magnitudes, an approximate indication of the extent of subjective reactions is available in a semantic scale of discomfort [e.g. British Standard BS 6841 (British Standards Institution (BSI), 1987) and International Standard ISO 2631 (ISO, 1997)].

Any limit to prevent vibration discomfort should vary between different environments (e.g. between buildings and transport) and between different types of transport (e.g. between cars and trucks) and within types of vehicle (e.g. between sports cars and limousines). The design limit depends on external factors (e.g. cost and speed) and the comfort in alternative environments (e.g. competitive vehicles).

Effects of vibration magnitude

The absolute threshold for the perception of vertical whole-body vibration in the frequency range 1–100 Hz is approximately 0.01 m s^{-2} r.m.s.; a magnitude of 0.1 m s^{-2} will be easily noticeable; magnitudes around 1 m s^{-2} r.m.s. are usually considered uncomfortable; magnitudes of 10 m s^{-2} r.m.s. are usually dangerous. The precise values depend on vibration frequency and exposure duration and they are different for other axes of vibration [see British Standard BS 6841 (BSI, 1987) and Griffin, 1990].

A doubling of vibration magnitude (expressed in m s^{-2}) produces an approximate doubling of discomfort. A halving of vibration magnitude can therefore produce a considerable improvement in comfort.

Effects of vibration frequency and direction

The dynamic responses of the body and the relevant physiological and psychological processes dictate that subjective reactions to vibration depend on vibration frequency and vibration direction. Frequency weightings are given in British Standard BS 6841 (BSI, 1987) and International Standard ISO 2631 (ISO, 1997).

Effects of vibration duration

Vibration discomfort tends to increase with increasing duration of exposure to vibration. The precise rate of increase may depend on many factors but a simple 'fourth power' time dependency is sometimes used to approximate how discomfort varies with exposure duration from the shortest possible shock to a full day of vibration exposure (i.e. $\text{acceleration}^4 \times \text{duration} = \text{constant}$). This time dependency is more consistent with available information and expectations than an 'energy time dependence' (see Health effects of whole-body vibration).

Vibration in buildings

Acceptable magnitudes of vibration in buildings are close to vibration perception thresholds. The effects of vibration in buildings are assumed to depend on the use of the building in addition to the vibration frequency, direction and duration. Guidance is given in various standards (e.g. ISO, 1989). British Standard BS 6472 (BSI, 1992) defines a procedure that combines the assessment of vibration and shock in buildings by using the 'vibration dose value'.

Health effects of whole-body vibration

Epidemiological studies have reported disorders among persons exposed to vibration from occupational, sport and leisure activities. The studies do not all agree on either the type or the extent of disorders and rarely have the findings been related to measurements of the vibration exposures. However, it is often assumed that disorders of the back (back pain, displacement of intervertebral discs, degeneration of spinal vertebrae, osteoarthritis,

etc.) may be associated with vibration exposure (Griffin, 1990, Chapter 5 and Appendix 5; Bovenzi and Hulshof, 1998). There may be several alternative causes of an increase in disorders of the back among persons exposed to vibration (e.g. poor sitting postures, heavy lifting). It is not always possible to conclude confidently that a back disorder is solely, or primarily, caused by vibration (Palmer *et al.*, 1999).

Other disorders that have been claimed to be due to occupational exposures to whole-body vibration include abdominal pain, digestive disorders, urinary frequency, prostatitis, haemorrhoids, balance and visual disorders, headaches and sleeplessness. Further research is required to confirm whether these signs and symptoms are causally related to exposure to vibration.

Evaluation of whole-body vibration

Epidemiological data alone are not sufficient to define how to evaluate whole-body vibration so as to predict the relative risks to health from the different types of vibration exposure. A consideration of such data in combination with an understanding of biodynamic responses and subjective responses is used to provide current guidance. The manner in which the health effects of oscillatory motions depend upon the frequency, direction and duration of motion is currently assumed to be similar to that for vibration discomfort. However, it is assumed that the 'total' exposure, rather than the 'average' exposure, is important and so a 'dose' measure is used.

Assessment of whole-body vibration

British Standard BS 6841 (BSI, 1987) and International Standard ISO 2631 (ISO, 1997) give guidance on the severity of exposures to whole-body vibration. There are some similarities between the two standards, but alternative methods within ISO 2631 offer confusingly conflicting guidance (see Griffin, 1998).

British Standard BS 6841 (BSI, 1987)

British Standard BS 6841 defines an 'action level' for vertical vibration-based 'vibration dose values'. The vibration dose value uses a 'fourth

power' time dependency to accumulate vibration severity over the exposure period from the shortest possible shock to a full day of vibration:

$$\text{vibration dose value} = \left[\int_{t=0}^{t=T} a^4(t) dt \right]^{1/4} \quad (18.5)$$

where $a(t)$ is the frequency-weighted acceleration. If the exposure duration (t , seconds) and the frequency-weighted r.m.s. acceleration (a_{rms} , m s^{-2} r.m.s.) are known for conditions in which the vibration characteristics are statistically stationary, it can be useful to calculate the 'estimated vibration dose value', eVDV:

$$\text{estimated vibration dose value} = 1.4 a_{\text{rms}} t^{1/4} \quad (18.6)$$

The eVDV is not applicable to transients, shocks and repeated shock motions in which the crest factor (peak value divided by the r.m.s. value) is high.

No precise limit can be offered to prevent disorders caused by whole-body vibration, but standards define useful methods of quantifying vibration severity. British Standard BS 6841 (BSI, 1987) offers the following guidance.

High vibration dose values will cause severe discomfort, pain and injury. Vibration dose values also indicate, in a general way, the severity of the vibration exposures

which caused them. However there is currently no consensus of opinion on the precise relation between vibration dose values and the risk of injury. It is known that vibration magnitudes and durations which produce vibration dose values in the region of $15 \text{ m s}^{-1.75}$ will usually cause severe discomfort. It is reasonable to assume that increased exposure to vibration will be accompanied by increased risk of injury.

An action level might be set higher or lower than $15 \text{ m s}^{-1.75}$. Figure 18.5 compares this action level with guidance in ISO 2631 (ISO, 1997) (see below).

International Standard ISO 2631 (ISO, 1997)

In International Standard ISO 2631, two different methods of evaluating vibration severity are defined with respect to health effects, and for both methods there are two boundaries. When evaluating vibration using the vibration dose value, it is suggested that below a boundary corresponding to a vibration dose value of $8.5 \text{ m s}^{-1.75}$, 'health risks have not been objectively observed', between 8.5 and $17 \text{ m s}^{-1.75}$, 'caution with respect to health risks is indicated' and above $17 \text{ m s}^{-1.75}$, 'health risks are likely'. The two boundaries define a 'VDV health guidance caution zone'. The alternative method of evaluation in ISO 2631 uses a time dependency in which the acceptable vibration does not vary with duration between 1 and

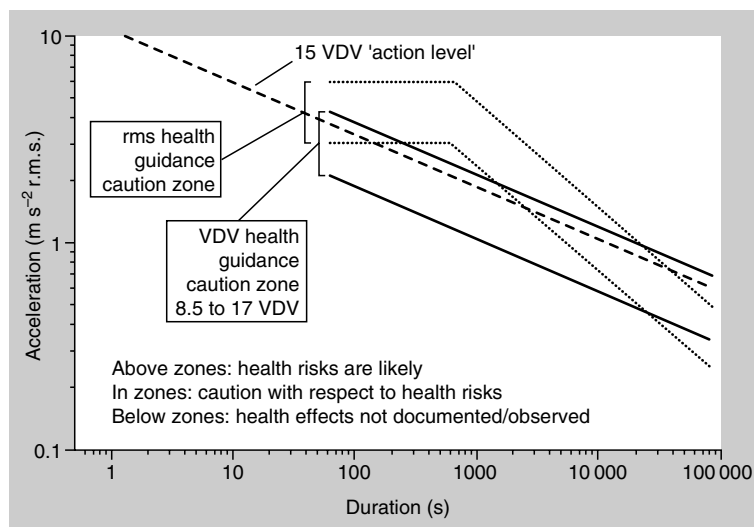


Figure 18.5 Comparison of the 'action level' corresponding to a vibration dose value (VDV) of $15 \text{ m s}^{-1.75}$ (in British Standard 6841) (BSI, 1987) and the r.m.s. and VDV 'health guidance caution zones' (in ISO 2631) (ISO, 1997).

10 min and then decreases in inverse proportion to the square root of duration from 10 min to 24 h. This method suggests a ‘r.m.s. health guidance caution zone’, but the method is not fully defined in the text; it allows very high accelerations at short durations, it conflicts dramatically with the vibration dose value method, and cannot be extended to durations below 1 min.

With severe vibration exposures, prior consideration of the fitness of the exposed persons and the design of adequate safety precautions may be required. The need for regular checks on the health of routinely exposed persons may also be considered.

Figure 18.5 compares the ‘VDV health guidance caution zone’, the ‘root-mean-square health guidance caution zone’ and the accelerations corresponding to the $15.0 \text{ m s}^{-1.75}$ ‘action level’ for exposure durations between 1 s and 24 h. Any exposure to continuous vibration, intermittent vibration or repeated shock may be compared with either the $15.0 \text{ m s}^{-1.75}$ ‘action level’ or the VDV ‘health guidance caution zone’ by calculating the vibration dose value. It would be unwise to exceed the appropriate action level without consideration of the possible health effects of an exposure to vibration or shock.

EU Machinery Safety Directive

The Machinery Safety Directive of the European Community (89/392/EEC) states that machinery must be designed and constructed so that hazards resulting from vibration produced by the machinery are reduced to the lowest practicable level, taking into account technical progress and the availability of means of reducing vibration. The instruction handbooks for hand-held and hand-guided machinery must specify the equivalent acceleration to which the hands or arms are subjected when this exceeds some stated value (for whole-body vibration this is a frequency-weighted acceleration of 0.5 m s^{-2} r.m.s. at present). The relevance of any such value will depend on the test conditions to be specified in other standards. Many work vehicles exceed this value at some stage during an operation or journey.

Standardized procedures for testing work vehicles are being prepared; the values quoted by manufacturers at present may not always be representative of the operating conditions in the work for which the machinery is used.

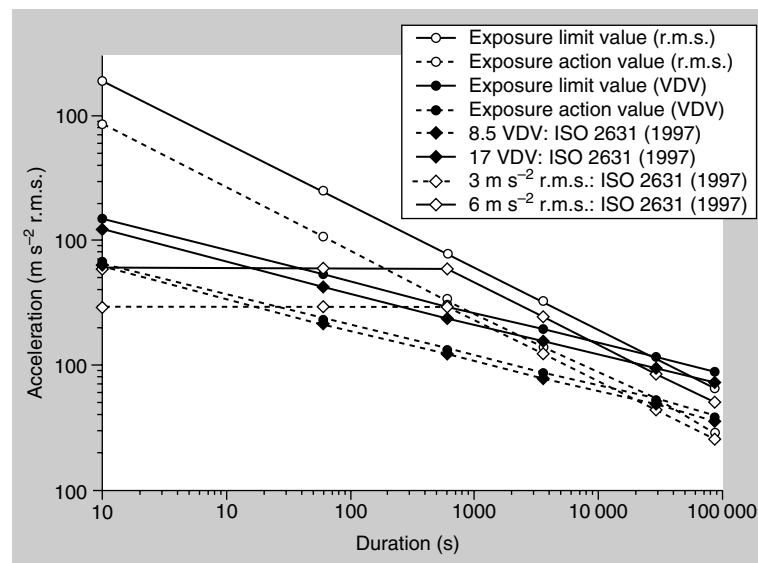
EU Physical Agents Directive (2002)

In 2002, the Parliament and Commission of the European Community agreed ‘minimum health and safety requirements’ for the exposure of workers to the risks arising from vibration. For whole-body vibration the Directive defines an 8-h equivalent ‘exposure action value’ of 0.5 m s^{-2} r.m.s. (or a vibration dose value of $9.1 \text{ m s}^{-1.75}$) and an 8-h equivalent ‘exposure limit value’ of 1.15 m s^{-2} r.m.s. (or a vibration dose value of $21 \text{ m s}^{-1.75}$). Member States of the European Union must bring into force laws to comply with the Directive by 6 July 2005.

The Directive says that workers shall not be exposed above the ‘exposure limit value’. If the ‘exposure action values’ are exceeded, the employer shall establish and implement a programme of technical and/or organizational measures intended to reduce to a minimum exposure to mechanical vibration and the attendant risks. The Directive says that workers exposed to mechanical vibration in excess of the ‘exposure action values’ shall be entitled to appropriate health surveillance, but health surveillance is not restricted to situations when the exposure action value is exceeded: health surveillance is required if there is any reason to suspect that workers may be injured by the vibration even if the ‘exposure action value’ is not exceeded.

The probability of injury arising from occupational exposures to whole-body vibration at the ‘exposure action value’ and the ‘exposure limit value’ cannot be estimated because epidemiological studies have not yet produced sufficient dose–effect relationships. However, it seems clear that the Directive does not define ‘safe exposures’ to whole-body vibration as the r.m.s. values are associated with extraordinarily high magnitudes of vibration (and shock) when the exposures are short: these exposures may be assumed to be haz-

Figure 18.6 Comparison of the 'health guidance caution zones' (in ISO 2631-1) (ISO, 1997); $3\text{--}6\text{ m s}^{-2}$ r.m.s. and $8.5\text{--}17\text{ m s}^{-1.75}$ with the 'exposure limit values' and 'exposure action values' for whole-body vibration (in the EU Physical Agents Directive, 2002) (The European Parliament and the Council of the European Union, 2002).



ardous (Fig. 18.6; see also Griffin, 2004). The vibration dose value procedure seems to suggest more reasonable vibration magnitudes for short-duration exposures.

Interference with activities by whole-body vibration

Vibration may interfere with the acquisition of information (e.g. by the eyes), the output of information (e.g. by hand or foot movements) or the complex central processes that relate input to output (e.g. learning, memory, decision-making) (Griffin, 1990). Effects of oscillatory motion on human performance may impair safety.

The greatest effects of whole-body vibration are on input processes (mainly vision) and output processes (mainly continuous hand control). In both cases there may be disturbance occurring entirely outside the body (e.g. vibration of a viewed display or vibration of a hand-held control), disturbance at the input or output (e.g. movement of the eye or hand) and disturbance affecting the peripheral nervous system (i.e. afferent or efferent system). Central processes may also be affected by vibration but understanding is currently too limited to make confident generalized statements.

The effects of vibration on vision and manual control are primarily caused by the movement of the affected part of the body (i.e. eye or hand). The effects may be decreased by reducing the transmission of vibration to the eye or to the hand, or by making the task less susceptible to disturbance (e.g. increasing the size of a display or reducing the sensitivity of a control). Often, the effects of vibration on vision and manual control can be much reduced by redesign of the task.

Simple cognitive tasks (e.g. simple reaction time) appear to be unaffected by vibration, other than by changes in arousal or motivation or by direct effects on input and output processes. This may also be true for some complex cognitive tasks. However, the scarcity and diversity of experimental studies allows the possibility of real and significant cognitive effects of vibration. Vibration may influence 'fatigue' but there is little relevant scientific evidence and none that supports the complex form of the so-called 'fatigue-decreased proficiency limit' offered in an old, and now withdrawn, version of International Standard ISO 2631.

Control of whole-body vibration

Wherever possible, reduction of vibration at source is to be preferred. This may involve

reducing the undulations of the terrain or reducing the speed of travel of vehicles.

Methods of reducing the transmission of vibration to operators require an understanding of the characteristics of the vibration environment and the route for the transmission of vibration to the body. For example, the magnitude of vibration often varies with location: lower magnitudes will be experienced in some areas. Table 18.9

Table 18.9 Summary of preventative measures to consider when persons are exposed to whole-body vibration (adapted from Griffin, 1990, Chapter 5).

<i>Group</i>	<i>Action</i>
Management	Seek technical advice
	Seek medical advice
	Warn exposed persons
	Train exposed persons
	Review exposure times
Machine manufacturers	Provide policy on removal from work
	Measure vibration
	Design to minimize whole-body vibration
	Optimize suspension design
	Optimize seating dynamics
	Ergonomic design to provide good posture, etc.
	Provide guidance on machine maintenance
	Provide guidance on seat maintenance
	Provide warning of dangerous vibration
	Measure vibration exposure
Provide appropriate machines	
Technical at workplace	Select seats with good attenuation
	Maintain machines
	Inform management
	Pre-employment screening
Medical	Routine medical checks
	Record all signs and reported symptoms
	Warn workers with predisposition
	Advise on consequences of exposure
	Inform management
Exposed persons	Use machine properly
	Avoid unnecessary vibration exposure
	Check that seat is properly adjusted
	Adopt good sitting posture
	Check condition of machine
	Inform supervisor of vibration problems
	Seek medical advice if symptoms appear
Inform employer of relevant disorders	

lists some preventative measures that may be considered.

Seats can be designed to attenuate vibration. However, most seats exhibit a resonance at low frequencies that results in higher magnitudes of vertical vibration occurring on the seat than on the floor! At high frequencies there is usually attenuation of vibration. In use, the resonance frequencies of common seats are in the region of 4 Hz. The amplification at resonance is partially determined by the ‘damping’ in the seat. Increases in the damping of the seat cushioning tend to reduce the amplification at resonance but increase the transmissibility at high frequencies. There are large variations in transmissibility between seats and these result in significant differences in the vibration experienced by people.

A simple numerical indication of the isolation efficiency of a seat for a specific application is provided by the ‘seat effective amplitude transmissibility’ (SEAT) (Griffin, 1990). A SEAT value greater than 100% indicates that, overall, the vibration on the seat is ‘worse’ than the vibration on the floor. Values below 100% indicate that the seat has provided some useful attenuation. Seats should be designed to have the lowest SEAT value compatible with other constraints.

A separate suspension mechanism is provided beneath the seat pan in ‘suspension seats’. These seats, which are used in some off-road vehicles, trucks and coaches, have low resonance frequencies (around 2 Hz) and so can attenuate vibration at frequencies above about 3 Hz. The transmissibilities of these seats are usually determined by the seat manufacturer, but their isolation efficiencies vary with operating conditions.

Conclusions

Mechanical oscillation of the human body (both hand-transmitted and whole-body vibration) can produce discomfort, interfere with the performance of activities and cause pathological and physiological changes in the body.

Those responsible for the health of workers should be aware of the relevant standards for the evaluation and assessment of hand-transmitted

vibration and whole-body vibration. The standards, guides and legislation provide valuable information and should not be ignored. Nevertheless, the complexity of the interactions between oscillatory motion (i.e. vibration and shock) and the functions of the body are great. It is helpful to be aware of the uneven scientific support for some of the information.

References

- Bovenzi, M., and Hulshof, C.T.J. (1998). An updated review of epidemiologic studies on the relationship between exposure to whole-body vibration and low back pain. *Journal of Sound and Vibration*, **215**, 595–611.
- Brammer, A.J., Taylor, W. and Lundborg, G. (1987). Sensorineural stages of the hand-arm vibration syndrome. *Scandinavian Journal of Work, Environment and Health*, **13**, 279–83.
- British Standards Institution (1987). Measurement and evaluation of human exposure to whole-body mechanical vibration and repeated shock. British Standards Institution BS 6841.
- British Standards Institution (1992). Evaluation of human exposure to vibration in buildings (1 Hz to 80 Hz). British Standards Institution BS 6472.
- Faculty of Occupational Medicine of the Royal College of Physicians, Working Party on Hand-transmitted Vibration (1993). Hand-transmitted vibration: clinical effects and pathophysiology, Part 1. Report of a working party, The Royal College of Physicians of London.
- Gemne, G., Pyykko, I., Taylor, W. and Pelmear, P. (1987). The Stockholm Workshop scale for the classification of cold-induced Raynaud's phenomenon in the hand-arm vibration syndrome (revision of the Taylor-Pelmear scale). *Scandinavian Journal of Work, Environment and Health*, **13**, 275–8.
- Griffin, M.J. (1990). *Handbook of Human Vibration*. Academic Press, London.
- Griffin, M.J. (1998). A comparison of standardized methods for predicting the hazards of whole-body vibration and repeated shocks. *Journal of Sound and Vibration*, **215**, 883–914.
- Griffin, M.J. (2004). Minimum health and safety requirements for workers exposed to hand-transmitted vibration and whole-body vibration in the European Union: a review. *Occupational and Environmental Medicine*, **61**, 387–97.
- Griffin, M.J., Bovenzi, M. (2002) The diagnosis of disorders caused by hand-transmitted vibration: Southampton Workshop 2000. *International Archives of Occupational and Environmental Health*, **75**, 1–5.
- Griffin, M.J., Bovenzi, M. and Nelson, C.M. (2002) Dose-response patterns for vibration-induced white finger. *Journal of Occupational and Environmental Medicine*, **63**, 16–26.
- Health and Safety Executive (1994) *Hand-arm Vibration*. Health and Safety Executive, HS(G) 88.
- International Organization for Standardization (1988). *Hand-held Portable Tools – Measurement of Vibration at the Handle – Part 1: General*. International Standard ISO 8662–1.
- International Organization for Standardization (1989). *Evaluation of Human Exposure to Whole-body Vibration – Part 2: Continuous and Shock-induced Vibration in Buildings*. International Standard ISO 2631–2.
- International Organization for Standardization (1997). *Mechanical Vibration and Shock – Evaluation of Human Exposure to Whole-body Vibration. Part 1: General Requirements*. International Standard ISO 2631–1.
- International Organization for Standardization (2001). *Mechanical Vibration – Measurement and Evaluation of Human Exposure to Hand-transmitted Vibration – Part 1: General Requirements*. International Standard ISO 5349–1:2001(E).
- Loriga, G. (1911). Il lavoro con i martelli pneumatici. [The use of pneumatic hammers.] *Bollettino Ispett. Lavoro*, **2**, 35–60.
- Palmer, K.T., Coggon, D.N., Bednall, H.E., Pannett, B., Griffin, M.J. and Haward, B. (1999). Whole-body vibration: occupational exposures and their health effects in Great Britain. Health and Safety Executive Contract Research Report 233/1999. HSE Books, London.
- Taylor, W., Pelmear, P.L. and Pearson, J. (1974) Raynaud's phenomenon in forestry chain saw operators. In *The Vibration Syndrome, Proceedings of a Conference on the Medical Engineering and Legal Aspects of Hand-Arm Vibration* (ed. W. Taylor), The University of Dundee, 12–14 July 1972. Academic Press, London.
- The European Parliament and the Council of the European Union (2002). *On the Minimum Health and Safety Requirements Regarding the Exposure of Workers to the Risks Arising from Physical Agents (Vibration)*. Directive 2002/44/EC; Official Journal of the European Communities, 6 July 2002; L177/13–19.

Chapter 19

Light and lighting

N. Alan Smith

Introduction	Lighting systems and lighting design
Fundamentals	Lighting systems
The electromagnetic spectrum, visible radiation and light	Visual task lighting
Infrared and ultraviolet radiation	Lumen method of lighting design
Terms and definitions	Inspection lighting
The eye and vision	Emergency lighting
Construction of the eye	Exterior lighting
Sensitivity of the eye	Lighting for areas containing visual display units
Retinal fatigue	General lighting requirements
Visual perception	Reflections and glare
Characteristics of vision	Luminaires for VDU areas
Accommodation	Use of uplighters
Adaptation	Lighting surveys and survey techniques
Acuity	Preliminaries
Visual fatigue	Preliminary report sheet
The visual task	Determination of the minimum number of measuring points
The visual environment	Measuring equipment
Glare	Illuminance meters
Glare index	Luminance meters
Lamps and luminaires	Interpretation of data
Filament lamps	Typical illuminance levels
Discharge lamps	References
Luminaires	Suggested further reading
Daylight	

Introduction

Light is a natural phenomenon required to enable everyday activities to be carried out. It is essential for our basic existence yet, ironically, it is very often taken for granted. Modern lifestyles are such that humans have become less dependent upon natural light since the introduction of artificial light sources.

The role of light and lighting in the field of occupational hygiene has become progressively more important and the necessity to overcome problems associated with ill-conceived lighting installations is more readily apparent.

This chapter gives the practising occupational hygienist an insight into the basic concepts of

light and lighting with the underlying emphasis on its application to the workplace. It is not the purpose of the chapter to introduce the reader to in-depth detail of the science of illumination; this can be found in specialist reference texts.

Fundamentals

The electromagnetic spectrum, visible radiation and light

Light is a form of energy. This form of energy is known as 'radiation' and is electromagnetic in character. This means that the radiation has both an electric and a magnetic field, both of which vary

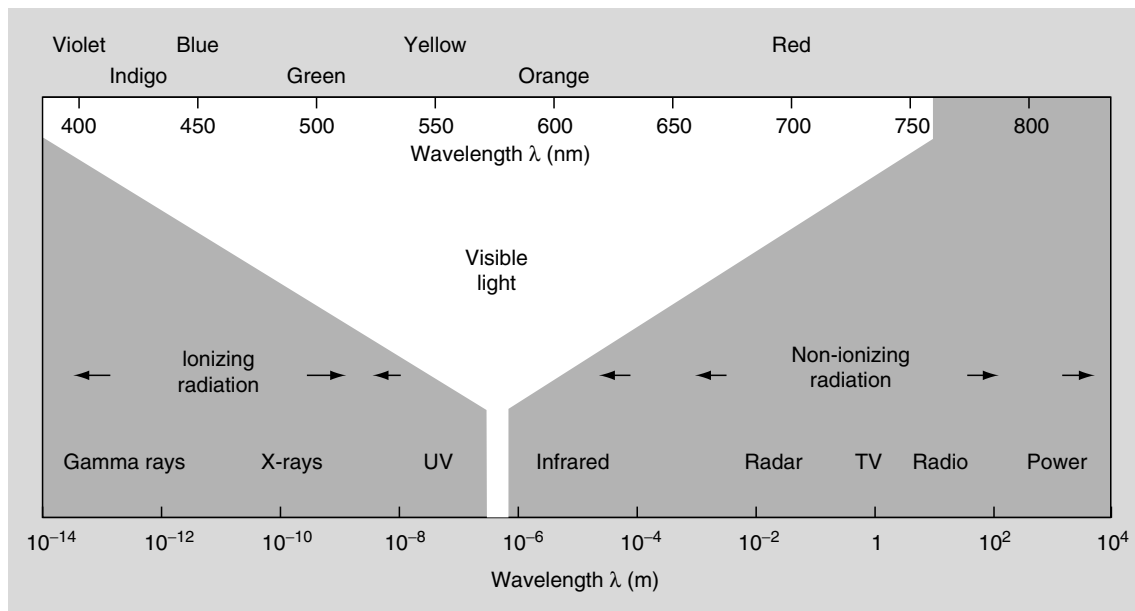


Figure 19.1 The electromagnetic spectrum.

sinusoidally with time. All electromagnetic radiation propagates with a velocity of 3×10^8 m s⁻¹. The electromagnetic spectrum is shown in Fig. 19.1 and it can be seen that the visible spectrum occupies the approximate wavelength range between 380 and 760 nm. In essence, radiation is the cause and light is the effect, with changes in wavelength within the visible spectrum producing changes in the perceived colour of the light output.

Infrared and ultraviolet radiation

At wavelengths at the extremities of the visible part of the electromagnetic spectrum, the radiation becomes progressively infrared (IR) at the long-wavelength end and ultraviolet (UV) at the short-wavelength end. Both IR and UV radiation have been subdivided into three groups: A, B and C.

Terms and definitions

There are many terms and definitions associated with lighting of which the following is an abridged list.

- *Candela*. The SI basic unit of luminous intensity (cd).
- *Colour rendering index (CRI)*. A measure of how colours under a particular light source compare with a standardized set of conditions.
- *Daylight factor*. The ratio of illuminance due to daylight at a particular point in a building to the simultaneous horizontal external illuminance from an unobstructed sky, expressed as a percentage. Sunlight is excluded.
- *Glare index*. A mathematical value assigned to the degree of discomfort glare.
- *Illuminance*. The quantity of luminous flux falling on a surface divided by the area upon which it is falling, measured in lumen per square metre (lm m⁻²) or lux.
- *Lumen*. The SI-derived unit of luminous flux (lm).
- *Luminaire*. The apparatus that controls the distribution of light and contains all the components for fixing, protecting and connecting the lamp.
- *Luminance*. The flow of light in a given direction (measured in candela per square metre, cd m⁻²) from a surface element.

- *Luminance contrast*. A measure of contrast as a ratio of luminance difference (task to background) to luminance of background.
- *Luminous efficacy*. The ratio of luminous flux produced by a light source to the electrical power consumed by the source in order to produce the luminous output, measured in lumen per watt (lm W^{-1}).
- *Luminous flux*. The light emitted by a source or received by a surface.
- *Luminous intensity*. A measure of the luminous flux emitted within a small conical angle in the direction of a surface, measured in candela.
- *Maintenance factor*. A factor which takes into account the reduction in output from a luminaire due, among other things, to: (1) depreciation in output from a lamp due to ageing; and (2) cleanliness of luminaire optical system.
- *Spacing-to-height ratio (SHR)*. The ratio of spacing between centres of adjacent luminaires to the mounting height above the working plane.
- *Utilization factor*. The ratio of the amount of luminous flux falling on the working plane to the total flux emitted by the luminaires.
- *Working plane*. The horizontal, vertical or inclined plane on which the task lies, normally assumed to be a horizontal plane 0.85 m above floor level, unless specified otherwise.

The eye and vision

Construction of the eye

The eye, a cross-section of which is shown in Fig. 19.2, is effectively an instrument that collects light rays and subsequently focuses them into an image on its rear surface.

Initially, light enters the eye through the cornea, which by virtue of its rounded shape behaves like a convex lens. Behind the cornea is the iris, a coloured annular structure that opens and closes analogous to the diaphragm of a camera in order to control the amount of light entering the eye. Within the centre of the iris is the pupil. Light having passed through the cornea passes through the iris and then into a transparent body called the lens. The shape of this lens is variable and can be changed by the ciliary muscles.

The ciliary muscles control the lens in order to accurately focus the light entering the eye on the retina. The retina is a light-sensitive layer at the back of the eye. Photosensitive cells within the retina convert the light falling on the retina into signals, which are carried via the optic nerve to the brain. The photosensitive cells are subdivided into cones and rods.

The cones are active at high levels of illuminance and allow us to see detailed colour in daylight. This

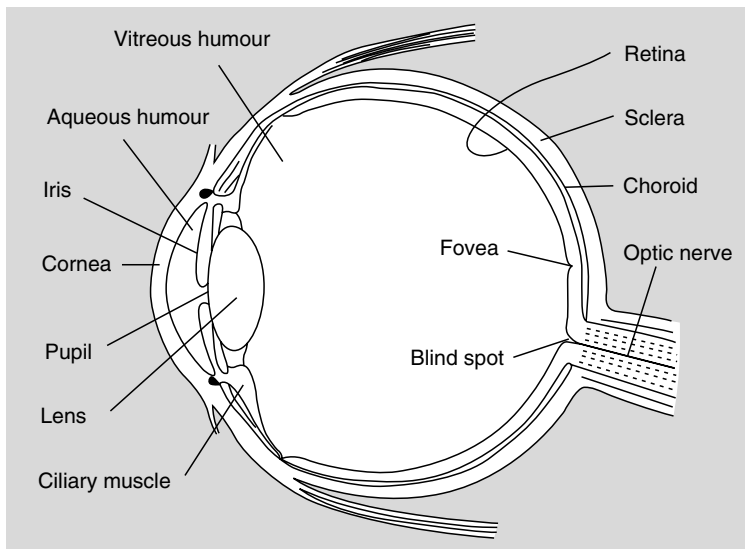


Figure 19.2 Cross-section through the eye.

type of vision is referred to as 'photopic'. At lower illuminance levels, as might be typified under road lighting, the cones become progressively less sensitive, with a simultaneous lessening in appreciation of colours. Such vision is referred to as 'mesopic'. At even lower levels of illuminance, typified by approaching darkness, the cones become insensitive and only the rods operate leading to a 'grey' vision often referred to as 'scotopic' vision.

Essentially, there are three types of cone, sensitive to red, blue and green light. The signals generated in the cones are subsequently transmitted, via the optic nerves, to the brain, which interprets the signals. Any variation in either the spectral distribution of the light source, or the colour-sensitive elements of the eye, will influence the final sensation of colour. In the centre of the retina is a small dimple termed the fovea, which contains only cones. This concentration of cones makes the fovea the centre of the eye's sharpest vision.

The aqueous humour is a water-like fluid that washes the front of the eye in the space between the cornea and the lens. The vitreous humour is a transparent jelly-like fluid that occupies the interior of the eye and helps the eye to keep its shape.

Sensitivity of the eye

The sensitivity of the eye is not constant over all wavelengths within the visible spectrum. The sensitivity, for the light-adapted eye, is greatest at a wavelength of 555 nm and diminishes at the extremes of the visible spectrum as shown in Fig. 19.3. If lights of different colours but of the same intensity are directed into the eyes of an observer, the colours at the middle of the visible spectrum will appear brighter than those at the extremes. Thus, light sources whose output is within the yellow-orange region of the spectrum (e.g. low-pressure sodium lamps with a monochromatic output at 589 nm) will have a greater efficacy than light sources whose output is more biased towards the red or blue ends of the visible spectrum.

Retinal fatigue

If the eye concentrates for a short time on an intense source of red light and then the gaze is transferred very quickly to a sheet of white paper,

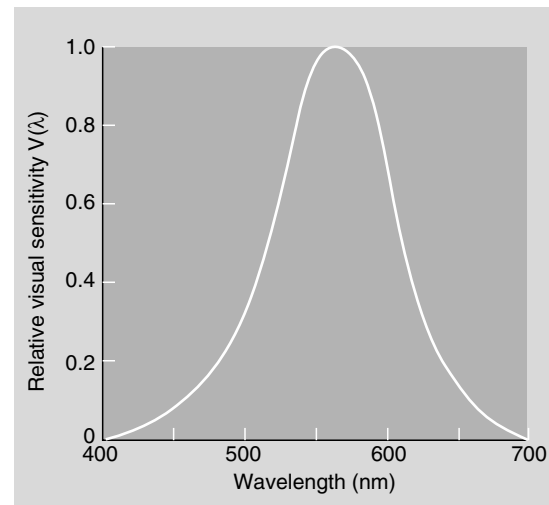


Figure 19.3 Sensitivity of the eye.

the observer perceives a blue/green image of the original source. On looking at the white background, the complementary colour of the original source will be detected. The nerves of the retina that detected the original red source of light have become 'fatigued'. When the observer transfers his or her gaze to the white paper, all the nerves of the retina are excited but, as in that part of the retina where the image of the red source fell those nerves are fatigued, this area of the retina will record a peacock-blue colour.

Visual perception

Visual perception can be thought of as a sequence of processes:

- the emission of light from a source or, alternatively, the reflection of light from an object;
- the radiant energy reaching the eye;
- the transfer of radiant energy in the retina into electrochemical signals for onward transmission to the brain;
- the creation in the brain, following receipt of the incoming signals, of an image that is a facsimile of the scene originally viewed.

Memory has a significant role to play in the process of interpretation. The information received on the retina is compared with similar recalled experiences. Comparison with previous

similar visual experiences, and their confirmation using evidence of other senses, leads to the brain linking the information received on the retina with the real world.

Characteristics of vision

There are three highly significant factors that influence the eye's ability to see: accommodation, adaptation and acuity.

Accommodation

Accommodation is the ability of the eye to focus on an object: a process that involves two separate and automatic operations, i.e.:

- 1 the adjustment of the lens so that the image subsequently formed on the retina is sharp;
- 2 the convergence of the signal received from each eye so that there is one 'real image' in the brain.

Figure 19.4 shows the two operations.

Adaptation

The eye will function over a brightness range of 1 000 000:1, but it is only capable of coping with brightness ranges of typically 1000:1 simultaneously. The eye responds to the brightness range

by varying its sensitivity to the brightness of the object being viewed. This change in the eye's sensitivity is termed *adaptation*. It does not occur instantaneously. In changing from a low brightness to a high brightness, the adaptation typically takes place in seconds, whereas in traversing from a high brightness to a low brightness the change may take several minutes.

Under normal daylight conditions, the sensitivity of the eye peaks at a wavelength of 555 nm. When the eye is adapted to darkened conditions the sensitivity of the eye peaks at a wavelength of 505 nm. This shift is referred to as the Purkinjé shift. The relationship between relative sensitivity and wavelength for both the light- and dark-adapted eye is shown in Fig. 19.5.

Acuity

Visual acuity is the ability to discern detail. It is influenced by the luminance of the object being viewed. The Snellen chart is used by optometrists and enables visual acuity to be determined using a combination of letter size and distance between the chart and the observer. Figure 19.6 shows a typical Snellen chart.

If the variation in acuity with task illuminance is plotted graphically, it will be shown that acuity

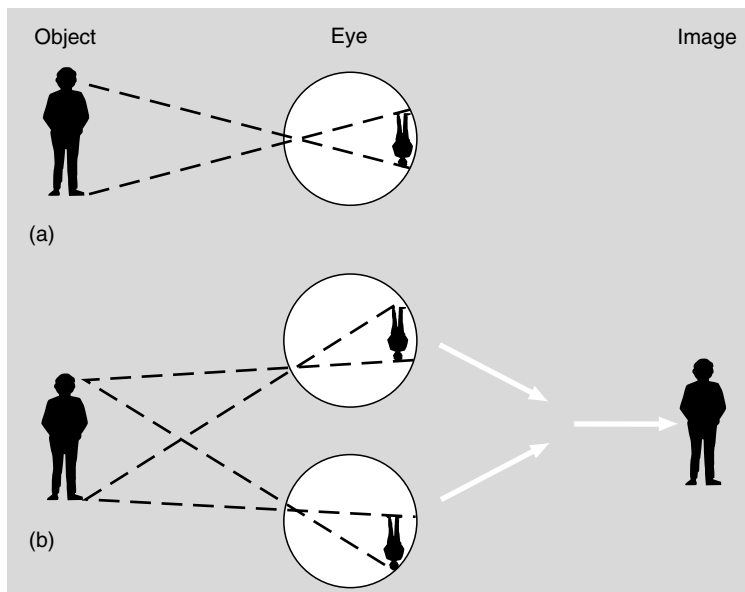


Figure 19.4 Accommodation.

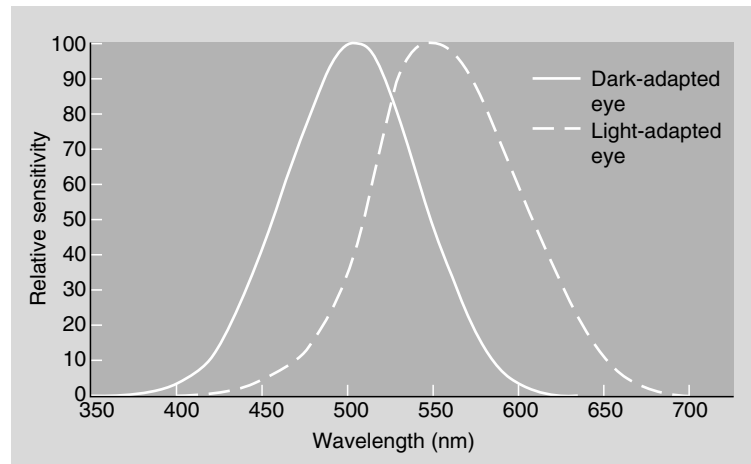


Figure 19.5 Variation in spectral sensitivity of the human eye.

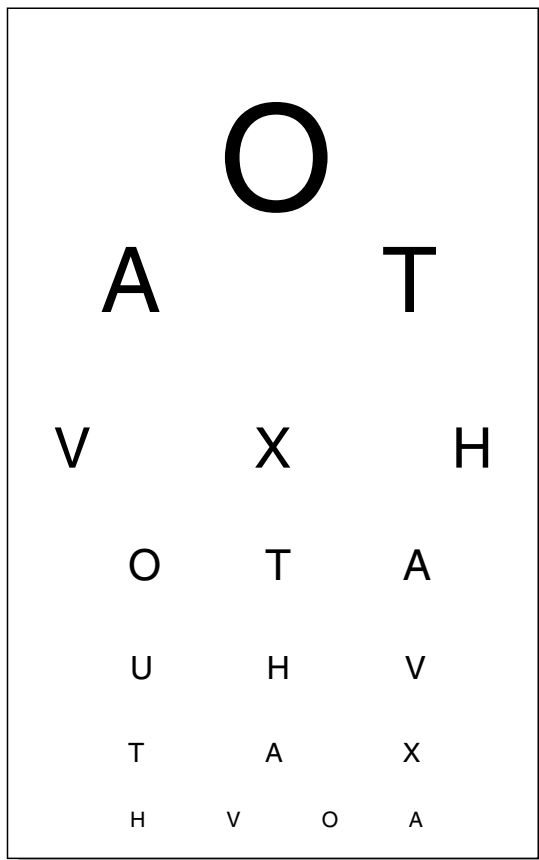


Figure 19.6 Typical Snellen chart.

initially increases rapidly, the initial rate of change decreasing until a point is reached at which the characteristic levels off. Figure 19.7 is a typical curve, relating to a particular task. If the task is a relatively simple one, then maximum performance (the point at which the characteristic levels off) will be achieved with a relatively low level of illuminance. It will be apparent that an excess of illuminance may ultimately lead to the development of glare.

By definition, acuity tends to be an academic indicator of the eye's ability to discern detail.

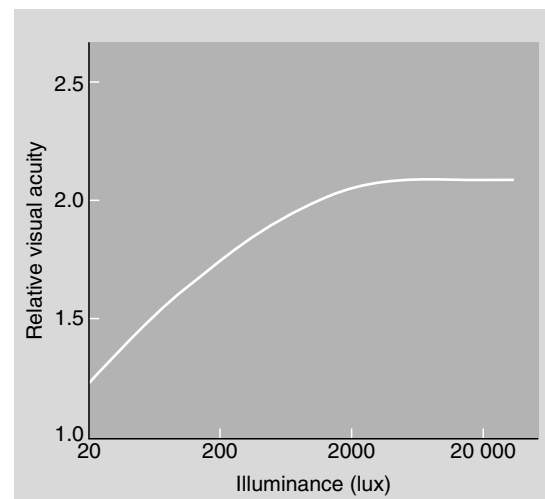


Figure 19.7 Typical relationship between visual acuity and prevailing illuminance.

When a visual task is under consideration it is necessary to determine not only whether the task can be undertaken, but also the ability of the eye to carry out the task with speed and accuracy, a combination referred to as *visual performance*.

There is a link between accuracy and speed and it can be shown that the plot of performance against illuminance is very similar to the plot of acuity against illuminance. Although there will be variation from task to task, these characteristics tend to level off at the point where the individual obtains maximum accuracy corresponding to the point where manual dexterity will not allow work at a faster rate. Figure 19.8 shows the typical relationship of speed and accuracy against illuminance. Figure 19.9 gives typical relative performance curves for three tasks having low, medium and good contrast. It will be evident that the low-contrast task requires a much greater illuminance in order for it to be performed adequately. Figure 19.10 shows the relationship between typical relative illuminance required for reading print on paper, as an example of a visual task, against the age of the individual reading the print.

Visual fatigue

The causes of visual fatigue can be conveniently divided into three categories:

- 1 *constitutional* – where the overall state of the health of the individual is important;
- 2 *ocular* – where the effects of deteriorating vision and/or the effects of ageing are significant;
- 3 *environmental* – where the levels of illuminance and the general characteristics of the immediate vicinity are influential. These environmental causes of visual fatigue can be subdivided into:
 - (a) those that are influenced by the visual task itself; and
 - (b) those that are influenced by the visual environment in which the task is carried out.

The visual task

When the light signal detected by the eyes is of low level, the brain attempts to amplify the signal by feedback to the eye. This process subsequently produces strain when the feedback becomes continuous. Characteristics of the visual task that subsequently lead to eyestrain include: minute detail, excessively low-contrast task/background, movement of task and surface finish of task.

The visual environment

Characteristics of the visual environment, which, either singularly or in combination, subsequently lead to eyestrain, include inadequate illuminance,

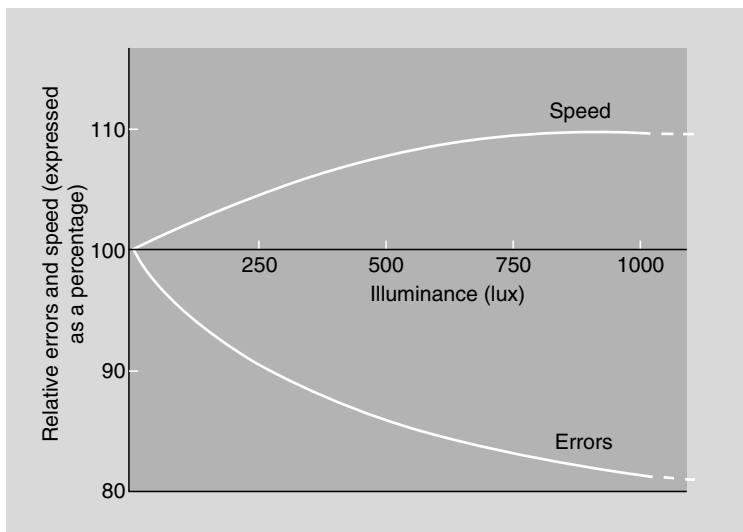


Figure 19.8 Typical relationship between speed and accuracy and prevailing illuminance.

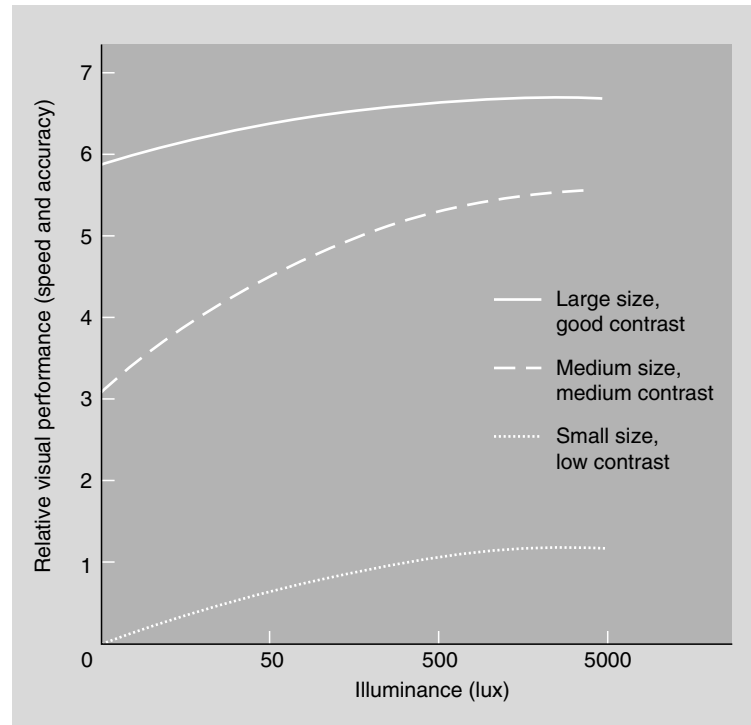


Figure 19.9 Typical relative performance versus illuminance for different tasks and contrast levels.

excessively high contrast (task/background), presence of glare, flicker from fluorescent sources and general feeling of lack of well-being within an environment.

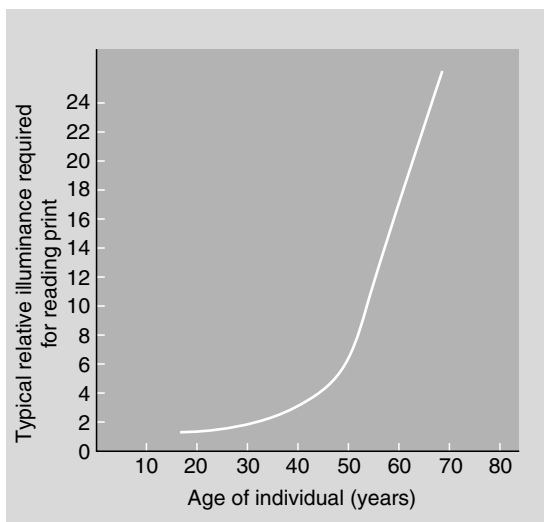


Figure 19.10 Typical relative illuminance required for reading print versus age of individual.

Glare

Several definitions of the term 'glare' have been put forward. The term is synonymous with the effect perceived when viewing the headlights of an oncoming vehicle, being more pronounced on a dark, wet night. In reality, glare is any situation where contrast within the field of view is excessive. The two main forms of glare are 'disability glare' and 'discomfort glare'.

Disability glare, as the name would imply, prevents or disables an individual from seeing a particular task. The previous reference to the headlights of an oncoming vehicle is an example of disability glare. In such situations it is almost impossible to discern the scene immediately surrounding the headlights. The disabling effect within the eye is due in part to a veil of scattered

light in the optic media within the eye and is likely to be more prevalent in elderly people. Disability glare is proportional to the intensity of the offending source.

Discomfort glare, which is more likely to occur in interiors, may not disable an individual from performing a particular visual task, but prolonged exposure to such an environment will cause discomfort. It may be that such discomfort will take several hours to materialize, often developing as a headache. Discomfort glare can be produced as a result of a badly designed lighting installation. Almost invariably, glare is an unwanted phenomenon, but occasionally it is possible to use glare to advantage. One such example is the use of high-powered lamps mounted on security checkpoints at factory entrances. In such circumstances would-be intruders may not be able to see if the checkpoint is manned.

Glare index

It is possible to assign a numerical value to the magnitude of discomfort glare. This value is known as the 'glare index'. Such values are then compared with limiting values of indices for typical interiors as specified in the Code for Lighting [1].

Lamps and luminaires

Lamps can be divided into two types: filament lamps and discharge lamps.

Filament lamps

These lamps rely on the principle of incandescence for the production of light. Essentially, a tungsten filament is heated by means of an electric current until it reaches incandescence, giving off visible (and other) radiation. The domestic lamp, often referred to as a general lighting service (GLS) lamp, is an everyday example of the filament lamp.

A progression from the GLS lamp is the tungsten-halogen lamp. Quartz glass is used for the envelope and the addition of a trace of one of the halogen elements, e.g. iodine, allows the iodine to

combine with the tungsten that has evaporated from the filament, to form tungsten-iodine. The tungsten-iodine vapour is then carried back to the filament by convection currents, where it separates out, the tungsten being redeposited on the original filament. The iodine is subsequently released and the cycle is repeated. The envelope of a tungsten halogen lamp should never be touched by human skin, as the greases and acids contained within the skin will attack the quartz glass and subsequently produce weak spots that may ultimately lead to premature failure.

Discharge lamps

A discharge lamp consists essentially of a discharge tube containing a gas or vapour.

When a voltage is applied to the lamp, energy is imparted to atoms resulting in the displacement of electrons to higher energy levels. The electrons subsequently fall back to their original levels, thereby releasing energy in the form of electromagnetic photons. The nature of the gas or vapour will determine the wavelength and thereby the perceived colour of the output from the lamp.

Discharge lamps can be further subdivided into low- and high-pressure lamps. Low-pressure lamps include fluorescent lamps (low-pressure mercury) and low-pressure sodium lamps. The latter emit monochromatic light at a wavelength of approximately 589 nm. These lamps have the advantage of almost instantaneous restrike in the event of momentary loss of electrical supply. The monochromatic output of the low-pressure sodium lamp is at a wavelength close to that of the maximum sensitivity of the eye and therefore the lamp has a very high luminous efficacy value. A further advantage is that low-pressure sodium lamps are extremely useful in foggy conditions as the water droplets in suspension in the atmosphere cannot readily disperse the monochromatic light.

A major disadvantage of the low-pressure sodium lamp is that its colour rendering index value is almost zero. Objects seen in the light from such a lamp will have their surface colours severely distorted.

Fluorescent lamps are used extensively in interiors. Electronic circuitry involving high-frequency

supplies allows regulation of the light output, which, if used in sympathy with levels of prevailing daylight, can be used in energy-saving installations. The ambient temperature influences the light output from fluorescent lamps.

High-pressure lamps include mercury vapour, metal halide and high-pressure sodium lamps. Such lamps are used extensively for road lighting, lighting for civic and amenity areas together with floodlighting for car parks, railway sidings, building sites and some sports grounds.

Mercury vapour and metal halide lamps, if broken, emit potentially dangerous ultraviolet radiation. Disposal of broken and spent lamps needs to be carefully controlled. Lamp disposal should be carried out following consultation with the relevant authorities [2]. Table 19.1 gives typical characteristics and applications of lamp types.

Table 19.2 gives details of typical colour rendering index values of lamps.

Luminaires

Formerly known as the 'light fitting', the luminaire supports the lamp and provides the necessary electrical connections. Luminaires control the flow of light, direct it towards the working plane and control the brightness of the output of the luminaire visible to the occupants within the interior. Additionally, the luminaire provides the means of fixing the lamp to the building fabric, together with providing a housing for the lamp control gear.

Luminaires have to operate in a variety of environments, including dusty, wet, corrosive and the generally hostile. The Ingress Protection (IP) system in BS4533 [4] for specifying the degree of protection, which is provided by an enclosure, also includes luminaires. A two-digit reference number designates the degree of protection. The first digit (0–6 inclusive) describes the degree of protection from the ingress of solid foreign objects. The second digit (0–8 inclusive) describes the degree of protection against the ingress of water.

Daylight

In the UK, the exterior illuminance due to daylight typically reaches 35 000 lux at noon during July, whereas at noon in December the exterior illuminance typically reaches 8000 lux. When the

Table 19.2 Typical colour rendering index (CRI) values of lamps.

<i>Lamp type</i>	<i>CRI index value</i>
General Lighting Service (GLS)	90
Tungsten halogen	90
Fluorescent	70–95
High-pressure mercury	40–65
Metal halide	50–85
High-pressure sodium	40–75
Low-pressure sodium	0+*

*The CRI value of low-pressure sodium lamps approaches, but is not equal to, zero.

Table 19.1 Characteristics and applications of lamp types.

<i>Lamp type (and ILCOS* symbol)</i>	<i>Typical lamp efficacy (lumens per watt)</i>	<i>Typical lamp life (h)</i>	<i>Typical applications</i>
Tungsten filament (I)	8–18	1000–2000	Domestic, display
Tungsten halogen (HS)	18–24	2000–4000	Display, traffic signals, overhead projectors (OHP)
Mercury vapour (QE)	40–60	5000–10 000	Industrial, road lighting
Metal halide (M)	65–85	5000–10 000	Floodlighting, area and amenity lighting
Fluorescent:(FD) tubular,(FS) compact	50–100	5000–10 000	General, domestic, commercial
Low-pressure sodium (LS)	100–175	6000–12 000	Road lighting
High-pressure sodium (S)	65–120	6000–20 000	Industrial, road lighting, civic and amenity lighting

*ILCOS, International Lamp Coding System [3].

exterior illuminance falls below 5000 lux, daylighting is generally accepted as being too weak to provide adequate lighting within an interior. The corresponding interior illuminance due to daylight is significantly lower, with typically less than 10% of the exterior illuminance filtering through to interiors. Furthermore, this figure is variable and is influenced, for example, by distance from windows.

The relationship connecting values of internal illuminance and external illuminance, both values being restricted to daylight as a source, is referred to as the daylight factor (DF), where:

$$\text{Daylight factor} = \frac{\text{daylight illuminance at point within a room}}{\text{simultaneous illuminance on a horizontal plane outside the building from a completely unobstructed sky (excluding sunlight)}} \quad (19.1)$$

The daylight factor is essentially a geometrical characteristic of an interior/window combination. Its value is not influenced by changes in external illuminance and, at any point in an interior, its numerical value is constant.

The daylight factor calculation is relative to an individual point within an interior. In order to develop a more meaningful pattern of the effects of daylight within an interior it would be necessary to develop a grid of such values throughout the building. The concept of an average daylight factor overcomes the necessity to perform multiple calculations from point to point.

Lighting systems and lighting design

Lighting systems

Systems used in commercial and industrial interiors can be divided into three major groups: general lighting, localized lighting and local lighting. General lighting installations aim to provide, as far as is practical, an approximately uniform illuminance over the whole of the working plane. Near-uniform illuminance is often achieved using the

Lumen method of design (see below). Localized lighting is a system designed to provide the required illuminance on the work areas, together with a reduced level of illuminance in adjacent areas. Local lighting is lighting for a small area surrounding the task, typically provided by small fluorescent luminaires. Figure 19.11 shows the essential differences between the three systems. The Chartered Institution of Building Services Engineers (CIBSE) *Lighting Guide 7: Offices* 2003 [5] is a useful reference for office lighting applications.

Visual task lighting

When analysing particular tasks an assessment should be made of the adaptation and accommodation involved in the task at hand, together with a further assessment of the frequency at which the task is being performed. Determining the objects that require viewing and establishing under what conditions viewing needs to take place, together with a detailed survey of the existing lighting outlining any deficiencies, will be highly beneficial in forming an overall view of task lighting requirements.

Many methods of design for lighting systems aim to achieve a relatively uniform distribution of illuminance at the working plane level. Unfortunately, this often simultaneously produces unacceptable visual task lighting. When visual task lighting is considered, and good task lighting conditions ultimately achieved, it is likely that a more acceptable working environment will be established which avoids the effects of visual distraction and/or visual fatigue. Relatively bright reflections within the task will limit its visibility and may subsequently lead to discomfort.

The human eye acknowledges details within a visual task by discriminating between the darker and lighter parts of the task. The variation in 'brightness' of a visual task is determined from the luminance contrast. Luminance contrast (C) is given by:

$$C = \left| \frac{L_t - L_b}{L_b} \right| \quad (19.2)$$

where L_t is the luminance of the task and L_b is the luminance of the background with both quantities

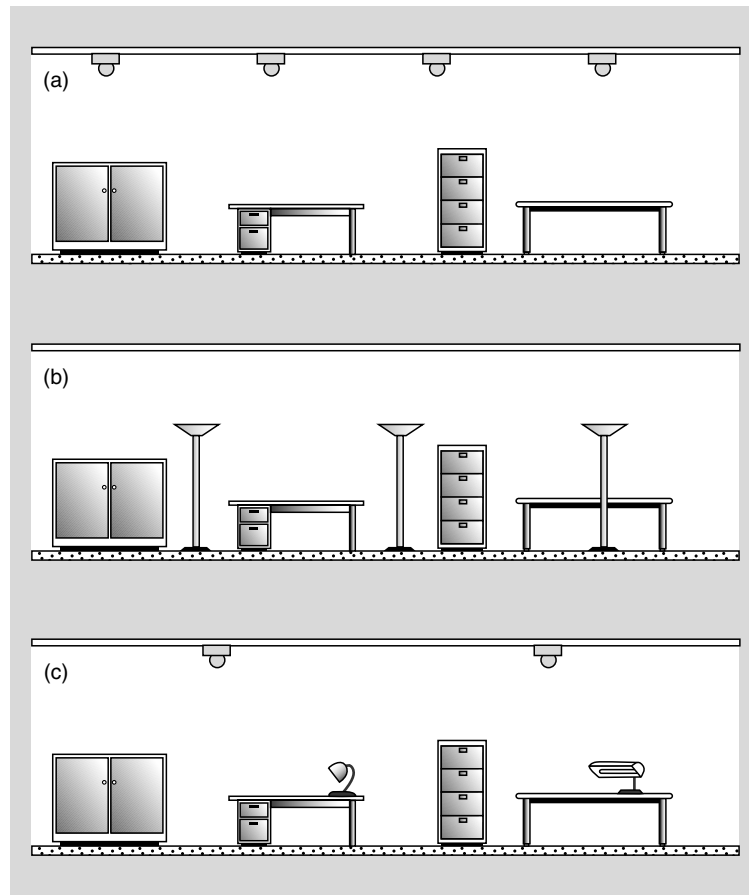


Figure 19.11 Types of lighting systems: (a) general lighting; (b) localized lighting; (c) local lighting.

expressed in candelas per square metre (cd m^{-2}). The vertical (modulus) lines indicate that all values of luminance contrast are to be considered as positive.

The contrast of a visual task will be influenced by the reflectance properties of the task itself. If the task material has a matt finish, then incident light upon the task will be reflected equally in all directions. This implies that the direction of the incident light will be insignificant.

It is more often the case, however, that the task has a non-matt or specular (mirror-like) finish. In such cases, defocused images of high-luminance sources, e.g. luminaires, will be seen reflected in the task. Such images produce ‘veiling reflections’, a form of indirect glare. ‘Veiling reflections’ are so termed because they produce a veil of light in front of the task.

Reflecting glare fools the eye into thinking that the environment is brighter than it is. This results in the eye failing to adapt correctly to the value of illuminance required for the visual task and leads to visual fatigue.

Lumen method of lighting design

The main aim of the lumen method of design is to achieve an average general level of illuminance on a working plane within an interior. This method takes no account of the task(s) likely to be performed in the interior.

With the lumen method, the illuminance is calculated from:

$$\begin{aligned} \text{Illuminance } E \text{ (in lux)} \\ &= \text{luminous flux (in lumens)} \\ &\quad \times \text{MF} \times \text{UF} / \text{area (in square metres)} \end{aligned} \quad (19.3)$$

where MF = maintenance factor and UF = utilization factor.

Inspection lighting

The purpose of inspection lighting is to highlight inconsistencies in a product so that it can be rejected if the defect is considered unacceptable. The techniques used are shown in Table 19.3.

Emergency lighting

The purpose of emergency lighting is to enable people to move in relative safety to an escape route, and subsequently to vacate the premises in the event of an interruption of supply to the main lighting system. Emergency lighting may be in use at all times, known as a 'maintained system', or it may come into operation only in the event of the failure of the main electrical supply. Such a system is referred to as 'non-maintained'. BS 5266 [7] relates to emergency lighting.

Exterior lighting

It is essential to appreciate that exterior lighting and interior lighting present different problems. Such differences include reflectance, size and mounting height. Reflectances are critical when considering interior lighting but with exterior

lighting there is an almost total absence of reflectances. Interiors are relatively small compared with the size of exteriors to be lit. The mounting heights in interiors tend to be low compared with exteriors, where often poles or towers are used.

The lighting of exteriors typically involves providing illumination for a small number of persons, often occupied in carrying out work of little visual difficulty. Such installations require a lower level of illuminance than is required for interior lighting. Typical examples of exterior lighting installations include: building and civil engineering sites; car parks; factory yards; railway yards and sidings; loading bays; gantry and crane yards; and storage areas.

For railway yards, sidings, factory yards and car parks, it is usual to use high-powered lamps at high mounting heights. On building and civil engineering sites, lighting is often required 24 h per day. On these sites it is usual to use portable lighting equipment that can be relocated as the site work progresses. When locating luminaires in loading bays, it is necessary to take account of the likely positions of vehicles during loading, in order to avoid the vehicles becoming obstructions and creating unwanted shadows. Storage areas can cause special problems. For areas free from obstruction it is more advantageous to use multiple luminaires on a single support. Increasing the mounting height can reduce unwanted shadows caused by obstructions.

Table 19.3 Inspection lighting techniques.

<i>Inspection</i>	<i>Technique</i>
Scratches on polished or glossy surfaces	Directional light applied to surface. Reflections from irregularities appear light on dark background
Surface flatness	Monochromatic light (typically from low-pressure sodium lamps) used in conjunction with optical flats creates optical fringes revealing defects
Defects in transparent materials, e.g. glass, plastics	Output from lamp is deliberately polarized, transmitted through the product being inspected and then analysed using a second polarizer. Defects will produce variations in the pattern of transmitted light
Surface finishes	Ultraviolet radiation causes some materials to fluoresce. Products to be inspected are coated with fluorescent material, which will reveal darkened areas where there are irregularities or discontinuities in the surface finish
Rotating components	Rotating components appear stationary using the stroboscopic effect
Colour matching*	Special fluorescent lamps in accordance with BS 950 [6]

*Care should be taken to avoid confusion with 'metamerism', i.e. when the colours of two articles appear identical under one light source but appear different under another source of light.

However, it will then be necessary to use higher rated lamps. For gantry and crane yards, structural members supporting the gantries can be used for the mounting of luminaires. Attention must be paid to the possibility of shadows being created by the working movement of the equipment. Further details of the lighting requirements for exterior installation can be found in CIBSE LG 06 *The Outdoor Environment* [8].

Lighting for areas containing visual display units

In order to exercise control over the working environment in which visual display units (VDUs) are used, legislation has been introduced which affects both employed and self-employed workers who habitually use VDUs for a significant part of their normal work. The Health and Safety (Display Screen Equipment) Regulations [9] came into force on 1 January 1993. The Regulations implement a European Directive on minimum safety and health requirements for work with display screen equipment. Employers have a requirement to ensure that VDU workstations comply with the Regulations.

General lighting requirements

In accordance with the Schedule to the Regulations, any room lighting or task lighting provided shall ensure satisfactory lighting conditions and produce an appropriate contrast between the VDU screen and the background environment, taking into account the nature of the work and the visual requirements of the operator or user. Any possible disturbing glare and reflections on the VDU screen, or other equipment, shall be prevented by coordinating workplace and workstation layout with the location and technical characteristics of the artificial light sources.

Reflections and glare

In accordance with the Schedule to the Regulations, workstations shall be designed so that sources of light such as windows and transparent or translucent walls cause no direct glare and no

distracting reflections on the screen. Windows shall be fitted with a suitable system of adjustable covering so as to attenuate the daylight that falls on the workstation.

Luminaires for VDU areas

The CIBSE (Chartered Institution of Building Services Engineers) LG 03 [10] gives recommendations for the use of luminaires in areas where there are VDUs.

The three-option category system of rating downlighter luminaires previously used in display screen equipment areas has now been superseded by a single system that limits the luminance to 200 cd m^{-2} above a cut-off angle of 65° to the downwards vertical (Fig. 19.12). In special circumstances, this angle may be reduced to 55° .

Use of uplighters

Uplighters, producing a form of indirect lighting, are often used in areas containing VDUs. Essentially, the light is directed onto the ceiling from where it is then reflected downwards. This method of illuminating an area overcomes the problem of direct vision of the light source itself and thereby the effects of glare on display screens are almost totally eliminated. Unfortunately, in order to prevent adverse lighting conditions, the use of up-

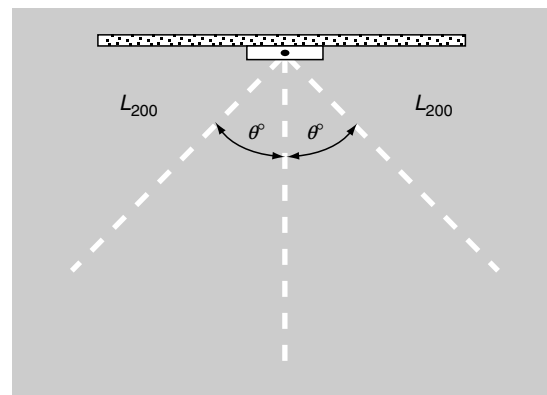


Figure 19.12 Luminaire cut-off angle. L_{200} represents a luminance limit of 200 cd m^{-2} .

lighters places certain restrictions on the finish and colour of the room fabrics. In extreme cases it is possible to inadvertently create a large luminous ceiling that amplifies the oscillations in light output produced by the alternating electrical supply to the lamps used. This will create an unwanted effect that may lead to visual fatigue when subjected to prolonged exposure.

Lighting surveys and survey techniques

Preliminaries

Scale drawings should be produced, giving details of all principal working surfaces, windows, luminaires and other structural and/or decor features. When confusion is likely to develop, sectional views will be beneficial.

Preliminary report sheet

Figure 19.13 shows a typical preliminary report sheet.

Determination of the minimum number of measuring points

For the results of a lighting survey to be meaningful there must be a minimum number of measuring points within the interior being surveyed. The number of points will be influenced by the geometry of the interior. One method of determining the minimum number of measuring points is to calculate the room index (RI) of the interior and then calculate the corresponding minimum number of measuring points. To calculate RI:

$$\text{RI} = \frac{\text{length} \times \text{width}}{(\text{length} + \text{width}) \times \text{mh}} \quad (19.4)$$

where 'mh' is the mounting height of the luminaires above the working plane.

The minimum number of measuring points is subsequently found by:

$$\text{Minimum number of measuring points} = (X + 2)^2 \quad (19.5)$$

where X is a parameter whose value is dependent upon the RI. For all values of RI less than 3.0, the value of X is taken as the next highest integer. For all values of RI equal to or greater than 3.0, the value of X is fixed at 4.0. Table 19.4 gives examples of the relationship between RI, X and the minimum number of measuring points.

The procedure outlined describes the minimum number of measuring points. However, it may be considered beneficial to select a higher number of measuring points if the room geometry so dictates.

Measuring equipment

Illuminance meters

The spectral response of the cells used in the instruments differs from the response of the human visual system. The response is typically corrected by the use of filters and when filters are incorporated the instrument is referred to as 'colour corrected'. A further correction is applied to take into account the direction of incident light falling upon the detector cell. Instruments that are capable of accurately measuring illuminance from differing directions of incident light are said to be 'cosine corrected'.

Luminance meters

Luminance meters use photovoltaic cells similar to those used in illuminance meters. Such instruments also have to be colour corrected.

Interpretation of data

Levels of illuminance should be compared with CIBSE requirements [1] for the type of interior under consideration. 'Patchy' lighting should be avoided wherever possible. To this end the uniformity ratio is a useful indicator:

$$\text{Uniformity ratio} = \frac{\text{minimum illuminance}}{\text{mean illuminance}} \quad (19.6)$$

The minimum acceptable value of the uniformity ratio is 0.8.

Lighting survey sheet

Date Time

Location Address

Survey Person

Reason for Survey.....

Room Dimensions: L = W = H =

Window Dimensions: H = W =

Daylight Availability: Side Glazing Roof Glazing

Artificial Lighting: Luminaires/Lamps

Luminaire Type

Lamp Type

Lamp Rating

Date of Lamp Change

Condition of Equipment and Room Fabrics:

Ceiling

Walls

Floor

Windows

Luminaires

Principal Visual Tasks

Principal Planes of Interest

CIBSE Recommended Illuminance Values

Figure 19.13 Typical preliminary report sheet.

Table 19.4 Relationship between room index, parameter X and minimum number of measuring points.

Room index	Parameter X	Minimum number of measuring points
0.9	1	9
1.9	2	16
2.0	3	25
3.0	4	36
3.4	4	36
5.1	4	36

Table 19.5 Typical illuminance levels.

Location or event	Typical illuminance levels (lux)
Starlight	0.2
Moonlight	2.0
Side road or estate road lighting	5.0–10.0
Domestic interior lighting	100–300
Workshop benches	400–500
General offices	500–750
Drawing offices	500–750
'Bad light' stops play at cricket	1000
Sports ground lighting for colour	500–2000
TV transmission	
Operating theatres	10 000–50 000
Bright sunlight	50 000–100 000

Note: In *Code for Lighting* (Chartered Institution of Building Services Engineers, 2002) [1] details are given of the recommended illuminance levels for domestic, commercial and industrial applications.

Typical illuminance levels

Typical illuminance levels relative to everyday events are listed in Table 19.5.

References

- Chartered Institution of Building Services Engineers (CIBSE) (1994). *Code for Lighting*. (See also New Code for Lighting, 2002.)
- Smith, N.A. (2000). *Lighting for Health and Safety*. Butterworth Heinemann, Oxford.

- International Electrotechnical Commission (IEC) (1993). *International Lamp Coding System (ILCOS)*, IEC Document no. 123–93.
- British Standards Institution (1990). *Electric Luminaires*. British Standards Institution BS 4533 (see also EN 60–598, 1989).
- Chartered Institution of Building Services Engineers (CIBSE) 7: Offices 2003, which is included in the CIBSE Code for Lighting 2002 [1]. Lighting Guide.
- British Standards Institution (1967). *Specification for Artificial Daylight for the Assessment of Colour. Illuminant for Colour Matching and Colour Appraisal*. British Standards Institution BS 950–1.
- British Standards Institution (1999). *Emergency Lighting. Part 1. Code of practice for the emergency lighting of premises other than cinemas and certain other specified premises used for entertainment*. British Standards Institution BS 5266.
- Chartered Institution of Building Services Engineers (CIBSE) (1992). *The Outdoor Environment*. Lighting Guide LG 06.
- Health and Safety Executive (1992). *Health and Safety (Display Screen Equipment) Regulations*. Statutory Instrument 2792.
- Chartered Institution of Building Services Engineers (CIBSE) (1996). *The Visual Environment for Display Screen Use*. Lighting Guide LG 03.

Suggested further reading

- British Standards Institution (1988). *Electrical Apparatus with Protection by Enclosure for Use in the Presence of Combustible Dusts, Part 2*. British Standards Institution BS 6467.
- British Standards Institution (1989). *Code of Practice for Selection, Installation and Maintenance of Electrical Apparatus for Use in Potentially Explosive Atmospheres (Other than Mining Applications or Explosive Processing and Manufacture)*. British Standards Institution BS 5345.
- Guidance on Provision and Use of Work Equipment Regulations 1998 (L22)*.
- Health and Safety Executive (1992). Display screen equipment work. In *Health and Safety (Display Screen Equipment) Regulations*. Guidance on Regulations (L26).
- Health and Safety Executive (1996). Safety Signs and Signals. In *Health and Safety Regulations*. Statutory Instrument no. 341.
- Health and Safety Executive (1996). Safety Signs and Signals. In *Health and Safety Regulations*. Guidance on Regulations (L64).
- Institution of Lighting Engineers (1997). *Guidance Notes for the Reduction of Light Pollution*.

International Labour Office (ILO) (1998). *Encyclopaedia of Occupational Health and Safety*, 4th edn, Vol. II. International Labour Office (ILO), Geneva.

International Occupational Safety and Health Information Centre (1977). *The Workplace*, Vol. I. International Occupational Safety and Health Information Centre, Geneva.

Lighting Industry Federation (LIF) *LIF Technical Statements 1 to 17* (included in the CIBSE Code for Lighting 2002 [1]).

Provision and Use of Work Equipment Regulations (1998). Statutory Instrument SI 2306.

Special Waste Regulations (1996) (as amended). Statutory Instrument SI 972.

Chapter 20

The thermal environment

Antony Youle

Introduction	Wind chill index
Thermal balance	Required clothing insulation
Heat transfer mechanisms	Surveying the thermal environment
Conduction	Objective measurements and instrumentation
Convection	Personal monitoring
Radiation	Air temperature
Evaporation	Radiant temperature
Overall balance	Globe thermometer
Evaluation of an environment	Humidity conditions
Thermal indices	Wet and dry bulb methods
Types of thermal indices	Air velocity
Rational indices	Katathermometer (cooling or 'down' thermometer)
Empirical indices	Cooling resistance/thermistors anemometer
Direct indices	Others
Selection of appropriate thermal indices	Integrating meters
Heat stress indices and standards	Principles of control
Heat stress – empirical and direct indices	Risk assessment
Wet bulb temperature	Control for hot conditions
Wet and dry bulb type	Planning
Effective temperature and corrected effective temperature	Environmental control
Wet bulb globe temperature	Control of the source
Heat stress – rational/analytical indices	Ventilation, air-conditioning and air movement
Required sweat rate – ISO 7933	Evaporative cooling
Heat stress index	Radiation shields and barriers
Predicted 4-h sweat rate	Managerial aspects
The use and application of heat stress indices	Protective clothing and equipment
Indices for comfort	Control for cold conditions
Empirical	Clothing
Analytical: Fanger analysis and ISO 7730	Work activity
Direct index: dry resultant temperature	Work–rest regimes
Indices for cold stress	Control and comfort
Still shade temperature	References
	Further reading

Introduction

Humans are warm-blooded animals; the body core temperature (i.e. the temperature deep in the body tissues) must be regulated normally to remain within a narrow range, typically $37.0 \pm 0.5^\circ\text{C}$ ($98.4 \pm 1^\circ\text{F}$). This process, termed 'heat homeo-

stasis', is required because many of the biochemical and cellular processes on which bodily functions depend take place efficiently and correctly only within this narrow range. External conditions that permit the narrow control conditions to be maintained are termed thermally 'neutral' or in the 'neutral zone'. Outside this zone, the

environment can be considered to be applying either heat or cold 'stress' to the body, with potential thermal 'strain' effects. The usual maximum deviation of core temperature that can be tolerated in fit people is approximately $\pm 2^\circ\text{C}$, but with potential strain effects. In the extreme, if the core temperature drops to about 31°C , strain is manifest by loss of consciousness and death can ensue rapidly; above 43°C the thermoregulation mechanism can fail, again with potentially fatal consequences (see Chapter 7).

The body generates heat continuously by the conversion of food to energy via the metabolic system, and using the energy in the form of work done. The majority of energy is converted to heat, which contributes to maintaining the body temperature, but may cause overheating. There must be an appropriate balance between the heat generated and the heat lost to, or gained from, the environment to maintain heat homeostasis.

Thermal balance

Whenever a temperature difference occurs between an object and its surroundings, heat transfer processes will occur to reduce this difference. The processes available, depending on the physical circumstances prevailing, are: conduction, convection, radiation and phase changes (e.g. evaporation). This transfer will apply to any such circumstances in occupations, e.g. heating effects of furnaces, solar heat gains through glazing in buildings, chilling from a cold wind or heat distribution from a heating system.

In particular, heat exchange mechanisms are continuously functioning between the human body and its surroundings. Under normal circumstances, there is a net loss from the body to its surroundings for the body's core temperature to remain constant, i.e. there is an equilibrium between internal heat production and heat loss from the surface (Fig. 20.1). This thermal balance can be expressed in the form of the equation:

$$M = \pm K \pm C \pm R - E \quad (20.1)$$

where M is the rate of metabolic heat production, K , C and R are the loss or gain of heat by conduc-

tion, convection and radiation respectively and E is the heat loss from the skin and respiratory tract due to the evaporation of moisture. Other factors that may affect the balance are the external work (w) performed by or on the body (affecting M) and the rate of change in the store of heat (S) in the body.

The value of metabolic heat production in the basal state, i.e. with complete physical and mental rest, is about 45 W m^{-2} (i.e. per m^2 of body surface area) for a 30-year-old man. With the surface area of a typical male being 1.8 m^2 , this amounts to $\approx 80 \text{ W}$ for the whole body. As activity or physical work increases, the metabolic rate also increases. Typical values of M for differing activities are given in Table 20.1. Metabolic heat is largely determined by muscle activity during physical work but may be increased at rest in the cold by the involuntary muscle contractions involved in shivering.

Heat transfer mechanisms

Conduction

The rate at which heat is transferred by conduction depends on the temperature difference between the body and the contacting material, the thermal conductivity of the material and the area of contact.

In normal circumstances, conduction transfer from the body direct to objects is voluntarily minimized as sensations of discomfort arise, for instance when resting a bare arm on to a good conductor (e.g. metal). The value for K in the heat balance equation is thus usually ignored. However, immersion in cold water is an example where conduction losses (i.e. to the water) are significant and can quickly lower the body temperature to dangerous levels.

Conduction is an important consideration in relation to clothing. An arbitrary unit of resistance (i.e. the inverse of conduction) the 'clo', is used for expressing the insulation value of clothing. By definition 1.0 clo is the insulation provided by clothing sufficient to allow a person to be comfortable when sitting in still air at a uniform temperature of 21°C . 1.0 clo is equivalent to a clothing resistance (insulation) value of $0.155 \text{ m}^2 \text{ }^\circ\text{C W}^{-1}$. Examples of clo values are given in Table 20.2.

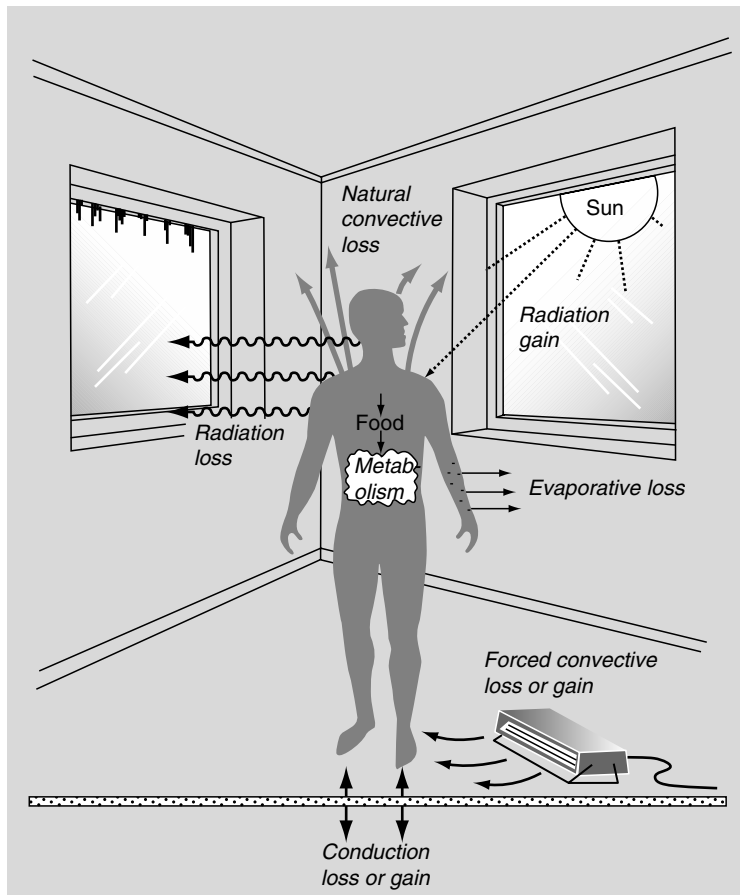


Figure 20.1 Heat balance mechanisms for the human body.

Table 20.1 Metabolic rates for differing activities.

Activity	Typical rate per person (W)	Rate per square metre of body surface (W)
Sleeping	75	43
Sitting	105	60
Light work	160	90
Walking	280	154
Heavy work	450–650	250–360
Shivering	1000	600

Note: a typical male adult has a body area of 1.8 m².

Table 20.2 Typical values for the insulation of clothing (in clo units).

Clothing assembly	Clo
Naked	0
Shorts	0.1
Light summer clothing	0.5
Typical indoor clothing	1.0
Heavy suit and underclothes	1.5
Polar clothing	3–4
Practical maximum	5

Convection

Heat exchange by convection is dependent on the temperature difference between the surface and the air and the nature of the surface via a factor known as the ‘convective heat transfer coefficient’.

For still air, the process is termed ‘natural’ or ‘free’ convection. Air velocity (v), due to the body moving through the air or air moving across the body, will significantly increase the convection loss. This is termed ‘forced convection’, with the

heat transfer coefficient now strongly dependent on 'v'.

Radiation

Radiant heat emission from a surface depends on the absolute temperature T (in kelvin, K, i.e. $^{\circ}\text{C} + 273$) of the surface to the fourth power, i.e. proportional to T^4 .

Because of the T^4 relationship, radiation exchange is particularly relevant when dealing with objects of high surface temperature, e.g. energy from the sun (surface temperature approximately 6000 K) and molten or 'red-hot' steel (≈ 1000 –1500 K). Similarly, if significant heating effect by radiation is required from a small source (for heating purposes) then it must be at a high temperature (e.g. a bar electric fire).

For many indoor situations, the surrounding surfaces are at a fairly uniform temperature and the radiant conditions can be described by the mean radiant temperature (MRT). MRT can be estimated from the temperature of the surrounding surfaces or alternatively obtained by measurement (see Surveying the thermal environment).

Evaporation

At rest in a comfortable ambient temperature, an individual loses moisture by evaporation of water diffusing through the skin (cutaneous loss) and from the respiratory passages. Total water loss in these conditions is approximately 30 g h^{-1} , with a corresponding heat loss due to evaporation. Such heat loss is known as 'insensible' (or 'latent') loss, with a corresponding insensible water loss. This term refers to heat loss without an associated temperature change, in contrast with 'sensible' heat loss (or gain), where heat transfer (by conduction, convection or radiation) causes temperature changes that are detectable by the 'senses'.

The latent heat of vaporization of water is 2453 kJ kg^{-1} at 20°C , resulting in a heat loss equal to approximately 10 W m^{-2} . This loss is increased by the process of sweating. For example, a high sweat rate, e.g. of 11 h^{-1} , will dissipate about 680 W over the whole body, corresponding

to the metabolic rate for high activity (see Table 20.1). This value of heat loss is only obtained if all the sweat is evaporated from the body surface; sweat that drips or is removed from the body is not providing effective cooling.

Evaporation loss results from the vapour pressure difference skin to air, i.e. it depends on the ambient temperature, humidity and air movement conditions.

Overall balance

Under normal temperature conditions providing comfort for sedentary activity, i.e. with a metabolic rate of typically 100 – 120 W and light to medium clothing, the relative losses by the four heat transfer mechanisms are:

- conduction 0%
- convection 25%
- radiation 45%
- evaporation 30% [i.e. overall 30% insensible (latent) and 70% sensible].

The precise relative values will depend on the exact conditions prevailing.

As the ambient conditions increase in temperature, the loss mechanisms by convection and radiation decrease, as the temperature difference between the body surface and the surroundings decreases. Hence to maintain balance, the evaporative loss increases by sweating. When the temperature difference reaches zero, or is reversed, convection and radiation losses cease and become gains, and the only loss mechanism is evaporation. Thermal balance may not then be maintained, leading to a rise in core temperature and associated physiological effects.

Conversely, as temperatures fall below comfort values, convection and radiation losses increase. In the extreme, the balance is lost and the core temperature will fall. These effects are summarized in Fig. 20.2.

Four of the thermal parameters affecting the body heat transfer mechanisms are related to the environment: air temperature, MRT, humidity conditions and air velocity, and two are related to the individual: activity and clothing. The time exposed to the prevailing conditions is of importance when balance is not maintained, as this will affect

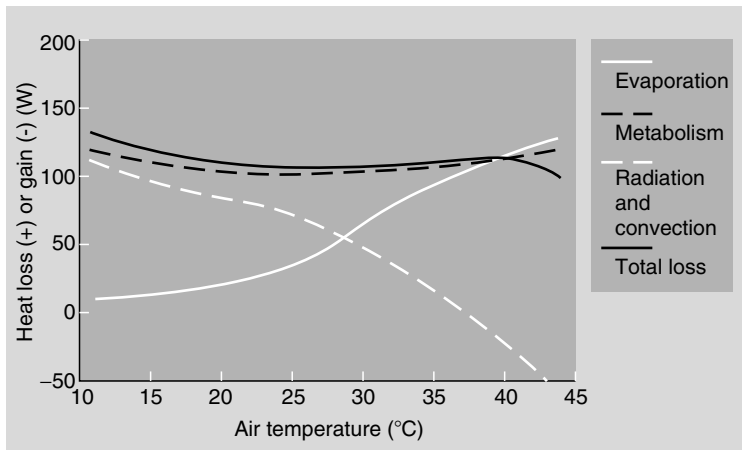


Figure 20.2 Relative rates of heat transfer from the body with ambient conditions (air temperature).

the magnitude of the change in body temperature and the resulting health risks.

Evaluation of an environment

Thermal indices

The thermal environment can be assessed in either subjective terms or by objective measurements. In the former case, individuals are asked to give an opinion on the thermal environment that they are experiencing. This is usually undertaken by the use of a standard subjective scale with which they are provided (e.g. British Occupational Hygiene Society, 1996). This is particularly relevant in assessing thermal comfort in office-type environments, for instance, in which a large number of people may be experiencing a similar objective environment. However, in extreme environments such an approach would yield little useful or reliable data except perhaps to warn that a problem is imminent.

An alternative approach is to quantify the objective parameters making up the environment, i.e. air temperature, radiant temperature, humidity and air velocity, together with activity and clothing. However, specification of these six parameters to define the thermal environment is both cumbersome and of little use to all except the experienced worker in the field. Thus, much effort over the years has been put into developing ‘thermal in-

indices’ that summarize the prevailing conditions by a single number, which can then be related to criteria for the index to indicate the severity of the environment in question.

The purpose of an index is to ‘sum up’ the inter-relation of the six objective parameters in a single figure in relation to the thermal performance (or heat balance) of the human body. The various parameters may change, but in such a way as not to affect the thermal balance, and hence the value of the index itself would not change. For instance, a rise in air temperature may be compensated for by a fall in radiant temperature or a rise in air velocity (or vice versa); an increase in activity may be balanced by a reduction in clothing or in air/radiant temperature. In such cases, the index value would remain constant although the separate parameters are changing. Alternatively, where a change in parameters (whether one or all six) causes a change in the thermal balance then this would be reflected in a change in the single parameter, i.e. the index.

Indices do not necessarily take into account all six parameters. For simplicity, they may consist of a dominant factor only (e.g. wet bulb) as a simple assessment of heat stress for a given type of activity (e.g. mining). Alternatively, some of the parameters may be assumed to be fixed or constant (e.g. activity and clothing), thus simplifying the inter-relation of the remaining four. This may occur for particular industries or occupations

which the index has been developed for, or applied to, specifically. Therefore, when using an index, care must always be taken to ensure that it is not being applied out of context.

Indices have been developed by many researchers, over many years, for different circumstances and different applications. In general, three types of approaches have evolved, namely 'rational', 'empirical' and 'direct' (e.g. BOHS, 1996; Parsons, 2003).

Types of thermal indices

Rational indices

Rational (or analytical) thermal indices incorporate the principles of heat exchange and heat balance in assessing human response to hot, neutral and cold environments. If a body is to remain at a constant temperature, the heat inputs to the body need to be balanced by the heat losses as previously quoted (Equation 20.1).

Analysis procedures require the values represented in the equation to be calculated from a knowledge of the physical environment, clothing and activity. Rational thermal indices use heat transfer equations and, sometimes, mathematical representations of the human thermoregulatory system to 'predict' human response to thermal environments.

A comprehensive mathematical and physical appraisal of the heat balance equation represents the approach taken by Fanger (1970) in relation to thermal comfort. This is the basis of ISO 7730 *Moderate Thermal Environments* (International Organization for Standardization, 1994). This approach enables the prediction of conditions that should provide 'comfort' or neutrality for differing levels of activity and clothing.

Similarly, ISO 7933 *Hot Environments – Analytical Determination of Thermal Stress using Calculation of Required Sweat Rate* (ISO, 1989) assesses heat stress conditions by the ability of the body to lose sufficient heat to the environment by the evaporation of sweat via a thorough analysis of the heat balance equation.

Empirical indices

Empirical thermal indices are based upon data collected from human subjects who have been exposed to a range of environmental conditions. Examples are the effective temperature (ET) and corrected effective temperature (CET) scales. These scales were derived from subjective studies on US marines. They are not fully comprehensive in relation to the six thermal parameters – partly to simplify the approach and partly because the scale was originally devised for particular circumstances (i.e. marines on ship decks in warm/hot conditions).

For this type of index, the index must be 'fitted' to values predicted by experience to provide 'comfort' or a degree of stress, e.g. a certain value (or range of values) of CET is recommended for a given occupational activity.

Direct indices

Direct indices are measurements taken by a simple instrument that responds to similar environmental components to which humans respond. For example a wet, black globe with a thermometer placed at its centre (often termed a 'botsball') will respond to air temperature, radiant temperature, air velocity and humidity. The temperature of the globe will therefore provide a simple thermal index, which can provide, with experience of use, a method of assessment of hot environments. Other instruments of this type include the temperature of a heated ellipse and the integrated value of wet bulb temperature, air temperature and black globe temperature (the WBGT scale). The application of direct indices is empirical in nature.

Selection of appropriate thermal indices

The first action in the selection of a thermal index is to determine whether a heat stress, comfort or cold stress index is required. Numerous thermal indices have been developed and most will provide a value that will be related to human response (if used in the appropriate environment). An important point is that experience with the use of an index should be gained in a particular occupation

or industry. In practice, it is advisable to gain experience initially with a simple direct index; this can then be used for day-to-day monitoring. If more detailed analysis is required, a rational index can be used (again experience should be gained in a particular industry) and, if necessary, both subjective and objective measurements taken. There may be circumstances when conditions are particularly extreme and lie outside the range of the index selected. In such cases a suitable alternative index must be selected, or the problem should be examined from first principles and direct physiological measurements made (e.g. of core temperature and heart rate).

Heat stress indices and standards

The purpose of a heat stress index is to provide the means for assessing hot thermal environments to predict their likely effect on people. In theory, a heat stress index will take account of all environmental factors to produce a single index number that will enable the stress, strain or risk to them to be assessed when considered in relation to a person, their clothing and metabolic rate. In practice, many indices will not consider all factors, or will deal with fixed values of one or more of the parameters in order to simplify the approach.

Empirical indices do not readily permit detailed consideration of the individual components of the thermal environment but, being practically derived, they are more widely used as the basis for standards. Theoretically derived standards allow detailed consideration of the factors controlling the body's heat balance and are therefore useful when assessing changes or control measures, but because their basis is theoretical they are often more complex to apply.

There are a number of indices that have been used to assess heat exposure. They are principally intended to prevent the deep body temperature from exceeding 38°C; they do not necessarily protect against the milder or chronic effects of heat. Control is achieved either by limiting environmental conditions in which exposure or work is permitted, or by limiting the time of exposure.

No index or standard is applicable in all circumstances as they have varying quantitative effects to changes in the individual components of the thermal environment. Thus, the application of different indices or standards to a particular environment often produces differing results. Also, a standard which is intended to provide a safe environment for all people will be far more restrictive than one aimed at young, fit people. A number of the more commonly used indices are described.

Heat stress – empirical and direct indices

Wet bulb temperature

Wet bulb temperature alone may be used in certain circumstances. For example, a maximum wet bulb temperature of 27°C has been quoted as a standard for tunnelling work when radiant heat is rarely a problem but humidity is often high, but the approach is rather limited.

Wet and dry bulb type

The Oxford index combines dry bulb (DB) and wet bulb (WB) ($0.15 \text{ DB} + 0.85 \text{ WB}$) and has been used to predict tolerance times for fit men working for short periods in extreme conditions. It was derived for mines rescue personnel wearing breathing apparatus.

Effective temperature and corrected effective temperature

Effective temperature (ET), developed in the 1920s by subjective tests on US marines, takes account of wet bulb temperature, dry bulb and air velocity. The CET was 'corrected' to take into account radiation conditions, by incorporating the globe (150-mm) thermometer temperature in place of the dry bulb temperature. Two levels of clothing are considered ('normal', i.e. lightly clad, and 'basic', i.e. stripped to the waist) and subsequent developments also allowed for varying work rate. CET is not generally used nowadays to predict heat stress, although it is still applied, for instance, in coalmining industries, in preference to other thermal standards.

Wet bulb globe temperature

Wet bulb globe temperature (WBGT) is the most widely accepted heat stress index and forms the basis of many standards, most notably threshold limit values of the American Conference of Governmental Industrial Hygienists in the USA (ACGIH, 2004), ISO 7243 (ISO, 1982, revised 1989) also as BS EN 27243 (British Standards Institution, BSI).

WBGT is calculated from:

$$\text{WBGT} = 0.7 \text{ WB} + 0.3 \text{ GT indoors} \quad (20.2)$$

or

$$\text{WBGT} = 0.7 \text{ WB} + 0.2 \text{ GT} + 0.1 \text{ DB outdoors} \quad (20.3)$$

where WB is the wet bulb temperature (natural), GT is the globe thermometer temperature (150-mm-diameter globe) and DB is the dry bulb temperature. The outdoors formula reduces the influence of the globe contribution from direct sun.

The WBGT index was originally derived to reduce heat casualties in the USA during military training. It takes account empirically of radiant and air temperatures, humidity and low air velocities (principally via the natural wet bulb and partly via the globe reading). The index alone does not provide guidance to exposure; it must be used with

empirical recommendations based on the body core temperature not exceeding 38°C.

Table 20.3 reproduces the standards for exposure as specified by ISO 7243 (ISO, 1982, revised 1989) using WBGT. These figures are generally conservative and indicate a level that is likely to be safe for most people who are physically fit and in good health. It is to be noted that different values are quoted for persons acclimatized and not acclimatized to heat. This standard is based on clothing worn being light summer clothing (clo value 0.6) and does not allow for variations in clothing, in particular the wearing of personal protective equipment (e.g. full respirator suit). It is assumed that rest periods are at the same thermal conditions as the activity and there is adequate water and salt intake. If conditions of exposure fluctuate, an appropriate time-weighted average exposure value should be derived.

The ACGIH quote the WBGT index for heat stress in their annual threshold limit values (TLV) listings (e.g. ACGIH, 2004). Included in their guidance is a table for work–rest regimes, e.g. for a given WBGT value and work rate (for un/acclimatized workers) a regime of 25% work, 75% rest each hour is recommended. Correction factors are also given for different clothing regimes, as also discussed in *Ergonomics of the Thermal Environment* (BSI, 2000).

Table 20.3 Reference values of wet bulb globe temperature (WBGT) heat stress index from BS EN 27243 (1994).

Metabolic rate (<i>M</i>)			Reference value of WBGT	
Metabolic rate class	Related to unit skin surface area ($W m^{-2}$)	Total (for a mean skin surface area of $1.8 m^2$) (<i>W</i>)	Person acclimatized to heat (°C)	Person not acclimatized to heat (°C)
0 (resting)	$M < 65$	$M < 117$	33	32
1	$65 < M < 130$	$117 < M < 234$	30	29
2	$130 < M < 200$	$234 < M < 360$	28	26
3	$200 < M < 260$	$360 < M < 468$	25* (26 [†])	22* (23 [†])
4	$M > 260$	$M > 468$	23* (25 [†])	18* (20 [†])

*No sensible air movement

[†]Sensible air movement.

If the reference values are exceeded then measures must be taken to either reduce the WBGT value or implement a work–rest regime. (After ISO, 1982, revised 1989. Extracts are reproduced with the permission of BSI. Complete standards can be obtained by post from BSI Customer Service, 389 Chiswick High Road, London W4 4AL; tel. +44(0)20 8996 9001.)

Heat stress – rational/analytical indices

Required sweat rate – ISO 7933

Various indices have been developed to predict thermal strain on the body by applying the heat balance equation as described previously (Equation 20.1). The most developed of these is given in ISO 7933 (ISO, 1989), which can be used for predicting heat strain from a very wide range of factors by determining the required sweat rate. The procedure requires knowledge of all of the environmental parameters, work rate and clothing. It also takes into account the evaporative efficiency of sweat rate and uses the concept of skin ‘wettedness’ (wet: total skin area). Recommendations for exposure are based on limiting the rise of core temperature and on assessing the strain induced by the sweating process. Although being the most extensive and detailed of the methods for predicting heat strain and the effects of the different components of the thermal environment, ISO 7933 is complex and difficult to use. It is generally not suitable for occasional or casual use; a computer program is given in the standard to facilitate calculations.

This index is under review, the revised version being termed the *predicted heat strain* (PHS) (ISO, 2001).

Heat stress index

An earlier analytical index is the heat stress index (HSI) developed by Belding and Hatch (in the 1950s). This is orientated towards assessment via knowledge of environmental conditions and heat balance, and is often referred to as an ‘engineering’ approach, as individual components of the environment can then be modified to provide control. The method for calculating the HSI is based on simple heat transfer equations to determine the required evaporative (i.e. sweat) loss compared with the maximum evaporative loss in the environment.

HSI is the ratio of these losses expressed as a number between 0 and 100, representing stress and hence indicating strain. Conditions giving an HSI of below 40 are not considered to pose a risk

to health; above 40 the risk increases; and 100 is the maximum tolerated by fit, acclimatized young men, when heat gain matches the maximum heat loss by evaporation.

Over 100 there is a net heat gain to the body and the core temperature will rise unless the exposure time is limited. The maximum allowable exposure time (AET) can also be calculated.

Predicted 4-h sweat rate

In contrast with the HSI, the predicted 4-h sweat rate (P4SR) index is ‘physiologically’ based, enabling a nominal sweat rate (i.e. strain) to be predicted from criteria relating to the environment and individual. Limiting values for various circumstances are recommended by different organizations. Typically, the recommended upper limit of P4SR for fit acclimatized young men is 4.5 l, whereas for clothed industrial workers the limiting figure is 2.7 l. The procedure for determining the index is via an appropriate nomogram (e.g. Parsons, 2003).

The use and application of heat stress indices

An index does not predict working conditions that are completely safe, as even at moderately elevated temperatures there will be some risk of the milder medical effects, and also individuals vary considerably in their susceptibility. For most applications, ISO 7243 (ISO, 1982, revised 1989) (i.e. WBGT) provides a relatively safe baseline, below which the risk of serious medical effect is very small to most people, whether working continuously or for short periods. However, even the conservative standards of this index will not provide sufficient protection in all cases. In particular, ISO 7243 cannot be assumed to apply to work or activities carried out in impervious protective clothing, and it may not fully reflect the risk if radiant temperature or air temperature and air velocity are high.

In circumstances where ISO 7243 does not apply, or where exposure to more extreme environments may occur, other indices can be used. As the risk increases with temperature, they should

be applied with great caution and in conjunction with other precautions, particularly medical screening, supervision and, if advised by a doctor, medical monitoring. The most appropriate index or indices, which take adequate account of all relevant environmental conditions, clothing, metabolism, etc. should be used. If possible, it is advisable to compare the results of several indices. The required sweat rate approach (ISO 7933: ISO, 1989) will provide the most comprehensive assessment of conditions, although other indices may have practical benefits in particular situations. However, some exposure conditions lie outside the range of any of the indices, in which case the problem needs to be viewed from first principles and is likely to call for direct physiological monitoring. A summary of parameters involved, and recommendations given, for the more widely used heat stress indices is given in Table 20.4.

Indices for comfort

Empirical

The CET scale, previously described under heat stress indices, has also been used as a single figure index for comfort. There are recommended levels for comfort in differing occupations, e.g. CET should lie in the range of 16–18°C for comfort in typical office environments. However, the values

quoted depend on who quotes them, for instance those quoted in the UK are typically 2–4°C lower than those quoted in the USA.

CET has been superseded generally as it is considered to overemphasize the importance of humidity in relation to comfort, as well as not being sufficiently comprehensive.

Analytical: Fanger analysis and ISO 7730

A comprehensive mathematical, physical and physiological appraisal of the heat balance equation was made by Fanger (1970), and forms the basis of ISO 7730 *Moderate Thermal Environments* (ISO, 1994).

This approach enables conditions that should provide ‘comfort’ or neutrality to be predicted for differing levels of activity and clothing. This is usually presented graphically, when air temperature, MRT and air velocity are combined as variables, with relative humidity (RH) assumed to be constant at 50%. (Fanger established that RH variations from 30% to 70% – the range usually found in buildings – have very little effect on thermally comfortable conditions.) Comfort conditions can be expressed graphically for different levels of activity and clothing (Fanger, 1970).

Fanger also expressed comfort in terms of the ‘predicted mean vote’ (PMV), via the standard subjective voting scale:

Table 20.4 Summary of factors and recommendations involved with five heat stress indices.

Parameter	Index				
	WBGT (ISO 7243)	Required sweat rate (ISO 7933)	HSI	P4SR	CET
Air temperature	0.1 DB	Yes	Yes	G/DB	G/DB
Radiant temperature	0.2/0.3G	MRT	MRT	G	G
Humidity	NWB	VP	VP	WB	WB
Air movement	Indirect	m s ⁻¹	m s ⁻¹	m s ⁻¹	m s ⁻¹
Clothing	One level	Clo value	Two levels	Clo value	Two levels
Activity	Three levels	W m ⁻²	W m ⁻²	W m ⁻²	One level (corrections)
Recommendations	Work–rest regimes	Exposure time	Exposure time	Sweat rate limits	Empirical

CET, corrected effective temperature; DB, dry bulb; G, globe; HSI, heat stress index; MRT, mean radiant temperature; NWB, natural wet bulb; P4SR, predicted 4-h sweat rate; WB, wet bulb; WBGT, wet bulb globe temperature; VP, vapour pressure.

- -3 cold
- -2 cool
- -1 slightly cool
- 0 neutral
- +1 slightly warm
- +2 warm
- +3 hot.

For given thermal conditions, Fanger's work predicts what the average vote on the scale will be (for a group of persons), i.e. the PMV for the group, and also the percentage of persons satisfied or dissatisfied with the environment, i.e. the predicted percentage dissatisfied (PPD). For example, if the PMV is +1.0 or -1.0, then the PPD will be 26%, i.e. this percentage of the group is predicted to be dissatisfied with the environment; a PMV of +2.0 or -2.0 gives a PPD of 75%. It should be noted that if the PMV is 0.0 (i.e. the ideal thermal environment) then 5% are still dissatisfied. ISO 7730 suggests that for comfort the limits for PMV in practice should be +0.5 to -0.5, giving a PPD of 10%; the Annex of the standard quotes examples of particular conditions that are predicted to satisfy this limit.

Heat balance alone is not a sufficient condition for thermal comfort. Localized discomfort on parts of the body can still occur even though the body as a whole may be in thermal balance. Individual components of the environment (e.g. air movement) must remain within limits and asymmetry of conditions must be controlled. This applies particularly to vertical temperature gradients and vertical and horizontal radiation effects. The discomfort sensation of air movement is also dependent on the turbulence of the air. Recommendations for limits in relation to localized conditions, e.g. to air movement, temperature gradients, etc. are given in the Annex to ISO 7730.

Direct index: dry resultant temperature

This is an index quoted by the UK Chartered Institution of Building Services Engineers (CIBSE, 1999), and represents a straightforward approach that encompasses all aspects nevertheless. The dry resultant temperature (DRT), often referred to simply as 'resultant temperature', is the value taken up by a 100-mm globe thermometer in 'still

air', i.e. it integrates air temperature and MRT (with RH assumed to be 40–60%), with small corrections to be made for increasing air movement.

Indices for cold stress

Standards relating to work performance, thermal balance and exposure duration in cold environments are not as well developed and validated as those for heat exposure. It is generally more difficult to assess the stress of cold climates, possibly because of the potentially greater part played by behavioural thermoregulation in maintaining body temperature conditions. The objectives of standards for cold exposure are to avoid core temperature falling below 35°C and to prevent cold injury to the extremities such as hands and feet. More widely used examples are given.

Still shade temperature

The still shade temperature (SST) takes actual outdoor conditions and expresses them as an 'equivalent' temperature when there is no solar heat exchange and no wind effect. A correction is applied for the solar heat absorbed by the body (allowing for clothing type, posture, etc.), which in full sunshine can amount to two or three times the resting metabolic rate. Further corrections are required to convert conditions to 'still', i.e. with an air velocity of zero, taking metabolic rate into account. Thus, any set of conditions can be converted to a single index figure. It is still necessary to have empirical recommendations to relate exposure to the index.

Wind chill index

The wind chill index (WCI) is an index of heat loss from the body and was developed by Siple and Passel in 1945 in order to identify the potential risk resulting from the combined cooling effect of wind and cold conditions. It is an empirical approach, based on an artificial model for the human, namely on the cooling characteristics of a warm (33°C) water-filled tin cylinder hoisted on a pole in a station in Antarctica under different con-

ditions of wind speed and temperature. The index is found to correlate well with human reactions to cold and wind, and is successful in identifying conditions of potential danger when skin surfaces are exposed. It is of particular value in estimating the local cooling of hands, feet and head, which may produce deterioration of physical performance and cold injury.

The WCI does not take into account the amount of clothing worn, which is necessary to express theoretically the effect of wind on heat loss of subjects. The proven usefulness of the WCI in practice is probably because, given adequate nutrition, tolerance of cold conditions is ultimately determined by the reaction of parts of the unprotected body. Complete protection of exposed areas by the use of suitable facemask and gloves would, in effect, make the WCI inapplicable. This contrasts with the approach of required insulation (I_{REQ}) (see below).

Values can be expressed graphically or as an equivalent ‘chilling temperature’ (Table 20.5).

Required clothing insulation

The important role of clothing insulation, omitted in the WCI, is used in the required clothing insulation (I_{REQ}) approach to express cold stress in terms of general body cooling and the insulation required to maintain thermal balance. As there is an upper

limit to the amount of clothing insulation possible, a duration period for limiting exposure on the basis of acceptable levels of body cooling may also be calculated for the available clothing.

I_{REQ} is a rational approach to assessing cold stress, based on the heat balance equation. It is defined as the minimal thermal insulation required to maintain body thermal equilibrium under steady-state conditions when sweating is absent and peripheral vasoconstriction is present. The minimal value of I_{REQ} describes the net insulation required to maintain the body in a state of thermal equilibrium at normal levels of body temperature. The higher the value of I_{REQ} at any given activity level, the greater the cooling power of the environment. Alternatively, increasing energy expenditure in the working environment will reduce I_{REQ} . The procedure does not fully cover local cooling of the head, hands and feet; these may require separate consideration, e.g. via the WCI. I_{REQ} is the basis of an ISO standard under development. Examples of the application of cold standards in working environments is to be found in BS 7915 (BSI, 1998).

Surveying the thermal environment

Objective measurements and instrumentation

Assessment of the thermal environment requires the accurate knowledge of the physical quantities involved. The fundamental parameters describing the environment, with the usual units involved, are:

- air temperature ($^{\circ}\text{C}$);
- mean radiant temperature ($^{\circ}\text{C}$);
- relative humidity (%) or absolute humidity of the air (pressure, Pa)
- air velocity (m s^{-1}).

These can be measured individually, or in combined form, to give an ‘integrated’ parameter or index figure direct. Information is usually also required relating to the activity and clothing of personnel involved. These values are normally estimated rather than measured and usually relate to fixed ‘categories’ within the index being applied (e.g. low, medium or high rates of activity), although techniques are available for assessing them in detail, e.g. Parsons (2003).

Table 20.5 The ‘chilling temperature’ (t_{ch}), defined as the ambient temperature that produces the same cooling power as the actual environmental conditions under calm conditions ($< 1.8 \text{ m s}^{-1}$ wind speed).

Wind chill index (WCI) (W m^{-2})	t_{ch} ($^{\circ}\text{C}$)	Effect (typical)
1160	-12	Very cold
1392	-21	Exposed flesh freezes after 60 min
1624	-30	Exposed flesh freezes after 20 min
1856	-40	Exposed flesh freezes after 15 min
2088	-49	Exposed flesh freezes after 10 min
2320	-58	Exposed flesh freezes after 8 min
2552	-67	Exposed flesh freezes after 4 min
2784	-76	Exposed flesh freezes after 1 min

Values of corresponding wind chill index (WCI) are also given.

The thermal environment also varies with time owing to the action of controls, cyclic changes in processes or the influence of varying external conditions. Furthermore, variations throughout the space are likely, particularly near to windows and air inlet and extract grilles or localized sources of heat or cold. Therefore, measurements should normally be made at various positions throughout a room and at three heights: ankle height (0.1 m), abdomen level (0.6 m in sitting areas and 1.1 m in standing areas) and head height (1.1 m in sitting areas and 1.7 m in standing areas). It is usual to weight these values 1:2:1 to reflect the body burden. Knowledge of the variation of parameters with position is also of importance in assessing the asymmetry of conditions, especially in relation to comfort. Measurements should be carried out over the cycling period of the process (see Chapter 11), operation of the heating-cooling controls and, when external solar conditions affect the internal environment, at different times during the day.

Personal monitoring

In cases of heat and cold stress, individual circumstances need to be taken into account when selecting positions for measurement. In general, the position should reflect the thermal load on the individual concerned. The concept of 'personal' monitoring, when equipment is attached to the individual, is not generally applicable in the thermal environment for practical reasons, but may need to be undertaken in special circumstances, e.g. to assess the radiant heat load experienced by firefighters, or to obtain real-time information on the medical condition of the individual, e.g. heart rate and core temperature to give a measure of the 'strain' occurring. However, miniaturization of equipment for sensing and data transmission is leading to more widespread consideration of the personal monitoring approach (e.g. ISO, 2004).

Air temperature

Air temperature (i.e. dry bulb temperature) can be measured with a suitable thermometer: mercury/alcohol in glass, electrical resistance, thermistor and thermocouple or differential expansion type.

Simple glass thermometers are relatively low cost, accurate (in the case of the mercury type typically to $\pm 0.2^\circ\text{C}$) and reliable, but are fragile and inflexible in use. The three electrical types provide more versatility (e.g. with purpose-designed air or surface probes), but accuracy is often less than the mercury in glass type (although resolution to 0.1°C is often presented on the display) and regular calibration is required. The differential expansion principle is more commonly found in continuous recording devices.

Although the instruments themselves are relatively straightforward in operation, many precautions need to be taken when using a thermometer and measuring air temperature (e.g. BOHS, 1996; ISO, 1998).

Radiant temperature

Radiant temperature can be measured indirectly with a globe thermometer or directly with pyrometers, thermopiles or non-contact thermometers. It can also be assessed via surface temperature measurements and calculation. The MRT is the temperature of a uniform imaginary enclosure in which the radiant heat transfer to an object in the enclosure is equal to the radiant heat transfer in the actual space. It is usually measured by instruments that allow the radiation within the enclosure to be integrated into a mean value.

Globe thermometer

The globe thermometer temperature can be used to determine the MRT indirectly, or used as a measure of radiant conditions in its own right to be used directly in an index (e.g. as in the WBGT index).

The globe thermometer is a matt black, hollow copper sphere (original size 150 mm diameter) with a simple thermometer projecting into the centre of the sphere. The sphere is suspended freely and allowed to come into thermal equilibrium with the surroundings (this takes approximately 20 min) to give the globe or 'black bulb' temperature. This can be converted into the MRT value with knowledge of the air temperature and air velocity, either by using appropriate nomograms or by calculation (ISO, 1998). The globe heats

or cools due to radiation exchange, but its final temperature is also a function of heat transfer by convection.

In using the globe thermometer, it should be suspended in an appropriate position, and not influenced by the assessor's own body. As readings for 150-mm globes take typically 20 min, only a limited number can usually be taken. The 40-mm globe responds more rapidly but may give different readings from the 150-mm globe as it is more affected by air movement, i.e. convection losses.

The globe thermometer is not appropriate for obtaining the radiant temperature of a localized radiation source. It measures the overall radiation conditions of all its surroundings, i.e. the MRT. 'Non-contact surface thermometers' or 'infrared thermometers' are suitable for measuring the radiation temperature of particular surfaces, e.g. reaction vessels, furnaces, cold roofs. Alternatively, the surface temperature can be measured directly with an appropriate contact probe, but this technique is often impractical, hazardous or time-consuming.

Humidity conditions

Humidity conditions can be assessed via the relative humidity (RH), usually measured as a percentage figure directly (%), or via the wet and dry bulb temperatures. It can also be expressed as absolute humidity either in kilograms of water per kilogram of air, or as a vapour pressure (in pascals, Pa, or kilopascals, kPa). The inter-relation of these parameters is given by the psychrometric chart (Fig. 20.3). Common types of hygrometers are: wet and dry bulb, dew point, moving fabric, electronic and chemical.

Wet and dry bulb methods

Water evaporating freely from a thermometer will cause a cooling effect termed the 'wet bulb depression', with the resulting temperature known as the 'wet bulb' (compare with the 'dry bulb'). The extent of this cooling is a function of how freely the water can evaporate, i.e. on the relative humidity of the surrounding air and the local air movement. Slide rules, tables or the psy-

chrometric chart provide the RH value for given wet and dry bulbs.

There are two types of measured wet bulb, described by the terms: 'natural', i.e. unspirated, screen, Masons or sheltered; and 'forced', i.e. aspirated, sling, whirling or draught. The value given by the natural wet bulb is influenced by local air movement and is thus less predictable than the forced version. However, it is easier to measure, and in reflecting localized conditions it may be more applicable, e.g. as used directly in the WBGT index.

The most common instrument based on wet and dry bulb is the whirling hygrometer. The principle can also be used with mercury or electrical thermometer sensors. Precautions should be applied in using wet and dry bulb hygrometers (BOHS, 1996; ISO, 1998).

Other instruments used for humidity measurement are the dew point apparatus, moving fabric hygrometers and electrical resistance or capacitance hygrometers.

Air velocity

Anemometers based on a range of principles are available, i.e. katathermometer (cooling), electrical resistance and thermistor types (cooling), moving vane types (mechanical action) and tracer techniques (e.g. smoke). Air movement associated with thermal environments is often very low, i.e. $< 0.2 \text{ m s}^{-1}$, and this limits the selection of a suitable anemometer. Traditionally, the katathermometer satisfies this criterion.

Katathermometer (cooling or 'down' thermometer)

The katathermometer bulb is heated and its cooling time between two fixed temperatures is measured. This is converted to an air velocity using an appropriate chart knowing the surrounding air temperature and instrument calibration factor. If radiation sources are present, then a silvered bulb should be used. The instrument provides multidirectional airflow measurement, averages out fluctuations and can measure to very low air speeds and is thus suited to the measurement of general room air movement.

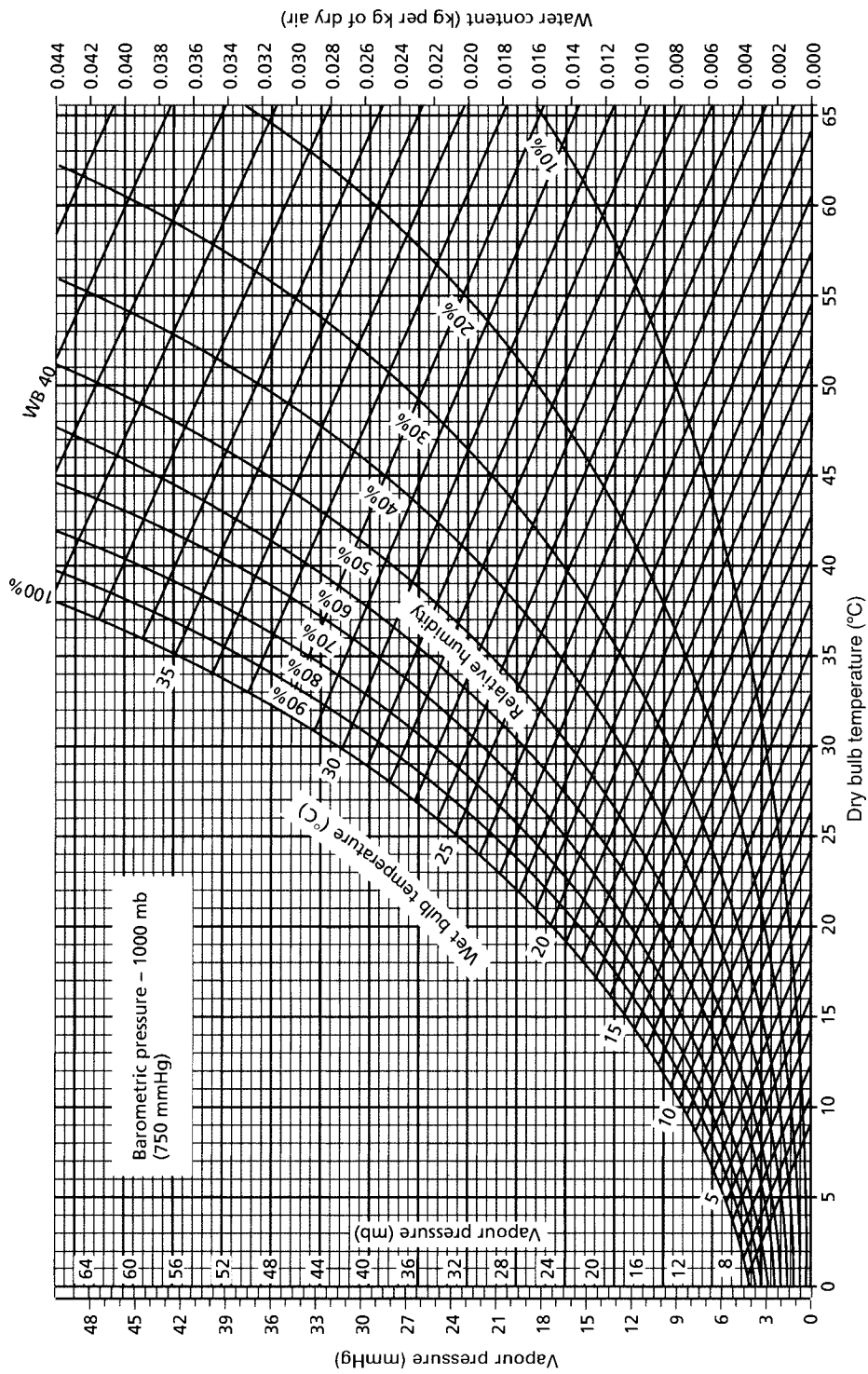


Figure 20.3 The psychrometric chart showing the relationship between dry and wet (sling) bulb temperatures and relative and absolute humidities. Note: for vapour pressure 1 mb = 100 Pa.

Cooling resistance/thermistor anemometer

A wire coil or thermistor bead is heated above the ambient temperature and the electrical current required to maintain it at this temperature is monitored. This is a function of the air velocity and air temperature. These instruments are widely used for measuring high air movement (e.g. $> 0.5 \text{ m s}^{-1}$), but can also be configured for lower values, at relatively high cost. Information on air turbulence, which is relevant for comfort studies, can also be provided.

Others

Moving vane type instruments (e.g. the rotating vane anemometer) are best suited to directional air movement of $> 0.3 \text{ m s}^{-1}$. Tracer methods (the most common being smoke) are essential for identifying and demonstrating airflow patterns in spaces, the effects of ventilation or air-conditioning systems and to identify draughts.

Integrating meters

There are meters that will measure all, or a selection, of these parameters and combine them as appropriate into a single figure scale. For example, a WBGT meter measures air temperature (dry bulb), globe temperature (40–150 mm, depending on the model) and un aspirated (and unshaded) wet bulb temperature, i.e. natural wet bulb (NWB). These are then combined to give a single figure WBGT value. They can be used for continuous remote monitoring and may be linked to alarm systems.

A further integrating device is the thermal comfort meter that is based on Fanger's work and enables the PMV or percentage of people dissatisfied (PPD) with a particular environment to be measured. Other devices, e.g. indoor climate analysers, are specifically designed to assess individual parameters, e.g. factors affecting discomfort, such as measurement of radiation asymmetry and air movement perceived as a draught.

Further details on measurement techniques are given in ISO 7726 (ISO, 1998).

Principles of control

Risk assessment

Typical occupational activities that can lead to heat or cold stress problems are shown in Tables 20.6 and 20.7. The relative contribution of the four environmental factors plus activity (metabolic rate) and clothing (where protective clothing, respirators, etc. can add to the thermal burden) are also given. In any such occupation, and others, a risk assessment is a requirement to identify hazards and risks (of all kinds) in the working environment. Examples of thermal environment type risk assessment procedures are given in *The Thermal Environment* (BOHS, 1996). The outcome of the assessment is likely to be the requirement of control (in one form or other) of the hazard. Principles of control for heat and cold stress conditions are discussed in the following sections. Control in relation to thermal comfort is generally an issue of the fine-tuning of conditions and is raised separately.

Control for hot conditions

The effects of heat stress can be controlled in a number of ways. The planning of work can minimize the length and extent of exposure. Modifying the environmental conditions can reduce the body burden. Appropriate supervision and training is essential for the health, safety and welfare of individuals. Finally, special protective clothing can be of value, whereas protective clothing for other hazards, e.g. toxic contaminants, can lead to exacerbation of heat stress problems.

Planning

Exposure to conditions that could lead to heat strain are best avoided or minimized by careful planning whenever possible. This applies in particular to work activities such as maintenance and repair of hot equipment, replacement of insulating materials on steam pipes etc. and other work of short duration that can often be planned ahead.

Table 20.6 Typical occupational situations where heat stress could occur.

	<i>Temperatures</i>					<i>PPE</i> [†]
	<i>Radiant</i>	<i>Air</i>	<i>Wet bulb</i>	<i>Air velocity</i>	<i>Metabolic rate</i>	
<i>Manufacturing</i>						
Tops of furnaces	High	High	Medium	Medium	Medium	+
Handling molten metal, rolling and forging	H*	Medium	Medium	Medium	High	+
Knockout and fettling	High	High	Medium	Medium	High	
Metal refining	High	Medium	Medium	Medium	High	
Welding, brazing, etc.	High	Medium	Medium	Medium	Medium	+
Glass-making	H*	High	Medium	Low	High	+
Boiler and furnace maintenance	High	High	Medium	Low	High	+
Metal finishing, pickling, galvanizing, degreasing	Medium	Medium	High	Medium	Medium	
<i>Mining and tunnelling</i>						
Face work, deep mines	Medium	High	Medium	Low–Medium	High	(+)
Face work, highly mechanized mines and tunnels	Medium	High	High	Low–Medium	High	
All work in very deep mines	High	High	Medium	Medium	Medium	
Mine rescue work, firefighting	Medium	High	High	Low	High	+
<i>Miscellaneous</i>						
Laundries	Medium	Medium	High	Low	Medium	
Kitchens	High	High	High	Low	Medium	
Firefighting	High	High	High	Low–Medium	High	(+)
Asbestos removal	Medium	High	(High)	Low	High	+
Boiler rooms, compressor houses, electricity generation	High	High	Medium	Medium	Medium	
<i>Shipping and armed services</i>						
Ships, boiler rooms, ships' guns	High	High	High	Low	Medium	
Tanks	High	High	Medium	Low	Medium	+
Fighting aircraft	Medium	High	Medium	Low	low	
<i>Outdoor work in hot places</i>						
Agriculture, quarrying, outdoor marketing	High	High	Medium	Medium	High	

*H, high with red heat or above.

[†]Situations in which required personal protective equipment may contribute significantly to heat strain are indicated by '+'.

If exposure is unavoidable then the risk should be controlled to an acceptable level, preferably by environmental control. The points to be considered when planning for hot work are described in BOHS (1996). When conditions are still likely to lead to heat strain despite all reasonable environmental control measures then additional precautions will be required to reduce personal risk, e.g. medical preselection and acclimatization, supervision and training, appropriate intake of fluids, restriction of work periods and thermal protective clothing.

Environmental control

Modifying the thermal parameters that are contributing to the heat stress conditions can be considered as follows.

Control of the source

Where heat is released by a particular process or source, the temperature of the source itself should be reduced. This may be done by direct temperature reduction, surface insulation, radiant heat emission control or a combination of these factors.

Table 20.7 Typical occupational situations where cold stress could occur.

	Temperatures					
	Radiant	Air	Wet bulb	Air velocity	Metabolic rate	Water*
<i>Outdoor</i>						
Quarrying, tipping, agriculture, stockyards, railways, local authority maintenance	Low	Low	Low	High	–	Medium
Sea fishing, oil rigs, shipping, armed services	Low	Low	Low	High	–	High
<i>Indoor</i>						
Deep-freeze stores	Low	Low	Low	Low to medium	–	Low
Mining in intake airways	Low	Low	Low	High	–	Low
Diving [†]	Low	Low	Low	Low	Low	High

*Refers to situations in which heat loss occurs due to water conduction or evaporation from wet clothing.

[†]Also has high respiratory heat loss.

Ventilation, air-conditioning and air movement

Ventilation can be used for thermal environment control either by removing or diluting hot/humid air and replacing it with cooler/drier air, or by increasing air movement over the body. Cooling effects from air movement result from heat loss/gain from convection and loss from evaporation. However, as the air temperature rises, losses are reduced and may become gains. As a rule of thumb, for hot conditions, if the wet bulb temperature is below 36°C, increasing air velocity over the body is beneficial, but above 36°C it is detrimental.

Evaporative cooling

Air temperature can be reduced *in situ* (via evaporative cooling) by the use of fine water sprays or wetted elements. Although this can reduce the air temperature, it also increases relative humidity and these two factors must be balanced when evaluating the potential benefits. Health risks associated with potential microbiological activity with such processes should also be assessed.

Radiation shields and barriers

Radiant heat from high-temperature sources can be reduced by radiation barriers positioned

between the source and the subject. Ideally, such barriers should be of a material with good insulating properties and have surfaces of low emissivity (i.e. high reflectivity) so that they do not themselves become hot, re-radiate and present a contact hazard. Reflected radiation from the heat source should be directed so as to avoid contributing to the heat load. Metal grids, chains or transparent reflective materials such as partially silvered glass or clear plastics can be used where it is necessary to view the heat source itself. Cold surfaces can also be used as radiation sinks, e.g. water-cooled panels.

Managerial aspects

Whenever work is to be carried out in hot environments, supervision and training are essential to ensure that potential heat casualties are detected quickly and removed immediately to a place for recovery. Personnel should not be allowed to work alone and unsupervised in such conditions.

Restricted work periods (work–rest regimes) invariably require to be implemented, e.g. as recommended by an appropriate heat stress index. Such recommendations may be conservative in nature, e.g. the rest area in ISO 7243 (ISO, 1982, revised 1989) is assumed to be at the same WBGT value as the work itself; for cooler rest areas recovery would be expected to be more rapid. Hence,

more pragmatic approaches based on physiological monitoring may be appropriate.

Protective clothing and equipment

Clothing, especially protective clothing (including respirators), often has an adverse effect on the body's heat balance in hot environments by insulating the body and also reducing evaporative heat loss. In particular, impervious clothing impedes heat loss and the wearing of such clothing may present some risk if physically demanding work or exercise is carried out at air (dry bulb) temperatures as low as 21°C, especially if the wearer is unfit, not acclimatized or otherwise susceptible (see BSI, 2000).

In some circumstances, clothing is required to provide protection against general heat, radiant heat and localized burns (e.g. from molten metal splashes). Heat-resistant protective clothing will only give protection for limited periods and may have a detrimental effect over long periods. If continued exposure is necessary in circumstances in which it would not otherwise be permitted, the use of cooled or conditioned protective clothing may allow longer periods of exposure. Examples are ice-cooled jackets, air-cooled suits and liquid (water)-cooled suits. It should be noted that wearing the jacket or suit can in itself lead to an increase in the metabolic rate and thus thermal strain. Because cooled or conditioned clothing is used in circumstances when exposure would not otherwise be permitted, its use should be restricted to those who are medically fit and there should be a high standard of supervision as the user will be exposed to an unacceptable environment and will need to be removed immediately in the event of its failure.

Control for cold conditions

The principles of control for thermal protection necessary to ensure comfort and well-being in the cold are determined by two sets of factors: personal and environmental. Personal factors include bodily activity (metabolic rate), clothing insulation worn and available, and duration of the exposure. Environmental factors are ambient air temperature and wind velocity in particular, but also

radiant conditions and the presence of precipitation.

Clothing

If shelter is not available, clothing is the most important means of protection against cold stress for people living or working in cold environments. The thermal insulation provided by clothing is a result of the fibrous structure of the clothing itself and air trapped between layers of clothing. Clothing also has to protect against wind, which can penetrate and negate the insulating property of the trapped air. It is therefore necessary for an effective cold weather assembly to be windproof by having an outside layer made of tightly woven or impermeable material.

Clothing that is waterproof is also essential in cold, wet environments because of the rapid cooling produced by the combined effects of evaporation and wind chill. However, a serious disadvantage of waterproofing is that the clothing is also impermeable to water vapour escaping from the skin surface. If it cannot escape, this water vapour will condense beneath the impermeable layer in cold weather and reduce the insulation effects of the trapped air, as well as being a source of discomfort. This effect is increased if the individual is physically active and sweating. In environments with temperatures below 0°C, trapped water in clothing may freeze. Impermeable clothing is mainly useful for people who are not very active. Loosely fitted, with openings around the neck and inbuilt air vents, the garments rely on a bellows effect to vent and reduce water vapour build-up. For active personnel, clothing with special external fabric that is both windproof and waterproof, but allows water vapour transfer, should be used.

The other important consideration with respect to clothing is protection of the extremities and head. Thick, insulating gloves are of little use when fine hand movements are required, and, furthermore, insulation around small-diameter cylinders, like the fingers, is difficult to achieve. Mitts, with all the fingers enclosed together and only the thumb separate, provide more effective insulation. Under survival situations these weaknesses in

insulation can be overcome by withdrawing the hands and arms into the body of the jacket (ensuring that loose sleeves are constrained and made airtight). Local cold injury to the hands and face is especially likely to occur as these areas are the most frequently exposed and particular care must therefore be applied, such as providing local heating to the areas.

Out of doors in snow or ice-covered terrain, eye protection should be provided from blowing ice crystals, whereas safety goggles and exposed skin treatment are needed to protect against ultraviolet radiation and glare.

Work activity

Clothing insulation must be balanced against work performed. However, if work is intermittent (the usual case) then problems can arise, e.g. a worker dressed for thermal protection during periods of inactivity will be overdressed for hard work, and hence providing clothing for operators engaged in intermittent work schedules can present some difficulty.

Work–rest regimes

Warm shelters should be available for rest and recovery for work performed continuously in a cold environment with an equivalent chilling temperature below -7°C . During rest periods it is recommended that dry clothing is provided as necessary and body fluids replaced (preferably by warm sweet drinks and soups) to combat dehydration. Alcohol and caffeine-containing beverages are not advisable as these have adverse diuretic and circulatory effects.

In environments of -12°C or below, e.g. in many cold stores, it is necessary for workers to be under constant observation or supervision. Work rates should not be so high as to cause heavy sweating, but if this is unavoidable, more frequent rest pauses in the warm for changing into dry clothes should be taken. Sitting or standing still for long periods in the cold should be avoided. Air movement from air blast coolers should be minimized by properly designed air distribution systems and should not exceed 1 m s^{-1} at the

worksite. For further discussion, see *Ergonomics of the Thermal Environment* (BSI, 1998).

Control and comfort

In recent years, much attention has been concentrated on the ‘quality’ of internal working environments, particularly offices (e.g. CIBSE, 1999). Thermal comfort represents one area contributing to such ‘quality’, or lack of, and standards for thermal comfort are well established (e.g. ISO 7730: ISO, 1984). However, buildings themselves and thermal control systems can influence conditions detrimentally, whereas the lack of perceived control by occupants can also be an important contributing factor. Issues that are likely to influence the thermal conditions occurring within a space or building include: the building location, orientation (particularly with respect to sun paths), fabric, the active internal temperature-modifying systems (heating, air-conditioning, etc.) with associated controls, and the influence of people, lighting and equipment.

Thus, thermal comfort in buildings is a complex issue in which any factor, or combination of factors, may require attention to improve a thermally unsatisfactory environment. A summary of typical actions to assess conditions can be found in *BOHS Technical Guide No. 12* (BOHS, 1996). It is likely that a range of personnel will be required for investigations of comfort, in particular those responsible for the design, operation and maintenance of the mechanical services plant (building services engineers) and for the building and fabric itself.

References

- American Conference of Governmental Industrial Hygienists (2004). *Threshold Limit Values for Physical Agents*. ACGIH, Cincinnati.
- BOHS (British Occupational Hygiene Society) (1996). *The Thermal Environment*, 2nd edn. BOHS Technical Guide No. 12. BOHS, Derby.
- BSI (British Standards Institution) (1998). *Ergonomics of the Thermal Environment – Guide to the Design and Evaluation of Working Practices for Cold Indoor Environments*. British Standards Institution BS 7915.
- British Standards Institution (2000). *Ergonomics of the Thermal Environment – Guide to the Assessment of*

- Heat Strain in Workers Wearing Personal Protective Equipment*. British Standards Institution BS 7963.
- Chartered Institution of Building Services Engineers (1999). *Chartered Institution of Building Services Engineers, Guide A1*. CIBSE Publications, Balham, London.
- Fanger, P.O. (1970). *Thermal Comfort*. Danish Technical Press, Copenhagen.
- ISO (International Organization for Standardization) (1982) (revised 1989). *Hot Environments – Estimation of the Heat Stress on a Working Man, Based on the WBGT-Index*. International Standard ISO 7243.
- ISO (International Organization for Standardization) (1989). *Hot Environments – Analytical Determination and Interpretation of Thermal Stress Using Calculation of Required Sweat Rate*. International Standard ISO 7933.
- ISO (International Organization for Standardization) (1994). *Moderate Thermal Environments – Determination of the PMV and PPD Indices and Specification of the Conditions for Thermal Comfort*. International Standard ISO 7730.
- ISO (International Organization for Standardization) (1998). *Thermal Environments – Instruments and Methods for Measuring Physical Quantities*. International Standard ISO 7726.
- ISO (International Organization for Standardization) (2001). *Ergonomics of the Thermal Environment – Analytical Determination and Interpretation of Heat Stress using Calculation of the Predicted Heat Strain*, International Standard ISO/CD 7933.
- ISO (International Organization for Standardization) (2004). *Ergonomics – Evaluation of Thermal Strain by Physiological Measurements*. International Standard ISO 9886.
- Parsons, K.C. (2003). *Human Thermal Environments*, 2nd edn. Taylor & Francis, London.
- Note: the ISO standards quoted are available as equivalent BS and/or EN standards.

Further reading

- British Occupational Hygiene Society (1996). *The Thermal Environment*, 2nd edn. BOHS Technical Guide No. 12. BOHS, Derby.
- Chrenko, F.A. (ed.) (1974). *Bedford's Basic Principles of Ventilation and Heating*. H.K. Lewis, London.
- Clark, R.P. and Edholm, O.G. (1985). *Man and his Thermal Environment*. Edward Arnold, London.
- Edholm, O.G. (1978). *Man – Hot and Cold*. Edward Arnold, London.
- International Organization for Standardization (1995). *Ergonomics of the Thermal Environment – Principles and Application of Relevant International Standards*. International Standard ISO 11399 (BS EN ISO 11399, 2001).
- National Institute for Occupational Safety and Health (1986). *Criteria for a Recommended Standard – Occupational Exposure to Hot Environments, Revised Criteria 1986*. NIOSH, Cincinnati, OH.
- Parsons, K.C. (2003). *Human Thermal Environments*, 2nd edn. Taylor & Francis, London.

Chapter 21

Non-ionizing radiation: electromagnetic fields and optical radiation

Philip Chadwick

Introduction	Physical interactions
What are electromagnetic fields and radiation?	Biological effects
Low frequencies	The eye
Physical interactions and established physiological effects	The skin
Static electric fields	Standards
Static magnetic fields	Control
Possible effects on health from 'low-level' exposures	Measurements
Control	Lasers
Measurements	Uses
Radiofrequencies	Hazards
Physical interactions and physiological effects	Classification
Possible effects on health from 'low-level' exposures	Standards
Control	Controls
Measurements	Measurements
Exposure and emission standards for low-frequency and RF electromagnetic fields	References
Optical radiation	Further reading

Introduction

Electromagnetism and the technology based upon it have brought immense benefits to modern society and have contributed in many ways to improvements in health and working conditions. This chapter provides detailed background material on the major segments of the non-ionizing part of the electromagnetic spectrum – low frequencies, radiofrequencies (RF) and optical wavelengths [infrared (IR), visible and ultraviolet (UV)] – noting, in the context of occupational hygiene, possible adverse effects. There is also a section on lasers.

What are electromagnetic fields and radiation?

Electromagnetic radiation has been around since the birth of the universe; light is perhaps its most

familiar form. Electric and magnetic fields are part of the spectrum of electromagnetic radiation, which extends from static electric and magnetic fields, through RF and IR radiation, to X-rays (Fig. 21.1).

All electromagnetic radiation can be characterized by a frequency and a wavelength. Wavelength is the 'distance between the two consecutive crests' of an electromagnetic wave. Frequency is the number of oscillations of the wave per unit time. Frequency is measured in hertz (Hz) (Hz = 1 cycle per second). For radio waves and microwaves, frequencies are very large and the units *kilohertz* (kHz), *megahertz* (MHz) and *gigahertz* (GHz) are used. 1 kHz is equivalent to 1000 Hz, 1 MHz is equivalent to 1000 kHz and 1 GHz is equivalent to 1000 MHz. Frequency, ν , and free-space wavelength, λ , are related by the equation

$$c = \nu\lambda \quad (21.1)$$

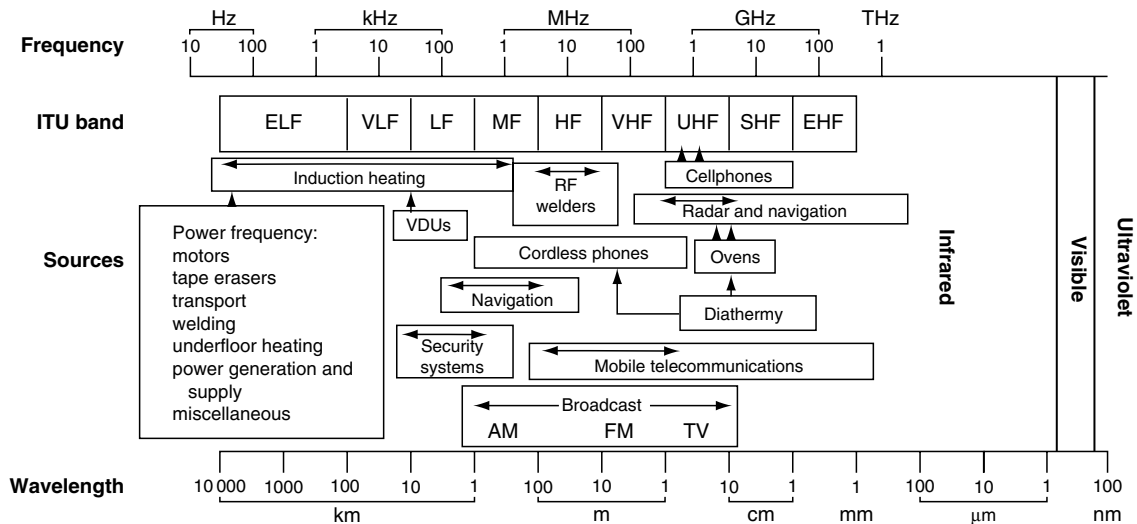


Figure 21.1 The electromagnetic spectrum covers an enormously wide range of frequencies – from 0 for static fields to 10^{22} Hz and beyond for the most energetic gamma rays.

where c is the velocity of light (3×10^8 m s⁻¹ in free space). When the electromagnetic wave is propagating through a medium that has electrical properties significantly different from free space, it will tend to be attenuated with distance and also propagate at a lower velocity than c ; because the frequency remains unchanged, Equation 21.1 predicts that the wavelength will be shortened. The wavelengths of electromagnetic fields in human tissue are significantly less than in free space, for example. At boundaries between different media, the waves undergo reflection and refraction and this can happen at tissue boundaries within the body as well as at its surface.

There is often confusion concerning the difference between electromagnetic fields, electromagnetic waves and electromagnetic radiation. Electromagnetic fields in general are non-propagating, localized around a source (such as a power line) and with spatial distributions and phase relationships determined by that source. Electric fields and magnetic fields, even though they are aspects of the same fundamental electromagnetic force, are considered separately. Under particular circumstances, electric and magnetic fields can be considered together as components of an electromagnetic wave or electromagnetic

radiation. Those circumstances are that the source is a significant fraction of a wavelength away:

$$r \geq \lambda / 2\pi \quad (21.2)$$

and that the source can be considered as a point source:

$$r \geq 2D^2 / \lambda \quad (21.3)$$

where r is the distance to the source, D is the greatest dimension of the source and λ is the wavelength. When these two equations are satisfied, under so-called *far-field* conditions, the electric and magnetic fields propagate together as a wave, the intensity (expressed in watts per square metre) of which follows an inverse-square law. The two fields are in phase and there is a constant and known relationship between their amplitudes. If the amplitude of either the electric or the magnetic field component of the electromagnetic wave is known, the other can be derived easily. The unit of electric field strength is the volt per metre (V m⁻¹) and the unit of magnetic field strength is the ampere per metre (A m⁻¹). The ratio of the electric field strength E to magnetic field strength H in a free-space electromagnetic wave has the dimensions of impedance and is known as the *impedance of free space*. It has a numerical value

of approximately 377 ohms (Ω). The intensity of the wave, S , in watts per square metre (W m^{-2}), is given by the product of the electric and magnetic field strengths:

$$S = EH = H^2/377 = 377E^2 \quad (21.4)$$

In contrast, the localized, non-propagating *near-field* electromagnetic fields close to sources have no fixed phase or amplitude relationships, and can exhibit strong spatial variability. As an example, the wavelength of the electric and magnetic fields from a European power line is 6000 km; practically, Equation 21.2 will not be satisfied and the electric and magnetic fields from the power line are considered as entirely separate entities. For a given line geometry, the electric field is related to the line voltage and the magnetic field to the current it carries.

In addition to the field/wave dichotomy, the well-known wave particle duality of the nature of electromagnetic radiation must be considered. Any propagating electromagnetic wave will have a photon energy, and a photon of any energy can be considered instead as a quantized electromagnetic wave. The Planck constant, h ($6.63 \times 10^{-34} \text{ J s}^{-1}$), relates the photon energy E to the wavelength:

$$E = h\nu \quad (21.5)$$

It is conventional to describe radio waves by frequency, or occasionally wavelength. Microwaves (radio waves with frequencies above 300 MHz) are described by either wavelength or frequency. IR, visible and UV radiations are described by wavelength, and ionizing radiations by photon energy. The definition of an ionizing radiation is a radiation with sufficient photon energy to cause ionization damage in tissues. The division between non-ionizing and ionizing radiation is defined as a quantum energy of around 11 electron volts (frequency of 3×10^{15} Hz, wavelength 100 nm) – in the UV part of the spectrum. Low-frequency electric and magnetic fields, radio waves, IR and visible radiations cannot directly damage genetic material by ionization, and so cause cancer, unlike higher energy UV, X- and gamma radiations. For example, mobile phone frequencies have wavelengths of around 30 cm and photon energies less than one-millionth of

those needed to cause ionization. Photon energies at power distribution frequencies are a million times lower again.

The International Telecommunications Union (ITU) has divided the radio spectrum into decade ranges as shown in Table 21.1.

Low frequencies

The definition of ‘low frequencies’ is to some extent arbitrary but this part of the spectrum is generally considered to extend from 0 (static fields) to about 100 kHz. The upper bound is chosen for two reasons. The first is that exposures below it are substantially in the near field: there are very few situations in which electric and magnetic fields can be treated together as components of a wave. The second reason is that above 100 kHz there is a change in the established effects of exposure as the thermal consequences of power absorption in the body begin to become more important than electrical effects on the nervous system.

Sources of static magnetic fields include the Earth, permanent magnets, magnetic resonance imaging (MRI) equipment, electrolytic processes using direct currents, some railway traction systems and the electromagnets used in guiding beams of nuclear particles. Static electric fields arise wherever there is an accumulation of electric charge, for example in thunder clouds, on synthetic fibres, on visual display unit (VDU) and TV screens, and near high-voltage DC (direct current) power systems.

Fields that oscillate at power frequencies – sometimes called ELF (extremely low frequency) (see Table 21.1) – are found wherever electricity is supplied and used. For example, they are produced by appliances, office equipment, electrical machinery, supply wiring, power lines and some electric railways. The magnetic field arises from the alternating currents (AC), whereas the electric field arises from the alternating voltage used. The frequency is usually 50 Hz, although 60 Hz is used in North America, parts of Japan and some other areas. Some European railways use $16\frac{2}{3}$ Hz, whereas aircraft power systems usually operate at 400 Hz. Higher frequency fields (tens of kilohertz)

Table 21.1 Ranges and selected uses of low and radiofrequencies.

<i>Frequency range</i>	<i>Uses</i>
30–300 Hz, ELF (extremely low frequency)	Electric power systems, railways, industrial processes, metal melting, motors, appliances
300–3000 Hz, VF (voice frequency)	Electric furnaces, induction heating, hardening, soldering, melting, refining, shop and library security systems
3–30 kHz, VLF (very low frequency)	Very long-range communications, radio navigation, induction heating, hardening, melting
30–300 kHz, LF (low frequency)	Radionavigation, radiolocation, electro-erosion, induction heating and melting, power inverters
0.3–3 MHz, MF (medium frequency)	AM broadcasting, marine radiotelephone, radionavigation, RF welding, industrial RF equipment
3–30 MHz, HF (high frequency)	Short-wave, citizens, and amateur radio, medical diathermy, MRI, dielectric heating, wood drying and gluing, RF sputtering and vacuum deposition, radiofrequency identification and access control systems
30–300 MHz, VHF (very high frequency)	FM broadcasting, police, fire, air traffic control, MRI, plastic welding, food processing, plasma heating
0.3–3 GHz, UHF (ultra high frequency)	TV broadcasting, mobile phones, microwave communications, mobile radio, radar, Bluetooth, wireless local area networks, medical diathermy, cooking
3–30 GHz, SHF (super high frequency)	Radar, satellite communications, microwave relays, anti-intruder alarms, wireless networking
30–300 GHz, EHF (extremely high frequency)	Radar, radionavigation, satellite communications, microwave relays

MRI, magnetic resonance imaging; RF, radiofrequency.

are produced by industrial equipment, such as in some metal heating and melting plants, some security and anti-theft systems and the line-scanning circuits of cathode-ray tube VDUs (these VDUs also produce ELF fields from their frame-scanning and power circuits).

Field waveforms are not always sinusoidal because of components at harmonics of the fundamental frequency in the current or voltage sources and because power control is sometimes achieved by ‘switching out’ part of each waveform, giving rise to higher frequency components. It may be necessary to analyse the field in terms of its harmonics or, for pulsed or transient waveforms, its emission spectrum. This is because exposure guidelines vary with frequency in such a way that the harmonic content can be a more significant contributor to exposure than the fundamental frequency.

As discussed above, electric fields are expressed in volts per metre (V m^{-1}) and magnetic fields in amperes per metre (A m^{-1}). However, for low-

frequency magnetic fields that are not components of an electromagnetic wave, it is conventional to describe exposures in terms of the magnetic flux density, in tesla (T). The tesla is a large unit, hence the milli, micro and nano sub-multiples (mT, μT and nT) are often used. The free-space relationship between these magnetic units is: $1 \text{ mT} = 796 \text{ A m}^{-1}$. For alternating fields, the root-mean-square (r.m.s.) value is almost always used, which, for sinusoidal waves, is the peak amplitude (measured from the zero line) divided by $\sqrt{2}$.

The field strengths encountered cover a very wide range. At 50 Hz for example, they range from about $1\text{--}100 \text{ V m}^{-1}$ and $0.01\text{--}1 \mu\text{T}$ in an office away from equipment, up to at least 10 kV m^{-1} and 1 mT near heavy electrical plant. In fact the field depends critically on the distance from the source – it falls off rapidly as one moves away – and this is illustrated in Fig. 21.2 in terms of a magnetic field. Electric fields also decay rapidly with distance, but the behaviour is more complex because of the presence of nearby objects and the ground.

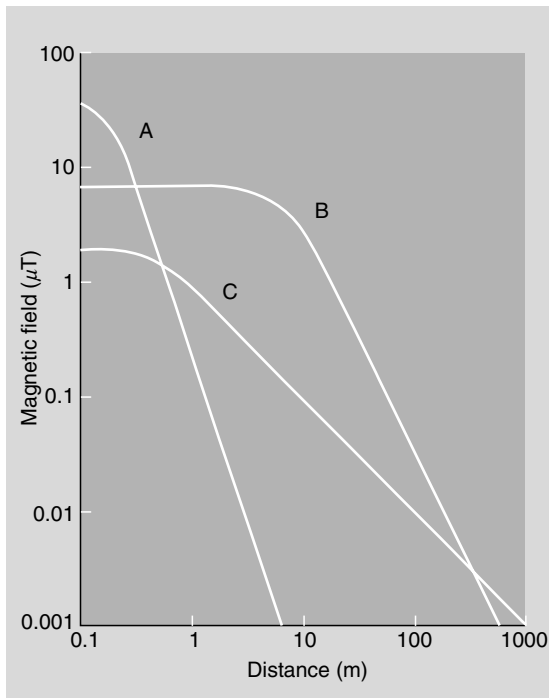


Figure 21.2 The magnetic field falls quite rapidly as one moves away from the source. Graph A is for a compact source such as an electric motor, B is for an extended source with equal ‘go’ and ‘return’ currents such as a power line, and C is for a single long current-carrying conductor. For each, the field is that along a path passing close to the source at a distance roughly equal to the centre of the curved portion of the respective graph. Well away from the source, the variation is $1/d^3$ for A, $1/d^2$ for B and $1/d$ for C, where d is the distance from the source. At the larger distances, therefore, there is likely to be a ‘background’ field composed of the various contributions from single currents – net currents in electrical cables for example.

Physical interactions and established physiological effects

Static electric fields

Static electric fields exert weak attractive forces by inducing charges on the surface of objects. If an object in the field is not well-connected electrically to the Earth, it will take on a voltage with respect to the Earth. Strong electric fields may sometimes be perceived through a tingling of exposed skin. Occasionally, small discharges between the edges of clothing or spectacle frames

and the skin may be experienced. These effects can be annoying but they do not have any lasting physiological consequences. Individual sensitivity varies greatly, but it is rare to reach the threshold for annoyance.

Low-frequency electric fields do not penetrate the body significantly but they do build up a charge on its surface. As a result, electric currents flow from the skin, through the body to the ground (earth). In an alternating electric field (AC current), the currents flowing in the body change direction as the surface of the body builds up a charge on it that is alternately positive and negative. In large alternating electric fields, for example beneath power lines, some people can feel the alternating charge when the hair on their body begins to vibrate. This is not harmful, but it can be annoying and it can be stressful if it happens often.

The voltage induced on a person by an electric field is likely to be different from that induced on some nearby object – for example, a steel structure or a vehicle – so that a small spark may occur at the instant of touching the object. Such spark discharges are akin to those experienced when touching a metal filing cabinet after walking across a synthetic carpet in dry conditions. If contact with the object is maintained, a current continues to flow. In practice, such currents are unlikely to exceed 1 milliamper (mA) at power frequencies, unless the field is particularly strong and the object large. Induced or contact currents may be felt if they are strong enough to stimulate nerve or muscle cells. The threshold for this is lowest in the range 10–1000 Hz at about 1 A m^{-2} , a level not reached by induction in any reasonable field. Earthing of the object, or other measures to equalize the potentials, generally removes any difficulties that could arise from such situations.

Static magnetic fields

Static magnetic fields exert forces on magnetizable materials. There is some debate about the possible presence of very small amounts of magnetic material inside certain cells of the human body, but in general tissues are not considered to be directly

affected by purely static fields except where the fields are high enough to cause paramagnetic or diamagnetic effects on non-magnetic materials such as blood or the oxygen in the lungs.

The primary interaction arises from *movement* of the body within a static field. This can give rise to circulating currents within tissues – a phenomenon that is discussed below.

Low-frequency magnetic fields can easily penetrate the body causing circulating currents to flow within it (Fig. 21.3). These currents do not necessarily flow to ground. If sufficiently large, the currents could cause stimulation of nerves and muscles and may affect other more subtle biological processes such as learning and memory.

The illusion of weak, flickering lights (magnetophosphenes) can result from stimulation of the retina in the eyes. This occurs from exposure to intense fields and is only found in a few occupational settings such as induction heating or arc welding. In even more intense fields, found in a few experimental and clinical situations, for example MRI, the induced currents may be large enough to cause muscles to contract and twitch.

Possible effects on health from 'low-level' exposures

For some years, the question has been asked whether or not there may be more subtle effects, including harmful effects on health, at the much lower levels of induced current and fields to which persons are ordinarily exposed in daily life.

In the 1960s and 1970s, the main concerns involving power lines were to do with their aesthetic impact, their interference with radio and television reception, and problems with noise and perception. In 1972, a variety of non-specific complaints were reported in a group of Soviet electricity substation workers exposed to very high electromagnetic fields. However, these studies were not widely disseminated until their translation into English a decade later. Meanwhile, in 1976 a report was published describing how the staff at the United States Embassy in Moscow had been irradiated with low-level microwaves by the Soviet authorities. The wide-scale survey of the possible health implications for staff that followed generated widespread public concern and interest in the subject of EMF. There had been research in the field of EMF and health before this, often under-

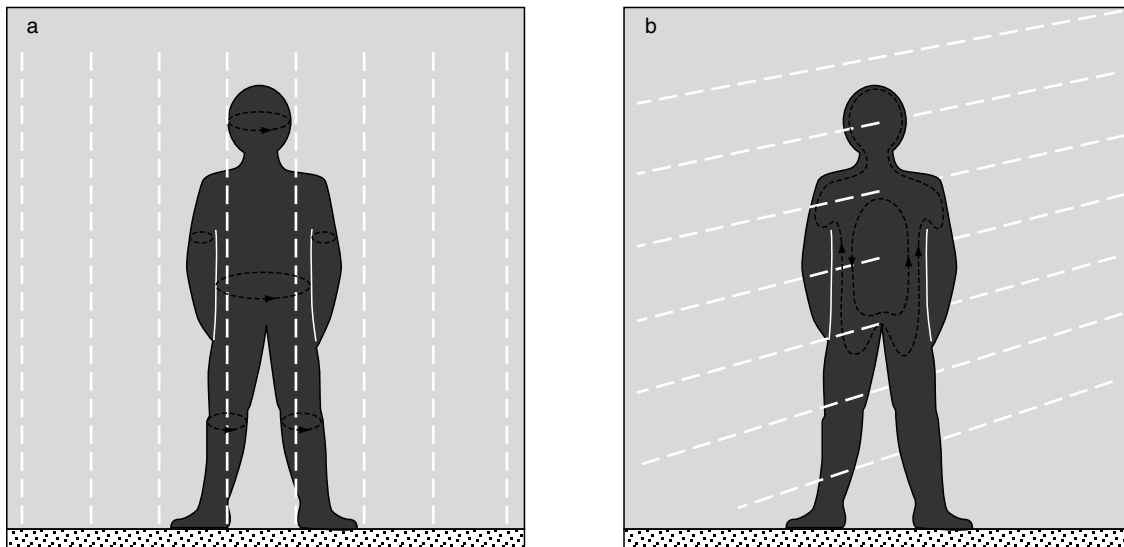


Figure 21.3 Small imperceptible circulating currents (signified by the dotted lines) are induced in a person when they are in an alternating magnetic field [dashed lines: (a) vertical field, (b) horizontal field]. The current flows in loops perpendicular to the direction of the field.

taken by the military, and some countries had exposure guidelines, but generally there was little public interest in the subject. In 1979, concern regarding the link between cancer and public exposures to magnetic fields arose because of a study on the incidence of childhood cancer in Denver by Ed Wertheimer and Nancy Leeper. This study, and the public and media interest that it generated, stimulated most of the scientific research that followed. Since then, many human health studies have been published around the world. Some have found an increased incidence of illness close to power lines, many have not.

The available scientific literature has been evaluated as part of the World Health Organization's International Electromagnetic Fields Project. It was concluded that although some gaps in knowledge about biological effects exist and need further research, current evidence does not prove that there are health consequences from exposure to low-level EMF, although results from a large UK study of childhood cancer indicate that amongst the very highest exposed children there might be some evidence of increased risk.

Although most research into the health effects of power lines has focused on the magnetic fields that the lines produce, there has also been interest in the possible direct and indirect effects of electric fields.

For persons fitted with active, implantable devices such as pacemakers, some caution may be needed in a few situations. Although these devices are designed to cope with electrical interference, strong fields may occasionally affect their operation. In a particular case, advice should be sought from the manufacturer of the device and from those responsible for implanting it.

Control

Electric fields are fairly easily reduced or screened, either near the source or near the person, by arrangements of metallic wires, mesh or sheeting. In extreme cases, such as when working directly on live high-voltage power lines, conducting suits are used. Magnetic fields in the low-frequency range are more difficult to control, sometimes requiring impracticably thick sheets of steel or aluminium, at least at power frequencies, to obtain a significant

reduction. A better approach is to arrange for all 'go' and 'return' currents to be as close together as possible (to increase field cancellation) and as distant from the working location as feasible. For electrical machines and transformers, good design can ensure that the leakage field is small. Occasionally, active screening systems are used in which an additional cancelling field is generated.

Measurements

The measurement of static electric fields requires quite sophisticated equipment and, as with alternating electric fields, considerable care since the field is so easily perturbed by the presence of nearby objects, including the person using the instrument. Most low-frequency electric-field meters consist essentially of two plates and a device to measure the capacitively induced current between them.

A variety of flux-gate magnetometers and Hall effect devices are available for measuring static magnetic fields and some of these have a response up to several hundred hertz and beyond. The simplest and most common meters for low-frequency alternating magnetic fields are based on measuring the voltage induced in a coil, which must be oriented so that its axis is parallel to the field. To avoid this necessity, devices are available which contain three orthogonal coils with electronic circuits to sum the three signals correctly. Such devices may also be conveniently employed to measure the 'elliptically polarized' fields produced by electrical systems using three-phase currents. Magnetic field meters, unlike electric field meters, are substantially undisturbed by the presence of the human body.

It is important to check that any meter that is used for a magnetic field is not adversely affected by the presence of an electric field (and vice versa), or by any radiofrequency fields that may also be present. Fields often vary markedly from point to point, so it is necessary to decide which is the most appropriate location for a measurement or over what volume it is appropriate to average the field.

There are many commercially available instruments for measuring human exposures to static and time-varying electric and magnetic fields.

Many of these can be interfaced to a computer or datalogger so that the time-variability of exposure can be assessed; there are also personal power-frequency magnetic field dosimeters available for epidemiological studies. These have also been used to assess the exposures of workers in industrial environments, such as steel works and foundries.

Radiofrequencies

The term 'radiofrequency' (RF) is used here for the region 100 kHz to 300 GHz, which includes microwaves. This region encompasses frequencies with a wide range of uses. These uses may roughly

be divided into those in which the radiation is intentional and generally directed – radio and television broadcasting for example – and those in which the emissions are either unintentional or a secondary result of a primary process such as industrial heating. See Table 21.1 for some applications of radiofrequencies. The strengths of the fields used in these applications can vary over many orders of magnitude.

At the lower frequencies in this range, the electric and magnetic fields must still be considered separately but, as the frequency rises, they become increasingly coupled together until true radiation predominates, as discussed in Section 20.1. Figure 21.4 illustrates the near- and far-field radiation regions of an RF source.

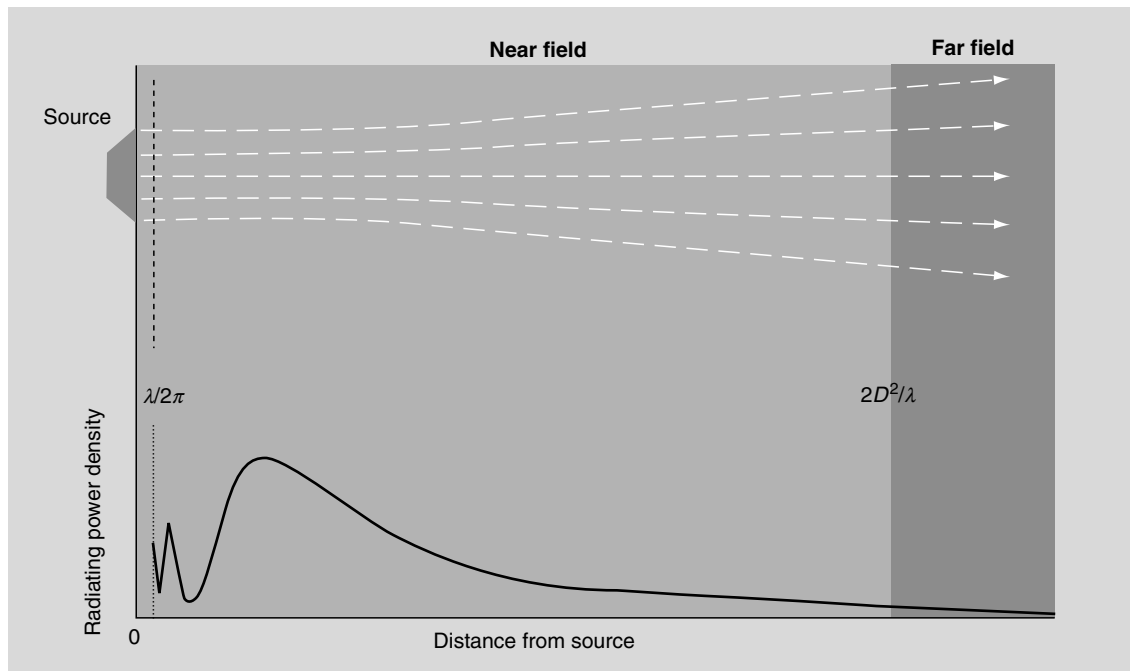


Figure 21.4 In the region close to a radiofrequency source – the near field – the radiating power density varies markedly both along the beam (as shown here for a square aperture) and across it. In the far field – beyond a distance of about $2D^2/\lambda$ from the source (λ is the wavelength and D is the largest dimension of the source aperture, usually several times λ) – the pattern is smooth, becoming almost a plane wave, with the inverse square law reduction in power density established by that distance. Very close to the source, within a distance of the order of λ , there are also non-radiating reactive fields that fall off more rapidly with distance than the radiating fields, but which dominate the latter within about $\lambda/2\pi$ of the source. In the context of human exposure, these reactive fields are usually of consequence only for wavelengths greater than about 1 m. In this figure the power density and distance scales are linear.

Physical interactions and physiological effects

Above 100 kHz, the established effects of exposure to electromagnetic fields are primarily thermal. Power is absorbed in the body from the field and at high level this can lead to heating of tissues. The dosimetric quantity is the *specific absorption rate* of energy (SAR). This is the power absorbed from the field per unit mass of tissue and its unit is the watt per kilogram (W kg^{-1}). Whole-body SAR is the total power absorption divided by the mass of the body, and high whole-body SARs could lead to increases in core temperature. The field levels to which most people could be exposed are normally very low and will not cause any detectable heating or rise in body temperature. However, heat stress could result from exposure levels above guideline levels, especially if subjects are irradiated during physical activity, or in hot and humid environments. In this case, work performance would decline and accident rates increase. Prolonged exposure at very high levels could even prove fatal. For modest inputs, the thermoregulatory system of the body can adjust and raise the heat lost by increasing the blood flow to surface areas, and by increasing the evaporation of moisture. Because of this, exposures to radiofrequency radiation can be averaged over time, usually a 6-min period. Occasional, short exposures exceeding exposure guidelines are permitted as long as the 6-min average is below the guideline levels. This time-averaging cannot be applied at lower frequencies.

Low-power RF sources such as radios and mobile phones may not be capable of producing significant whole-body SARs, but there is the potential for localized heating of tissue when the devices are operated close to the body. This is avoided by ensuring that localized SAR in a particular small mass of tissue (typically 1 or 10 g) is not excessive.

As frequency rises, the absorption mechanisms of electromagnetic fields in tissue become more effective and the penetration depth falls. Below a few megahertz, the penetration depth is greater than the dimensions of the body, but above 10 GHz the penetration depth is so small that absorption occurs substantially within the skin.

The idea of SAR becomes redundant when absorption is superficial, and the dosimetric quantity above 10 GHz is often taken to be incident power density, in watts per square metre. This is consistent with the way in which the interaction with the body of IR radiation is considered.

Two other factors become important as the frequency increases: (1) when the wavelength of the field becomes similar to that of the object, various resonances can occur; and (2) reflections from neighbouring objects can give rise to complex field patterns because of interference. The resonance effect has quite profound implications for the exposure of people.

It is convenient to divide the frequency range into four regions to correspond with the major features of the power absorption processes involved (Fig. 21.5). In the first, the *subresonance range*, the absorption increases with increasing frequency, proportionally to begin with and then more rapidly because of the increasing induced currents. Thus, the possibility of significant heating arises in strong fields, particularly in narrow regions such as the ankles where the current density is high. Next is the *resonance range*, in which the body dimensions are of the order of one-quarter to one-half of a wavelength, and in which strong absorption takes place, with a marked peak when the electric field component of the radiation is aligned with the major axis of the body. Partial body resonances, such as in the head, also occur. Then follows the *hot-spot range*, when local heating is more affected by the variation in the electrical properties of the body from point to point, and by shape and refraction effects. Finally, one reaches the *surface absorption range*, in which the fields penetrate only a few millimetres with the consequence that the heating is largely superficial.

Possible effects on health from 'low-level' exposures

The most common sources of exposure to RF are mobile phones and base stations, and there has been some public concern about possible health effects of exposure to their signals, even though the exposure levels are below national and international guidelines.

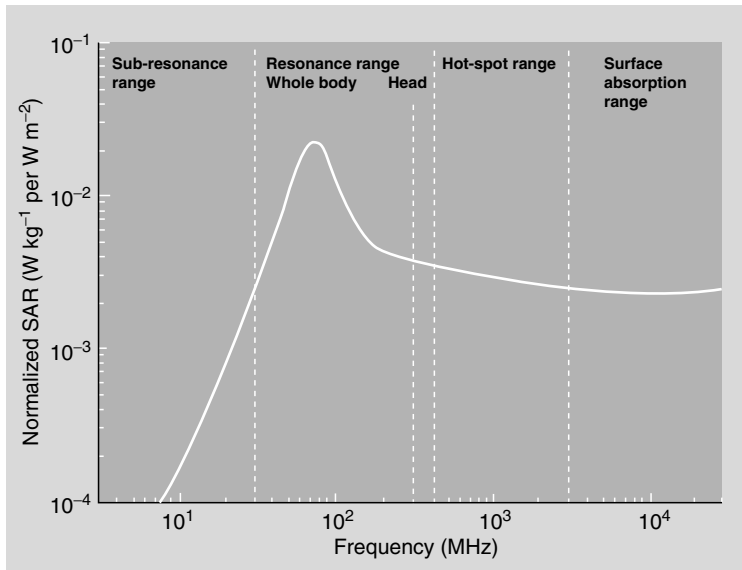


Figure 21.5 The absorption of radio-frequency energy in a person needs to be considered in terms of several frequency ranges corresponding to the different absorption processes involved. A typical variation of the specific energy absorption rate (SAR) with frequency is shown here. The peak in the resonance range occurs when the electric field is parallel to the long dimension of the body.

There is a substantial body of evidence about the health effects of microwaves and other RF electromagnetic fields, but most studies have examined the results of short-term, whole-body exposure to RF fields at levels far higher than those normally associated with wireless communications. Comparatively few studies have addressed the consequences of localized exposures to RF fields to the head.

There have been a few health studies of people who live close to radio and TV transmitters. These operate at similar frequencies to mobile phone base stations, although the signals themselves are different: radio and TV signals are emitted continuously, whereas the signals from mobile phones and base stations are emitted in short bursts (pulse modulated). Although some of these studies have found increased incidence of disease close to radio and TV transmitters, others have not, and the balance of evidence does not indicate a public health risk.

National and international *expert groups* have provided reviews of the scientific literature that are a useful resource for the occupational hygiene professional. These reviews summarize the scientific evidence relating to low-level exposure and produce independent, authoritative overviews of their significance. Both the International Commission

on Non-Ionizing Radiation Protection (ICNIRP) and the World Health Organization's International EMF Project have produced excellent reviews and recommendations. European national authorities who have completed EMF reviews include those in the UK, France, the Netherlands, Germany, Scandinavia and Canada. In the USA, Congress asked the National Academy of Sciences to establish an expert committee to review the possible health effects of electric and magnetic fields. In 1996 this committee concluded: 'the current body of evidence does not show that exposure to (electric and magnetic) fields presents a human-health hazard'.

In the UK, an independent enquiry (the Independent Expert Group on Mobile Phones, or Stewart Group) was set up to examine the implications of the operation of mobile phone systems for public health. The Expert Group's report made helpful recommendations on measures to address public concern about the health effects of mobile telecommunication technologies. Importantly, it provided information to help the consumer to make informed choices concerning personal and family use of these technologies. In its foreword the report states that 'the balance of evidence does not suggest that mobile phone technologies put the health of the general population of the UK at risk'.

However, because gaps remain in scientific knowledge, the Expert Group felt that the possibility of harm cannot yet be ruled out with confidence. The report proposed that precautionary approaches should be adopted until more robust scientific information becomes available.

The operation of some cardiac pacemakers and other active, implantable devices may be affected by radiofrequencies below 'thermal' levels, but this is primarily an electromagnetic compatibility problem and is being addressed progressively by the manufacturers.

Control

It is comparatively easy to install shielding around industrial heating equipment to lower electric and magnetic field emissions. However, it is a specialist task, and if incorrectly installed the result may be large circulating currents in the shielding. These currents can actually lead to an increase in magnetic field exposure and may also contribute to contact currents and the risk of RF burn.

Because exposures can be time-averaged, administrative and access controls can be very effective in ensuring that exposure guidelines are met; it is possible to enter high-field regions for short times, although working practices would need to be defined sufficiently tightly that overexposure could not occur. In certain cases it is necessary to exclude personnel from high-field regions. To achieve effective control, measurements need to be made regularly in areas where exposure levels could approach or exceed guidelines, and personnel need to be trained in correct work practices.

Broadcast and communications sources are unique in that the antenna is designed to produce high levels of RF. Proper access control should ensure that exposure close to the front of the antennae cannot occur, but there can be stray fields from feeder cables. Perhaps the biggest difficulty is that many RF broadcast sites are shared between many users and can be congested with sources whose status and characteristics may not be known.

Measurements

Equipment for measuring radiofrequency radiation is broadly of two types. In the first, the field

(usually the electric field) is sensed and the result displayed in terms of power density assuming plane-wave conditions. In the other, power is absorbed and the resulting temperature rise is determined by a thermistor or thermocouple and displayed as a power density. Often, three orthogonal sensors (e.g. three small dipoles) are incorporated to give an isotropic response, thus eliminating the need to orientate the device according to the direction and polarization of the incident wave. For near-field measurements, instruments that specifically determine the electric and/or magnetic field are required. Most instruments respond over a fairly broad band of frequencies, which is convenient, but care is needed if it is necessary to distinguish between more than one source. Some devices are easily overloaded and even damaged so frequent checks on correct operation as well as calibration are advisable. For pulsed and modulated sources, consideration should be given to appropriate averaging times.

There is a wide range of commercially available RF hazard probes, which, with the above provisos, are quite simple to use.

Exposure and emission standards for low-frequency and RF electromagnetic fields

There are some national standards for exposure to electromagnetic fields and radiation, but many countries now use the exposure guidelines of the ICNIRP (ICNIRP, 1998) or the Institute of Electrical and Electronics Engineers (IEEE, 1992, 2002). The ICNIRP guidelines for public exposure have been incorporated into a Recommendation of the European Commission, which has been agreed by Member States (European Council Recommendation, 1999). A European Directive limiting occupational exposure was published in 2004 (EU, 2004). It is based on the ICNIRP occupational guidelines.

These standards provide an explicit rationale for restricting exposure, focusing on established interactions. At low frequencies, restrictions are set on induced currents or internal electric field strength,

a factor of 10 to 100 below the levels corresponding to nerve and muscle stimulation. At high frequencies, restrictions on whole-body and localized SAR are set to avoid the hazards associated with increases in core temperature or localized tissue heating.

Dosimetric quantities such as induced current or SAR are difficult or impossible to measure in people. It is possible to use simple physical models of the human body (phantoms) to assess compliance with exposure guidelines, and this approach is used in the assessment of exposure from mobile phones. However, the use of phantoms is primarily a laboratory technique with little practical application in occupational hygiene. To facilitate their application, exposure standards include a set of field strengths (called *reference levels* by ICNIRP), which is derived from the SAR and current restrictions using conservative dosimetric models. As long as the reference levels are met, the underlying restrictions on SAR and current will also be met, even under worst-case exposure conditions. Under less-than-worst-case conditions, it may be possible to exceed the reference levels as long as compliance with the underlying restrictions can be demonstrated directly.

In addition to the human exposure guidelines of ICNIRP and IEEE, there are also *product emission standards* produced by the European Committee for Electrotechnical Standardization (CENELEC) and the Electrotechnical Standardization Commission (International Electrotechnical Commission, IEC). Published CENELEC standards at the time of writing cover domestic appliances, mobile phones, base stations and anti-theft devices. Standards under development cover industrial heating, broadcast, studio devices (radiomicrophones etc.) and welding. There is also a Generic Standard that allows assessment of any product that does not have its own dedicated product standard. IEC will publish similar standards on an international basis in the future. The CENELEC and IEC standards are intended primarily for use by employers and by manufacturers and operators of equipment, and they are effectively measurement protocols for specific devices and situations. They are intended to be used in conjunction with a human exposure guideline such as the European Recommendation or the ICNIRP or IEEE guidelines.

Optical radiation

On the optical part of the electromagnetic spectrum, the radiation is usually characterized by its wavelength. This is partly because the wavelength is easier to determine than the frequency, but also because some important optical processes, such as diffraction, depend critically on the relationship between the wavelength and the size of the features of an illuminated object. The optical range is customarily taken to run from a wavelength of 1 mm (the top of the radiofrequency range) through the IR, visible and UV regions to a wavelength of about 10 nm, overlapping the bottom of the soft X-ray region. However, the region from 100 to 10 nm is generally taken to belong to the ionizing part of the spectrum.

Both the IR and UV regions may be subdivided in various ways, but the most appropriate here is that defined by the International Commission on Illumination (CIE) and based broadly on the biological effects of the radiation. The various ranges are given in Table 21.2. The Earth's atmosphere absorbs all incident solar UV radiation that is shorter than 280 nm and, in addition, absorption in the ozone layer limits the amount of UVB that reaches the Earth's surface. Below about 180 nm, propagation of UV radiation in air is not possible because of strong absorption.

All bodies emit optical radiation over quite a wide spread of wavelengths, depending on their temperature. For a so-called 'black body' (a fair

Table 21.2 The principal wavelength regions of optical radiation as defined by the International Commission on Illumination.

Region	Wavelength
UVC	100–280 nm
UVB	280–315 nm
UVA	315 to 380–400 nm
Light*	380–400 to 760–780 nm
IRA	760–780 to 1400 nm
IRB	1.4–3.0 μm
IRC	3.0 μm to 1 mm

*The limits for the human eye vary among individuals over the ranges indicated.

approximation for many hot bodies), the peak emission occurs at a wavelength (λ in micrometres) given by Wien's law:

$$\lambda_{\max} = 2900/T \quad (21.6)$$

where T is the absolute temperature. Thus, a relatively high temperature is needed to achieve significant optical output. For a conventional tungsten-halogen lamp, for example, $T \sim 2800$ K, so the peak is in the near IR, although some tungsten-halogen lamps can emit biologically significant amounts of UV radiation. Even for the sun ($T \sim 6000$ K), by far the most important source of optical radiation for mankind from a hygiene point of view, about one-half of the radiation reaching the Earth's atmosphere is still in the IR.

There are also sources whose output tends to be concentrated in narrow spectral regions. These use an electrical discharge in a gas or vapour. A fluorescent lamp is a typical example, in which mercury vapour produces a strong emission in the UV at 253.7 nm (4.9 electronvolts, eV). Most of this is absorbed by the phosphor coating, which then fluoresces at several wavelengths in the visible range. Multicomponent phosphors are needed to give a good approximation to white light. Sodium lamps radiate directly in the visible yellow region at 589 nm and, by increasing the pressure of the vapour, the energy levels are broadened, thereby broadening the range of the wavelengths produced and improving the colour.

Some indication of the range of intense sources that may be encountered is given in Table 21.3. Lasers are considered in a later section. It should be noted that when short-wavelength (less than about 250 nm) UV radiation is transmitted through air, ozone is produced.

Physical interactions

The way optical radiation is absorbed depends on the wavelength; in the far IR (wavelengths greater than a few tens of micrometres), the electric field of the radiation interacts with induced or residual electric dipole moments of molecules to increase rotational oscillations, whereas at the shorter IR wavelengths, the action is to increase internal motions such as flexing and stretching vibrations.

Table 21.3 Typical sources of optical radiation.

Lamps	Incandescent, including tungsten-halogen Low-pressure gas discharge Fluorescent Low-pressure sodium 'Blacklight' Mercury vapour High-pressure sodium Metal halide Xenon
Industrial processes	Arc and gas welding Hot and molten metal Glassworking Electrical discharges
Natural	Solar

In all of these cases, the energy absorbed is progressively shared with adjacent molecules and amongst all the modes of oscillation and motion. This is manifest as a rise in temperature of the absorbing material. Thermal effects are thus to be found throughout the IR and into the red end of the visible region. Apart from transmission to the back of the eye, the absorption of optical radiation by the human body is almost entirely superficial. Therefore, local cooling by thermal diffusion and by nearby blood flow are important in determining whether the temperature reached is high enough to cause damage. At shorter wavelengths, beginning towards the blue end of the visible region, but particularly in the UV regions, the energy absorbed may raise electrons to higher energy states within the molecules. This can lead to photochemical reactions in which the energy alters chemical bonding.

Quantities frequently used in characterizing optical radiation include irradiance (the power per unit area incident on a target surface), radiant exposure (the energy per unit area on the target) and, for extended sources, radiance (the power emitted per unit area per unit solid angle by a source). The corresponding quantities expressed per unit wavelength are used in more detailed spectral analyses. These can be multiplied by a spectral weighting function, variously known as a 'response', 'action' or 'hazard' function, which

characterizes the relative variation of the biological response with wavelength, to obtain a measure of the actual biological response to be expected from the radiation in question.

Biological effects

There are many vital and life-sustaining benefits of optical radiation that should not be forgotten, but here the focus is on the potentially adverse effects that may arise from excessive exposure to certain sources. Many sources radiate over a broad band of wavelengths and any hazards they pose need to be considered for each region of the spectrum separately and also collectively to allow for any synergistic effects.

For prolonged exposure to far IR radiation, particularly in the B and C regions, there is a possibility of thermal stress when the temperature-regulating system of the body can no longer maintain adequate control. However, most attention

with regard to optical radiation needs to be given to possible direct effects on the eye and the skin. The main effects that may occur are summarized in Fig. 21.6 in terms of the relevant spectral regions.

The eye

In the far IR (IRC), corneal burn is the only effect that may arise as the cornea is essentially opaque in this region, thus shielding the rest of the eye.

The natural tear film provides an important cooling mechanism at the corneal surface, so those persons who suffer from 'dry-eye' are at a disadvantage in the presence of strong IRC sources. In the IRA and part of the IRB ($< 2 \mu\text{m}$) regions, the corneal absorption is lower so that significant absorption occurs in the lens and adjacent material. Prolonged exposure in this region may lead, years later, to enhanced lenticular opacities – commonly referred to as 'cataracts'

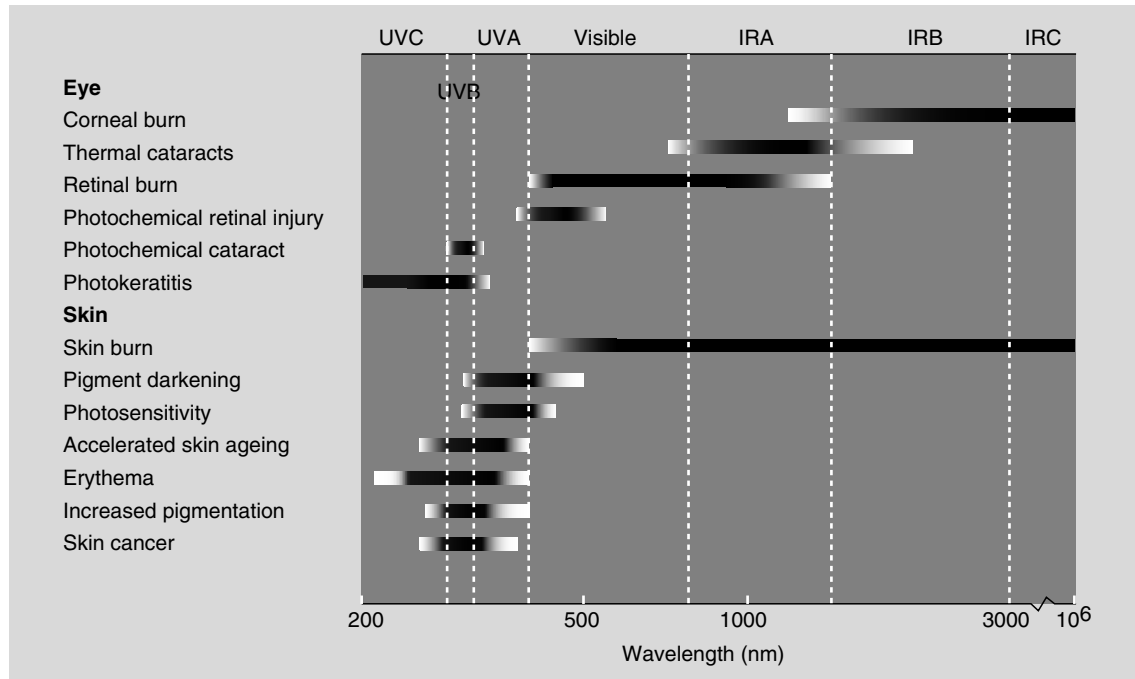


Figure 21.6 The predominant biological effects of optical radiation are listed here according to the wavelengths at which they may occur. The first three effects for the eye are thermal in origin and the shift of them to progressively shorter wavelengths arises from the diminishing absorption with reducing wavelength of first the cornea and then the lens.

and first recognized as a hazard for glass-blowers many years ago. Improved working conditions have apparently eliminated the incidence of this affliction in this occupation and there is little evidence that more modern sources, such as welding arcs, are a cause of cataracts.

In the IRA and visible regions, the radiation reaches the retina and is usually focused on it, giving a much increased irradiance. Therefore, an intense source can lead to retinal damage through local thermal damage. At the blue end of the spectrum, there is also the possibility of photochemical injury to the retina. However, the normal reflex reactions to a bright source (blinking, eye movement and turning of the head) provide good protection, except for intense laser sources. Of course, vision itself depends on photochemical reactions in the rod and cone photoreceptors of the retina.

Data from animal studies show that UVB wavelengths in the range of 295–315 nm can induce both temporary and permanent lenticular opacities, the latter only at a radiant exposure of about 5000 J m^{-2} and above. Human data are sparse. More significantly, photokeratitis (inflammation and damage to the surface of the cornea) and conjunctivitis may occur throughout the UV region, but especially in the UVB and UVC ranges, with a threshold of about 30 J m^{-2} . ‘Snow-blindness’ (from the strong, scattered blue and UV radiation) and ‘welder’s flash’ are typical instances. For exposures just above the threshold, the effects appear some hours after the exposure, with the symptoms perhaps persisting for only a few days. Fortunately, the outermost epithelial cells of the cornea are constantly renewed on a time scale of a day or two, but deeper damage takes longer to be repaired. Effects of this type are characterized by a dose, essentially the exposure level multiplied by the exposure time. However, there is usually some lower threshold for the effect and at low levels of exposure repair can keep pace. For direct sunlight (but not necessarily for industrial sources of UV) the cornea is reasonably protected by the recessed position of the eye and the glancing incidence of the radiation. Little UV radiation reaches the retina because of absorption in the anterior part of the eye. For those persons (aphakes) who have had

a lens removed (and for younger children), there can be a retinal hazard from UVA radiation.

The skin

Thermal skin burn is possible throughout the IR and much of the visible regions at high irradiances. UVA photosensitive reactions may occur in some individuals when certain chemicals (e.g. in cosmetics or therapeutic agents) either have been applied to the skin or have been ingested and subsequently transported sufficiently close to the skin surface to be able to absorb the incident radiation. Some pigment darkening may result from exposure at the shorter wavelength end of the visible region and in the UVA region. Increased pigmentation (‘tanning’) – the production of extra melanin and the migration of melanin from deeper to shallower layers in the skin – occurs primarily as a result of UVB exposure. Much of the ageing of skin – the drying, coarsening and wrinkling – is now attributed to the effect of chronic exposure to sunlight, principally the UV fraction. Sunburn is, however, the most obvious result of excessive exposure to UV radiation, the UVB component being mainly responsible. A complex set of photochemical and biochemical reactions produces the initial reddening (erythema), usually accompanied by some tanning. This depends on the skin type (‘phototype’). Type 1 never tans, for example.

The most serious long-term consequence attributed to UV exposure is an increase in skin cancers. It is now generally accepted that solar UV radiation is a causal agent in the production of human non-melanoma skin cancers. For the less common, but more serious, malignant melanomas, epidemiological evidence indicates that high levels of exposure to solar UV, particularly at an early age, may be a contributory factor.

Standards

ICNIRP has published *Guidelines on Limits of Exposure to Broad-Band Incoherent Optical Radiation* (0.38 to $3 \mu\text{m}$) (ICNIRP, 1997) and the American Conference of Governmental Industrial Hygienists (ACGIH) includes recommended

guidelines in its annual booklet of threshold limit values. Avoidance of thermal injury to the eye and photochemical injury to the retina forms the basis of these guidelines.

For UV radiation, ICNIRP has published comprehensive guidelines (ICNIRP, 1996), similar to those of the ACGIH. When examined in detail, these various guidelines are necessarily complex because they have to take account of the different biological effects within this part of the spectrum and the often marked dependence of each effect on the wavelength of the radiation (Fig. 21.7).

Control

The *protection hierarchy* ‘engineering controls first, administrative controls and special working procedures next, and personal protective equipment last’ applies to exposure to optical radiation for which the major concern is the protection of the skin and eyes.

Emission may be controlled by using enclosures, screens and, for UV especially, appropriate absorbing glass or plastic. Some light sources, such as tungsten–halogen lamps, need additional filters if

they are used near the body. If reflective screens are used, care must be taken to ensure that the problem is not just moved to another place. In general, at distances similar to the size of the source, the irradiance is inversely proportional to the distance from the source, whereas with increasing distance there is a gradual change to an inverse square law. Exposure can often be reduced substantially, simply by being further away from the source. Outdoors, appropriate sunglasses, hats, clothing, sun blocks or broad-band sunscreens and shading should be used if prolonged exposure to sunlight is likely. Not all fabrics provide adequate shielding for UV, especially when wet.

For some tasks, personal protective equipment is essential, such as reflective clothing for extreme radiant heating, and goggles for use where there are intense unshielded light sources. Occasionally, the energy absorbed by the glass in the goggles may raise their temperature sufficiently to radiate longer wavelength heat directly to the cornea. An outer reflective coating can be advantageous in these circumstances. Helmets with simple filter windows have long been used when arc welding or plasma spraying. Now, improved designs with auto-darkening filters are available, which allow the welder still to see the work immediately prior to striking the arc. Their use also removes the necessity for constantly raising and lowering the helmet. It is necessary to ensure that the darkening speed is fast enough for the work being done. Switching times as short as 100 μ s are possible.

Measurements

Because the biological effects of optical radiation depend greatly on the wavelength, the complete assessment of a source or a location may require the use of quite sophisticated spectroradiometers to obtain the irradiance as a function of the wavelength. In many situations simpler equipment can be used, either in conjunction with filters to separate the spectral regions or directly if its response is tailored to match the appropriate ‘action spectrum’ or biological spectral response factors. The angular response of the instrument can also be an important feature. Instruments that effectively

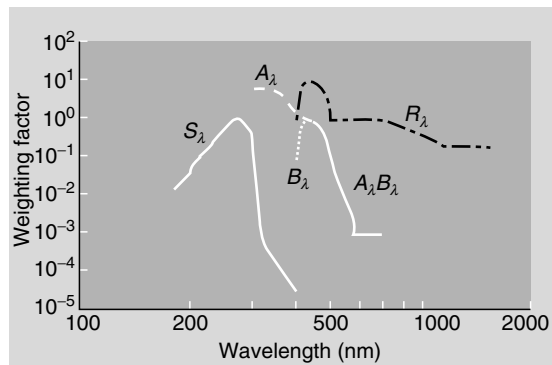


Figure 21.7 To take account of the sharply varying biological effectiveness of optical radiation with wavelength, especially for broad-band sources, various spectral weighting factors are used. S_λ is the relative spectral effectiveness of UV radiation for both the eye and the skin; R_λ is the retinal thermal hazard function; B_λ is the blue-light hazard function (for retinal photochemical injury) and A_λ is the aphakic hazard function (i.e. B_λ adjusted for persons who have had a lens removed).

integrate the received radiation over a period of time are useful in providing a measure of dose for those conditions in which this is a suitable quantity, for instance when assessing the likelihood of erythema from a UV source. The simplest of these is like an old-fashioned ionizing radiation film badge and incorporates a polysulphone film to indicate UV dose through a change in coloration, although its spectral response does not accurately match the erythemal action spectrum. Being worn means that better account is taken of the person's movements with respect to the source, which, through self-shielding for example, may influence the effective dose.

For the longer IR wavelengths, measuring devices are normally 'thermal' (e.g. thermopiles and bolometers) in that they rely on the absorption of the incident energy increasing the temperature of the sensing element. This in turn may be detected electrically to produce a signal proportional to the incident radiation power. Such instruments are inherently broad band and tend to be slower in response and not particularly sensitive. 'Photonic' detectors – e.g. semiconductor photodiodes or photomultipliers – respond to the photons of the incident radiation and can be very sensitive and fast but with narrower band widths. These can also have a non-linear spectral response, although this is often accommodated by internal processing software.

Lasers

Optical radiation usually consists of a large number of photons emitted spontaneously, and hence randomly in time, by the atoms or molecules in the source. The essence of laser action is the stimulation of the atoms or molecules so that the individual photons are emitted in phase with each other. The radiation oscillations are then to a great extent coherent in time and in space (i.e. both along and across the beam). The result is that the radiation has a very specific wavelength. Lasers are now available in most parts of the optical spectrum (Table 21.4). Laser is an acronym for light amplification by stimulated emission of radiation. In fact, most lasers act as oscillators – amplifiers with internal positive feedback – but 'losers' is hardly an appealing label.

A laser has three basic components: the 'lasing' medium, a 'pump' and an optical cavity. The medium contains particular atoms or molecules that have a set of internal energy levels such that many of the atoms or molecules can be pumped up to a higher, metastable level and then stimulated to emit energy by returning to a lower level. The medium may be a gas (e.g. a mixture of helium and neon), a liquid (e.g. ethanol containing rhodamine 6G dye) or a solid (e.g. gallium arsenide or yttrium–aluminium–garnet containing neodymium). The pump is the initial source of the energy

Table 21.4 A summary of commonly used lasers and examples of their applications.

<i>Laser</i>	<i>Emission (nm)</i>	<i>Applications</i>
Excimer	193–351	Ophthalmic surgery, material processing
Helium–cadmium	325, 442	Alignment, surveying
Dye lasers	350–1000	Instruments, dermatology
Argon ion	350, 458–514.5	Holography, retinal surgery, displays
Krypton ion	568, 647	Instruments, displays
Helium–neon	632.8	Alignment, surveying, holography
Ruby	694.3	Ranging, dermatology
Gallium arsenide	850–950	Optical fibre communications, ranging
Neodymium glass/YAG	1060	Material processing, radar/ranging, surgery
Carbon dioxide	10 600	Material processing, radar/ranging, surgery

and may be optical (e.g. a flash tube), electrical (e.g. a current through the medium) or chemical (a chemical reaction leaving the lasing molecules in the higher energy state). The optical cavity, which resonates at the wavelength to be produced, is usually formed by placing a mirror at each end of the medium. The radiation then travels back and forth many times, being amplified as it does so. One of the mirrors is only partially reflecting, thus allowing part of the energy to emerge in a well-collimated beam. This last feature means that lasers have an irradiance or radiant exposure. Some lasers operate as continuous wave (CW) sources, others are pulsed, some producing extremely short pulses of very high peak power.

Uses

In a little over 30 years, lasers have progressed from invention to widespread application in industry, business, commerce, construction, transport, medicine, research and the home, as well as use by the military. They are widely used in the entertainment world, for example laser displays at outdoor venues. Each of these exploits one or more of the key features of lasers: narrow, almost parallel beams; narrow wavelength spread; coherence and high powers and power densities. Some of the references at the end of this chapter contain more details.

Hazards

Some lasers are potentially very hazardous because of their high irradiance or radiant exposure, and the high energies that they can deliver to the body, particularly the eye, and also because of their long range. The coherence and narrow-band properties, except in so far as they contribute to the narrow beams, do not pose any new hazard. Outside the visible region, however, the absence of any associated broad-band and partly visible radiation means that the laser may be more easily overlooked. Inadvertent reflections must also be avoided, otherwise the beam may be redirected into what would have been a safe zone.

All the biological interactions of optical radiation described previously apply to laser radiation, with the obvious proviso that damage from absorption may be more severe because of extreme localization and high-energy densities. Absorption of intense, short pulses can generate thermoacoustic pressure transients, which may lead to ablation of tissue. At extremely high intensities (usually only met in sub-nanosecond pulses), the electric field of the optical radiation can interact directly with cells and molecules and cause their disruption, even in transparent material.

Classification

The hazards of lasers were recognized early, and over the years a classification system has been developed which greatly contributes to their safe use and to reducing the effort that might otherwise be needed on the part of those responsible for safety when the laser is operated. Most users will not need to make exposure measurements and can concentrate on establishing safe working practices appropriate to the classification of the laser being used. The hazard classification scheme (IEC 60825-1 Edition 1.2 2001: International Electrotechnical Commission) (IEC, 2001) is akin to an emission standard and takes account of the capabilities of the laser itself and also the protective features provided, for instance the extent to which the radiation is confined within the installation. Manufacturers must determine the class to which a particular laser belongs, label it accordingly and provide the necessary user information. The classes may be summarized as shown below.

- *Class 1* lasers are those that are inherently safe because of their low power or are safe by virtue of the protective features incorporated into the design of the product.
- *Class 1M* lasers are sources with large area or widely diverging beams (e.g. LEDs), which are potentially hazardous when viewed using an optical instrument (e.g. a magnifying optic).
- *Class 2* lasers are low-power devices that emit visible radiation; they are not intrinsically safe but eye protection is normally afforded by aversion

responses including the blink reflex. They are not capable of causing injury to the skin.

- *Class 2M* lasers are visible sources with large area or widely diverging beams (e.g. LEDs), which are potentially hazardous when viewed using an optical instrument.
- *Class 3R* allows some relaxations on the requirements for class 3B lasers.
- *Class 3B* lasers are capable of causing eye injury either because their output is invisible and therefore aversion responses are not activated or because the beam power is such that damage is done in a time shorter than the blink reflex (~ 0.25 s). Higher power lasers in this class may also cause skin burns but, for lasers other than those that emit UV radiation, it would be usual to expect that sufficient discomfort would arise from skin exposure to cause withdrawal of the exposed skin.
- *Class 4* is the most hazardous laser classification. Laser products in this class are high-power devices capable of causing an immediate injury to the eyes and skin. Exposure to diffuse reflections may even be hazardous. Class 4 lasers can also damage objects and set fire to flammable materials.

Figure 21.8 shows these classes for continuous-wave lasers. Variations in the detail of this scheme may be found in some national recommendations.

Standards

Standards defining limits of exposure to laser radiation are complex and, consequently, sometimes difficult to interpret. This is because of the wide range of biological interactions possible, their dependence on wavelength, energy, power, pulse duration, beam geometry and so on. Standards or guidelines have been drawn up by various international (namely IEC/CENELEC, ICNIRP) and national bodies.

Controls

In addition to ensuring that a laser is properly classified and that any associated requirements are met, it is advisable to consider several other matters as follows, depending on the classification.

- *Class 2*. The positioning and mounting of the laser from the point of view of the likelihood of the beam reaching a person's eye (the need for the beam to be above or below eye level); the provision of suitable warning labels.
- *Class 3R*. The requirements vary, depending on whether the beam is visible (400 to 700 nm) or not. For example, the requirements to terminate the beam and to appoint a laser safety officer apply

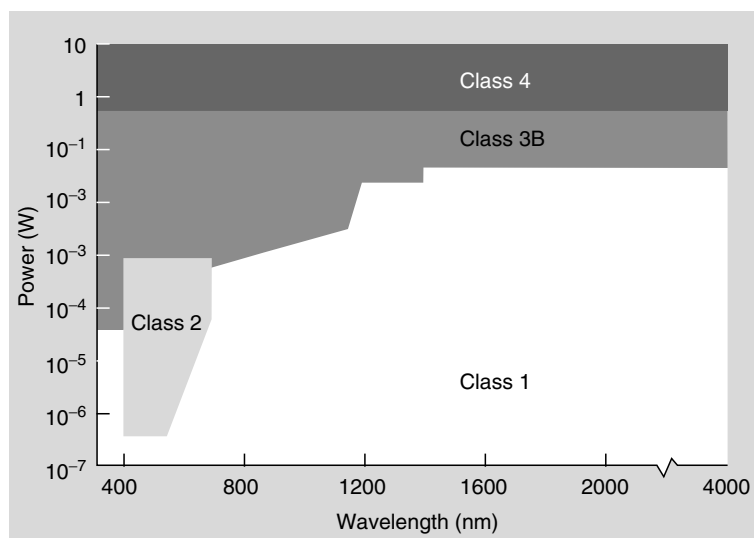


Figure 21.8 There is a well-established classification system for lasers based on the hazards which their use may entail. The details are complex but a simplified presentation for continuous-wave lasers is shown here in terms of the accessible power emission limits and the wavelength.

specifically to invisible class 3R beams, and as for class 2 above.

- *Classes 3B and 4.* Control of access to the area in which the laser is operated/special consideration needs to be given to outdoor use; the removal of specularly reflecting surfaces from near the beam path; the appointment of a laser safety officer, and as for class 2 above.

In some cases, personal protective equipment (special clothing and eyewear) may be required. If eyewear has to be used, it is important to ensure that it is acceptable and always used, it does not impair normal visibility seriously and it is adequate for the power and wavelength of the radiation in question.

There are some secondary potential hazards associated with lasers (particularly high-power devices) that should be borne in mind, such as fumes and flying debris from the target, electric shock from power supplies, possible explosions in capacitor banks and the toxic chemicals used in some lasers. These 'associated hazards' are often much more significant than the laser beam itself. By way of illustration, the reported deaths from incidents involving lasers have all been due to the associated hazards: electrocutions, endotracheal airway fires and a mechanical crush fatality.

Measurements

The measurement of laser radiation is not easy because of the wide range of conditions met and there are many possible sources of error. It is recommended that some of the references listed below are consulted. The general types of instrument needed have been outlined above but, in addition, special techniques are required to resolve the time course of the nanosecond and sub-nanosecond pulses that some lasers produce.

References

European Council Recommendation (1999). The limitation of exposure of the general public to electromagnetic fields 0 Hz to 300 GHz. *Official Journal of the European Communities*, 30, 7–99.

ICNIRP (1996). Guidelines on UV Radiation Exposure Limits. *Health Physics*, 71, 978.

ICNIRP (1997). Guidelines on limits of exposure to broadband incoherent optical radiation (0.38 to 3 μm). *Health Physics*, 73, 539–54.

ICNIRP (1998). Guidelines on limiting exposure to time-varying electric, magnetic and electromagnetic fields. *Health Physics*, 74, 494.

IEC (2001). *Safety of Laser Products, Equipment Classification, Requirements and User's Guide*. IEC60825–1, edn 1.2.

IEEE (1992). *IEEE Standard C95.1–1991 for Safety Levels with Respect to Human Exposure to Radio Frequency Electromagnetic fields, 3 kHz to 300 GHz*.

IEEE (2002). *IEEE Standard C95.6–2002 for Safety Levels with Respect to Human Exposure to Electromagnetic Fields, 0 to 3 kHz*.

European Union Directive of the European Parliament and of the European Council (2004). The minimum health and safety requirements regarding the exposure of workers to the risks arising from physical agents (electromagnetic fields). *Official Journal of the European Communities*, 159, 1–26.

Further reading

Advisory Group on Non-ionising Radiation (1994). Health effects related to the use of visual display units. *Documents of the NRPB*, 5, 1–75.

European Commission CD-GV (1996). *Public Health and Safety at Work. Non-ionizing Radiation. Sources, Exposure and Health Effects*. EC, Brussels.

ILO/IRPA-INIRC (1994). *Protection of Workers from Power Frequency Electric and Magnetic Fields*. International Labour Organization Occupational Health and Safety Series 69. ILO, Geneva.

Independent Expert Group on Mobile Phones (2000). *Mobile Phones and Health*. IEGMP, c/o NRPB, Chilton, UK.

International Non-ionizing Radiation Committee in collaboration with The International Labour Office (1993). *The Use of Lasers in the Workplace*. Occupational Safety and Health Series No. 68. International Labour Office, Geneva.

McKinlay, A.F., Harlen, F. and Whillock, M.J. (1988). *Hazards of Optical Radiation. A Guide to Sources, Uses and Safety*. Adam Hilger, Bristol.

NRPB (2002). Health Effects from Ultraviolet Radiation: Report of an Advisory Group on Non-Ionising Radiation. *Documents of NRPB*, 13, 1.

NRPB (2002). Advice on Protection Against Ultraviolet Radiation. *Documents of NRPB*, 13, 3.

Repacholi, M.H. (ed.) (1988). *Non-ionizing Radiations: Physical Characteristics, Biological Effects and Health Hazard Assessment*. IRPA, Melbourne.

Sliney, D. and Wolbarsht, M. (1980). *Safety with Lasers and Other Optical Sources. A Comprehensive Handbook*. Plenum, New York.

- Suess, M.J. and Benwell-Morison, D.A. (ed.) (1989). *Non-ionizing Radiation Protection*, European Series, no. 25. WHO Regional Publications, Copenhagen.
- US National Academy of Sciences (1997). *Possible Health Effects of Exposure to Residential Electric and Magnetic Fields*. National Academy Press, Washington, DC.
- World Health Organization (1979). *Environmental Health Criteria 14: Ultraviolet Radiation*. WHO, Geneva.
- World Health Organization (1982). *Environmental Health Criteria 23: Lasers and Optical Radiation*. WHO, Geneva.
- World Health Organization (1984). *Environmental Health Criteria 35: Extremely Low Frequency (ELF) Fields*. WHO, Geneva.
- World Health Organization (1987). *Environmental Health Criteria 69: Magnetic Fields*. WHO, Geneva.
- World Health Organization (1993). *Environmental Health Criteria 137: Electromagnetic Fields (300 Hz to 300 GHz)*. WHO, Geneva.

Chapter 22

Ionizing radiation: physics, measurement, biological effects and control

Ronald F. Clayton

Introduction

Sources of ionizing radiation

X-rays

Atomic structure

Radioactivity

Types of ionizing radiation due to radioactivity

Alpha radiation

Beta radiation

Gamma radiation

X-radiation

Bremsstrahlung radiation

Neutron radiation

Properties of ionizing radiation

Alpha radiation

Beta radiation

Gamma radiation and X-radiation

Neutron radiation

Radiological protection terminology

External radiation

Sealed sources or closed sources

Unsealed sources or open sources

Radioactive contamination

Measurement

Units of radioactivity

Units of radiation dose

Units of radiation dose rate

Instrumentation

Biological effects of radiation

Control

Legislative control

Administrative measures for controlling exposure to ionizing radiation

Practical measures for controlling exposure to ionizing radiation

Control of external exposure

Time

Distance

Shielding

Radiation monitoring

Control of internal radiation

Transport of radioactive material

Background sources of radiation

References

Further reading

Introduction

‘Ionizing’, or the alternative spelling ‘ionising’, radiation can be defined as radiant energy that produces ionization of the matter through which it passes either directly or indirectly. It is encountered both as part of the electromagnetic spectrum and as particulate matter. The term ‘radiation’ correctly applies to all forms of radiant energy: radio waves, heat, light, sound, etc. However, as this chapter is concerned only with ionizing radiation, for simplicity it will be referred to as ‘radiation’. Since the advent of nuclear weapons, nuclear energy and irradiation of foodstuffs by gamma rays, ‘radiation’ has become an emotive word with the general public, being associated with all the undesirable effects of ionizing radiations. There is

also widespread confusion between what is meant by the terms ‘radioactive’ and ‘radiation’. ‘Radioactive’ is the name given to the property of a substance that undergoes spontaneous disintegration and emits ‘radiation’ during the process of disintegration. Ionizing radiations include:

- X-rays generated by electrical devices;
- X-rays and gamma rays emitted by radioactive material;
- alpha and beta particles emitted by radioactive material;
- neutrons generated during nuclear fission and fusion processes;
- neutrons produced by spontaneous fission in some elements such as californium-252 and special sources such as americium/beryllium alloys;
- high-energy particles produced in accelerators.

Although much ionizing radiation is artificially produced, by far the greater proportion of the exposure of the general public is to naturally occurring radiations from radioactive materials in the Earth's crust, radioactive gases in the atmosphere and cosmic radiation from outer space.

To understand the production of radiation from radioactive material requires an elementary knowledge of atomic structure and the phenomenon of radioactivity. This concept will be explained in a following section.

Sources of ionizing radiation

X-rays

X-rays were discovered in 1895 by Dr Röntgen, who showed that they penetrated matter opaque to visible light and that they ionized air, enabling it to conduct electricity. They are part of the electromagnetic spectrum, having a shorter wavelength than ultraviolet radiation. Most X-rays are produced in apparatus specifically designed for that purpose, such as medical and industrial X-ray sets. They may also be produced in any electrical apparatus in which there is a heated cathode emitting electrons and a potential difference of a few thousand volts accelerating the electrons so that they bombard an anode. Most of the energy of the electrons absorbed in the anode appears as heat and a small proportion as X-rays. The energy of these X-rays lies between 60% and 70% of the accelerating voltage used in the apparatus. Thus, in an apparatus with an accelerating voltage of 50 kV (kilovolts), the X-ray energy will lie in the range 30–35 keV (kiloelectronvolts). Examples of such devices are: cathode-ray tubes (TV tubes), radio valves, valve rectifiers, electron beam welders, electron microscopes, mass spectrometers, gyrotrons and all types of accelerators. With few exceptions, any apparatus that has an accelerating potential in excess of 5 kV may be considered as having the potential to give rise to an ionizing radiation hazard.

Some X-rays may be emitted during the disintegration of radioactive material or when beta particles emitted by radioactive material are stopped in some absorbing material. These latter are a

special case and will be defined later in the chapter.

Atomic structure

As stated in the introduction, a simple explanation of the structure of the atom and nuclear disintegration is necessary to understand the phenomenon of radioactivity and the production of ionizing radiations from this source. An atom may be considered to consist of a nucleus that carries a positive electric charge of the order of 5×10^{18} coulombs per cubic centimetre ($C\text{ cm}^{-3}$), surrounded by negatively charged orbiting electrons, rather like a miniature solar system, with a sun (the nucleus) and orbiting planets (the electrons). The electrons are held in discrete orbital shells, each having its own energy level. The orbital shells are identified by letters, the innermost shell being called the 'K' shell, the next the 'L' shell, and so on. The nucleus is made up of densely packed, positively charged protons and neutrons that do not carry any electrical charge. It is very small, having a radius of about 10^{-12} cm, compared with the radii of atoms that are about 10^{-8} cm. As the nucleus contains virtually all of the mass of the atom, it must have a very high density of about 10^{14} g cm^{-3} . In an atom in its normal state, the number of orbiting electrons equals the number of nuclear protons so that the atom as a whole appears uncharged.

The neutron has the additional property of being capable of dividing into a positively charged proton and a negatively charged electron. This electron should not be confused with the orbital electrons, although they are physically similar. Other properties of the nuclear particles are summarized in Table 22.1.

The number of protons in the nucleus of a particular atom determines its atomic number (Z) and thus determines the element and its chemical nature. Examples are:

- $Z = 1$, hydrogen (H);
- $Z = 3$, helium (He);
- $Z = 24$, chromium (Cr).

The number of neutrons (N) in a nucleus is about the same as the number of protons (Z) in the case of the lighter elements (carbon-12, $Z = 6$,

Table 22.1 Properties of the nuclear particles.

Particle	Relative mass	Electrical charge
Proton	Approximately 1	Positive (+1)
Neutron	Approximately 1	Uncharged (0)
Electron	Approximately 1/1840	Negative (-)

$N = 6$; cobalt-59, $Z = 27$, $N = 32$), but somewhat greater in the case of the heavier elements such as uranium. There may be up to about 50 extra neutrons in the heaviest nuclei (uranium-238, $Z = 92$, $N = 146$; $N - Z = 54$). The number of neutrons can vary slightly between the atoms of the same element; the sum of the number of neutrons and protons in a nucleus determines the atomic weight of the element (symbol = A ; thus $Z + N = A$). The precise number of neutrons in a particular atom of an element also determines the isotope of the element or which nuclide is the nucleus. These isotopes or nuclides are not always radioactive; for example, oxygen isotopes with the atomic weights of 16, 17 and 18 are not radioactive, but oxygen-19 is. The Z number of an element is indicated by a subscript in front of the chemical symbol (thus ${}_1\text{H}$) and the atomic weight of the element by a superscript also in front of the chemical symbol (thus ${}^1\text{H}$). The two combined showing both the Z and A numbers are written thus: ${}^1_1\text{H}$. In reports and correspondence, instead of using the scientific representation as shown in the previous sentence, it is usual and acceptable to use the form cobalt-60 or carbon-14. A pseudo-scientific method is to use the chemical symbol and the isotope number, for example Co-60 or C-14.

Radioactivity

Radioactivity was discovered in the late nineteenth century by the Curies and other workers. Radioactivity may be defined as the property possessed by some atomic nuclei of disintegrating spontaneously, with the loss of energy through the emission of a charged particle such as an alpha or beta particle and electromagnetic radiation such as gamma and/or X-rays. For legislative purposes in national legislation such as the United Kingdom Radioactive Substances Act 1993 and the Ionising

Radiations Regulations 1999, radioactivity may be defined more narrowly. Radioactivity occurs when there is an imbalance of energy in the nucleus. This is usually due to a disparate proton–neutron ratio for the particular nuclide and the mass–energy relationship among the parent nucleus, the daughter nucleus and the emitted particle, which gives rise to divergent atomic packing fractions and nuclear binding energies. Virtually all nuclides with an atomic number of greater than 83 are radioactive.

Radioactive decay is a random process. Therefore, a radioactive nucleus is likely to decay at any time, however long it has existed. The probability of disintegration is not affected by external factors such as temperature, pressure, its state of chemical combination, etc. The probability per unit time of a nucleus decaying is called the ‘decay constant’ and is represented by the symbol λ . It varies enormously from one radionuclide to another. The decay constant has the dimensions of 1/time and is usually expressed in ‘ s^{-1} ’. To establish the basic formula for radioactive decay, it is assumed that at a given time there are N radioactive nuclei present and that the average number decaying per second is λN . Hence:

$$-dN/dt = -\lambda N \quad (22.1)$$

from which the basic formula for radioactive decay is derived:

$$N = N_0 e^{-\lambda t} \quad (22.2)$$

where N_0 is the number of atoms present when $t = 0$.

The mean life of the nuclei of a certain nuclide can be represented by the formula $\tau = 1/\lambda$. However, it is more usual to express the rate of radioactive decay in terms of the half-life (T). This is the time elapsing before one-half of the original radioactive nuclei have decayed. It can be derived from the formula $N = N_0 e^{-\lambda t}$ by putting:

$$\begin{aligned} N/N_0 &= 1/2 = e^{-\lambda T} \text{ therefore } \ln 2 \\ &= \lambda T, \text{ and } T = \ln 2/\lambda = 0.693/\lambda \\ &= 0.693\tau \end{aligned} \quad (22.3)$$

Thus, the mean life of the nuclei is 1/0.693 or 1.443 times as long as the half-life of the radionuclide.

Types of ionizing radiation due to radioactivity

The three basic types of ionizing radiations emitted by radioactive material are the alpha particle (α), the beta particle (β) and the gamma ray (γ). Each radioactive nuclide emits one or more of these radiations and each has an energy characteristic of the decaying nuclide. An alpha particle is composed of two protons and two neutrons, and therefore is the nucleus of a helium atom. A beta particle is essentially an electron, usually carrying a negative charge, but occasionally it may carry a positive charge and is then called a 'positron'. A positron is defined as the antiparticle of the electron and is the only antiparticle considered at present to have any significance in radiation protection or nuclear power.

Some radioactive materials may emit neutrons and some will emit X-rays. The latter are more likely to be emitted from the orbital electron shells rather than from the atomic nucleus. This occurs when an electron from the inner shell of 'K' electrons is captured by the nucleus. The excess energy arising from the loss of the electron into the nucleus appears as X-ray emissions. This form of radioactive decay is known conventionally as 'K capture'.

Neutrons are emitted from some of the heavier atoms, such as plutonium, due to spontaneous fission of the atom. Specially prepared material, such as americium/beryllium alloys in which the alpha particle emitted by the americium displaces a neutron from the beryllium atom, are available as discrete neutron sources.

Of all five types of radiation, only alpha and beta particles are directly ionizing as they carry an electrical charge. The other types of radiation do not carry an electrical charge and produce ionization by their interaction with the target material.

The energy of all types of radiation is measured in electronvolts (eV) but this being such a small quantity, the energy is more usually expressed in kilo-electronvolts (keV) or megaelectronvolts (MeV).

Alpha radiation

Alpha radiation is emitted mainly by radioisotopes of the heavy elements such as thorium, radium,

uranium and plutonium. Alpha radiation, from whatever source, consists of heavy, doubly charged particles and results in the original element being transformed into an element that is two atomic numbers (Z) lower and four atomic mass numbers (A) lower. For example, plutonium-239 ($Z = 94$, $A = 239$) emits an alpha particle and is transformed into uranium-235 ($Z = 92$, $A = 235$).

The alpha particle from each radioisotope has its own characteristic energy or energies. In most cases, all such emission is at one discrete energy level, but in some cases two or even three energy levels are detected. For example, the alpha particles from uranium-235 have energies of 4.18 and 4.56 MeV, whereas the alpha particles from uranium-238 have only one energy level at 4.2 MeV. Most alpha-emitting elements also emit either an X-ray or a gamma ray as an energy-adjusting mechanism when the newly formed atom settles from an excited state to a ground state.

Beta radiation

Beta radiation is emitted mainly by radioisotopes of the intermediate and lighter atomic weight elements, although some isotopes of the heavier elements also emit beta particles; examples are plutonium-241, uranium-237 and some of the radioactive daughters in the decay chain of the naturally occurring radioactive elements radium and thorium. A beta particle is essentially an electron, but originates from the transformation of a neutron in the nucleus into a proton and an electron and has a mass of $1/1840$ of a proton. They are not to be confused with the orbital electrons. The beta particle is usually represented by the symbol β , but sometimes by either β^+ or β^- ; these two symbols are able to differentiate the more common, negatively charged particle from the less common, positively charged particle (positron). The β^- is produced when there is a neutron excess in the originating nucleus and the β^+ when there is a neutron deficiency. Beta radiation, from whatever isotope, consists of light, singly charged particles. In the case of negative beta radiation, the original element is transformed into an element with a numerically one higher Z number but with the same A number. For example, cobalt-60

($Z = 27$, $A = 60$) emits a β^- and is transformed into nickel-60 ($Z = 28$, $A = 60$). In the case of an element that emits a positively charged beta particle, the original element is transformed into one with a lower Z number but with the same A number. For example, when zinc-63 ($Z = 30$, $A = 63$) emits a positron (β^+) it is transformed into copper-63 ($Z = 29$, $A = 63$). Beta radiation from each beta emitter has its own spectrum of energies, not a discrete energy level as in the case of alpha particles. Each spectrum, however, has a characteristic maximum, for example yttrium-90 (the daughter of strontium-90) has a maximum at 2.27 MeV, whereas phosphorus-32 has a maximum at 1.71 MeV. Most beta emitters also emit an energy-compensating gamma or X-ray.

Gamma radiation

Gamma radiation, which is part of the electromagnetic spectrum, is emitted as an accompaniment to most alpha and beta emissions; there are no 'gamma only' emitters. Gamma radiation, not being particulate and not carrying an electrical charge, does not ionize other matter directly. It only does so by its interaction with the molecular or atomic structure of the irradiated material. Its emission from a radioactive substance does not bring about any transformation in the emitting element. It occurs when the excess energy present in an atom following its transformation is released, when it settles from the excited state following the transformation into its ground state. For example, when cobalt-60 decays to nickel-60, the two gamma rays having energies of 1.33 and 1.17 MeV normally used to identify the presence of cobalt-60 are in fact emitted by the daughter nickel-60 atom settling to its ground state.

X-radiation

X-radiation is similar to gamma radiation but is less penetrating and has a longer wavelength. It is produced in radioactive material as a result of the release of excess energy when a daughter atom settles to its ground state. It may also be released when there is a disruption in the orbital shells, such

as when a 'K' shell electron is captured by the nucleus or when an electron falls from a higher energy shell to a lower energy shell.

Bremsstrahlung radiation

When beta particles are attenuated in matter, some of the energy appears as X-radiation, which in this case is termed 'bremsstrahlung' ('braking radiation', from the German 'bremsen', to brake or slow down and 'strahlung', radiation). Generally, bombardment of a material by X-rays or gamma rays does not result in the irradiated material becoming radioactive. When the energy of the radiation is in excess of about 6 MeV then some activation may occur, depending on the nature of the material being irradiated.

Neutron radiation

Neutron radiation is emitted from some fissile material when it undergoes spontaneous fission, resulting in the formation of fission products that are usually radioactive and a release of energy that may appear as heat, X-rays, gamma rays and beta rays. Examples of elements undergoing significant spontaneous fission are californium-252, plutonium-240 and curium-244. Other transuranic elements undergo spontaneous fission to a lesser degree. Neutrons are also emitted from radioactive sources specifically manufactured for this purpose, such as americium-241 alloyed with beryllium. Neutrons are also produced in nuclear reactors, some accelerators and nuclear fusion experiments.

Although not directly ionizing, they do bring about ionization by their interaction with matter. More importantly they make the irradiated material radioactive, thereby producing an extra source of ionization due to the radioactive decay of the isotopes produced.

Properties of ionizing radiation

Alpha and beta particles, because they are particulate and carry an electric charge, interact both mechanically and electrically with the atoms of the materials irradiated by them. Because they progressively lose their energy they have a finite

Table 22.2 Properties of types of ionizing radiation.

<i>Type of radiation</i>	<i>Nature and symbol</i>	<i>Relative mass</i>	<i>Electrical charge</i>	<i>Approximate energy range</i>	<i>Approximate maximum range in air (at STP)</i>
Alpha	α particulate	4	+2	2 to 8 MeV	Few centimetres
Beta	β particulate	1/1840	-1 or +1	keV to 5 MeV	Few metres
Gamma	γ electromagnetic	0	0	keV to 7 MeV	Very long range
X-rays	X electromagnetic	0	0	keV up to 20 MeV	Very long range
Neutrons	n particulate	0	0	eV to 20 MeV	Long range

range. Table 22.2 summarizes the properties of the various types of radiation discussed in this section.

Alpha radiation

An alpha particle has approximately 7360 times the mass of a beta particle (electron), twice the electric charge and travels more slowly. It loses its energy much more quickly and therefore deposits all of its energy over a very short path. It has a range of only a few centimetres in air at STP (standard temperature and pressure) and only a few micrometres in tissue. The short range and low penetrating power of an alpha particle mean that it is difficult to detect, can be shielded completely by very thin material such as paper and does not penetrate the dead, outer layers of the human skin. These properties mean that alpha particles originating from a source outside the human body do not present any radiological hazard. Because they deposit all their energy over a very short range in tissue they are extremely hazardous when released from radioactive material deposited inside the body.

Beta radiation

Beta radiation is more penetrating than alpha radiation and has a longer range in air (the range depending on the energy of the particle). The energy of beta particles varies enormously, from about 18 keV in the case of beta radiation from tritium (hydrogen-3) to about 3.6 MeV in the case of potassium-42. The energies of beta emissions from some fission products are even higher. Despite the high maximum energy of some beta particles, they can be efficiently shielded by relatively

light materials such as Perspex, aluminium and glass. Because their path in tissue is much longer than that of an alpha particle and therefore they do not deposit so much energy per unit length in the absorbing material, they do not represent such a great hazard as an alpha particle when released from radioactive material deposited inside the body. The hazard, however, is not insignificant.

A danger that must be guarded against when designing shielding for beta radiation is the production of bremsstrahlung X-rays in the material used for shielding. In the case of large beta sources with high-energy maxima, such as terabecquerel (TBq), petabecquerel (PBq) strontium-90/yttrium-90 sources, substantial thicknesses of heavy shielding may be required to attenuate the bremsstrahlung X-rays.

Gamma radiation and X-radiation

Both gamma and X-radiation are electromagnetic radiation and have short wavelengths. Being uncharged they are not deflected by magnetic fields and as a result have long ranges in air. They are very penetrating and will pass through the body, irradiating all internal organs in their path. Heavy shielding is required to attenuate them, often up to several centimetres of lead or a few metres of concrete.

Neutron radiation

Because they are uncharged, neutrons cannot be deflected by magnetic fields, but their energy is dissipated when they undergo collisions with the atoms of the irradiated material, which then becomes radioactive. Because of the manner in which

they interact with matter, they may deposit their energy in the target material over a relatively short distance and can produce significant doses of ionizing radiation. This is due first to their immediate atomic interactions in which they may produce electrons and heavy particles and then to radiation from the radioactive material produced in the target. Because they interact with targets and make them radioactive, storage and transport containers for neutron sources will become radioactive.

Radiological protection terminology

In addition to knowing the terms for describing the properties of radiation and radioactivity, persons working with ionizing radiation should become familiar with the various terms used in the field of radiological protection.

External and internal radiation hazards

The term *external radiation hazard* is used to identify radiation arising from sources outside the body, whereas the term *internal radiation hazard* is applied to sources of radiation deposited inside the body. X-ray sets can produce only external radiation hazards, even though the effects of the exposure to the radiation are on the internal organs of the body. Radioactive material produces an external radiation hazard when it is outside the body, but an internal radiation hazard when it is deposited inside the body as a result of an accident (e.g. when radioactive material is breathed in and deposited in the lungs).

Sealed sources or closed sources

These are sources that remain intact under all circumstances, whether in routine use or during emergency situations. They may be radioactive material sealed in some form of encapsulation or non-friable solid radioactive material. Before being defined as a sealed or closed source, they have to satisfy certain strict criteria such as not breaking under pressure or when dropped or when subjected to high temperatures. They must not show any sign of leaching or leakage when immersed in water for extended periods.

Unsealed sources or open sources

These are sources that are in a form capable of giving rise to radioactive contamination on or in any other material with which they come into contact. They include any source that does not satisfy the rigid criteria required of sealed or closed sources. They include powders, dusts, solutions, liquids, gases, vapours, friable solids or solids from which the surface such as an oxide layer can be removed easily.

Radioactive contamination

This is normally thought of as unwanted radioactive substances on surfaces or in the air in the workplace. Contamination is also present if the wrong material is present in an operating zone, such as a glove box or fume cupboard. X-rays, gamma rays or a sealed source in contact with a surface will not give rise to contamination. Neutron sources and very high-energy X-rays will make materials exposed to them radioactive, which is a different matter altogether from becoming contaminated.

Measurement

The relevant authority for advising on the use of units in radioactivity and radiation measurement and in the field of radiological protection is the International Commission on Radiation Units and Measurements (ICRU). In 1975, the ICRU advised that the International System of units formulated in 1960 (SI units) should be used throughout the field of radiological protection, radioactivity and radiation measurements. Since 1975, it has been a statutory requirement in the UK for all radiation dose-rate meters to have their readings displayed in SI units. This requirement does not apply to contamination monitors or instruments such as installed alarm monitors, the purpose of which is to detect the presence of ionizing radiations.

Although the use of the traditional units has been largely superseded, they are still occasionally encountered. Old radioactive sources may be labelled in the traditional units and the specifications

of old X-ray sets may be written in the old units. Some knowledge of them may still be required for persons who may be engaged in the decommissioning or decontamination of old installations, disposing of old radioactive sources or dealing with old X-ray sets.

Units of radioactivity

Materials demonstrating radioactivity disintegrate at a certain rate per unit of time, usually per second (s^{-1}), which is the reciprocal of the SI unit of time. The special unit of radioactivity is the becquerel (Bq). The radioactivity of a material that is decaying at a rate of one disintegration per second ($1 s^{-1}$) is 1 becquerel (1 Bq). Because this is such a small quantity, it is more usual to encounter amounts such as the kilobecquerel (kBq), the megabecquerel (MBq) or even higher values.

The traditional unit, the curie (Ci), was originally defined as the number of disintegrations per second (d.p.s.) associated with 1 g of radium-226 in equilibrium with its daughters. It was later defined as being that amount of radioactivity giving rise to 3.7×10^{10} d.p.s. Fractions and multiples of the curie were used, e.g. the millicurie (mCi = 3.7×10^7 d.p.s.) and the megacurie (MCi = 3.7×10^{16} d.p.s.). The becquerel is approximately equivalent to 2.703×10^{-11} of a curie; thus 1 MBq = $27 \mu\text{Ci}$ and 1 mCi = 37 MBq.

Units of radiation dose

The derived SI unit for the measurement of radiation dose absorbed in any irradiated material has the dimensions of joules per kilogram and is called a gray (Gy). One gray dose is equivalent to the deposition of one joule per kilogram in the irradiated material:

$$1 \text{ Gy} = 1 \text{ J kg}^{-1} \quad (22.4)$$

The special derived SI unit for describing the dose equivalent is the sievert (Sv) and is the absorbed dose in Gy multiplied by a quality factor (Q) to take account of the different degrees of harm likely to result from unit dose to tissue due to different types of radiation. The quality factor is dimension-

less and therefore the dose equivalent still has the dimensions of joules per kilogram:

$$1 \text{ Sv} = 1 \text{ Gy} \times Q \quad (22.5)$$

Multiples or fractions of the primary units are commonly used such as a millisievert (mSv) (1 mSv = one-thousandth of a sievert) or a microsievert ($1 \mu\text{Sv}$ = one-millionth of a sievert). The traditional unit of the *röntgen* is now very rarely encountered except in the use of X-rays and in the specification of old X-ray tubes. Strictly, it is a measure of X-ray exposure rather than absorbed dose. It is not applicable to any other type of radiation, except perhaps gamma radiation up to about 3.0 MeV. It was defined in terms of ionization produced in air at STP and was given the symbol R:

$$1 \text{ R} = 2.58 \times 10^{-4} \text{ C kg}^{-1} = 87.7 \text{ ergs g}^{-1} \quad (22.6)$$

Exposure to 1 röntgen resulted in an absorbed dose of 95 ergs g^{-1} of irradiated material.

The unit used for the measurement of absorbed dose was the *rad*, which are the initials of 'röntgen absorbed dose'. It is a measure of the absorbed dose, or deposited energy, in any material from any type of radiation and has the numerical equivalent of 100 ergs g^{-1} :

$$1 \text{ rad} = 100 \text{ ergs g}^{-1} \quad (22.7)$$

To measure the effectiveness or harm of a radiation dose to human tissue, the unit used was the *rem*, the initials of 'rad equivalent man'. This is the absorbed dose in rads multiplied by a quality factor. The quality factor here has the same function as that described earlier.

$$1 \text{ rem} = 1 \text{ rad} \times Q \quad (22.8)$$

By applying the conversion factors for ergs to joules and grams to kilograms, it will be seen that $1 \text{ Gy} = 100 \text{ rad}$.

Units of radiation dose rate

The above units are measures of accumulated dose and are the result of being exposed in a radiation field for a finite time. To be able to calculate a potential dose it is necessary to measure dose rate. This is measured in dose per unit time, usually

per hour, for example microsievert per hour ($\mu\text{Sv h}^{-1}$). Dose rates in the main beam of X-ray sets, for example, may be given in dose per minute or second (Sv min^{-1} or mSv s^{-1}).

Instrumentation

No single type of instrument can be used to detect or measure all types of radiation. It is necessary to make a distinction between those that detect radiation and require some interpretation of the readings and those that will read directly in dose rate or accumulated dose. Passive radiation detectors are the film badge and the thermoluminescent dosimeters, which, when processed, will indicate accumulated dose.

Radiation detection or measuring instruments are all based on one of four types of detector which, with associated electronic circuitry, may be used to indicate count rate, dose rate or accumulated dose:

- 1 ionization chamber;
- 2 proportional counter;
- 3 Geiger–Müller tube;
- 4 scintillation counter.

The instruments may be either portable or installed in fixed positions. All the instruments depend on their response to ionization produced due to the interaction between ionizing radiation and matter. Alpha and beta radiation are detected by their direct ionizing properties so that the media in which they move becomes ionized and thus electrically conducting. X-rays and gamma rays produce electrons when they pass through matter and it is these electrons that are detected. Neutrons, when reacting with matter, produce a variety of charged particles, and again it is these that are detected. The detectors are connected to electronic circuitry to amplify their signal to enable it to be displayed as either an audible or visual response.

An *ionization chamber* consists of a volume of air enclosed in a vessel with two electrodes with a potential difference between them. The incident radiation ionizes the air and the ions formed are attracted to the electrodes. The current that is flowing is measured and can be converted to give a direct reading of the absorbed dose.

A *proportional counter* functions in a way similar to an ionization chamber but operates at a higher potential difference and is more sensitive. Because it operates at the higher voltage, the ions produced are accelerated and cause further ionization: a phenomenon known as ‘gas multiplication’.

A *Geiger–Müller tube*, more commonly known as a *Geiger counter*, operates at an even higher potential difference, and as a result ions produced are accelerated to higher velocities, and therefore it produces many more secondary ions than either the ionization chamber or the proportional counter. The current produced is no longer proportional to the dose because the enormous number of ions produced results in a relatively large pulse arriving at the electrode in comparison with that produced by the initial incident radiation.

Scintillation detectors rely on the ability of certain substances to re-emit energy they absorb from incident ionizing radiations as light. The emitted light is amplified by a photomultiplier before being fed into the electronic circuitry of the instrument and displayed as either a visual or audible signal. Scintillators may be either solid or liquid, the latter being used to detect very low-energy beta particles such as those emitted by tritium (hydrogen-3). Alpha radiation is difficult to detect because of its short range in air and in solid or liquid matter. Alpha detectors usually utilize a scintillation detector coupled with a photomultiplier. The window of the detector must be light tight to prevent spurious readings being obtained, but thin enough to allow the passage of alpha particles. The window is extremely fragile and therefore great care is required when using an alpha detector because even a minute hole in it destroying its light-tightness will render the instrument inoperative. The scintillator is also very thin, usually being a coating of a substance, such as zinc sulphide, on a screen transparent to light. Because of the thinness of the scintillating material, it is not significantly affected by other radiations.

Scintillation detectors are also used to detect beta radiation, X-radiation and gamma radiation, but because these are more penetrating, the windows can be made of thicker material and consequently they are less fragile than those used in

alpha particle detectors. They must still maintain light-tightness.

A sophisticated type of instrument incorporating a scintillation detector and suitable electronic circuitry is used to distinguish between the energies of incident radiations.

Instruments using any one of the types of detectors described are available in various forms for a variety of uses. Those used essentially for external radiation assessment may be described as radiation meters, dose rate meters, rate meters, etc. Pocket-sized instruments such as the quartz fibre electro-scope based on a miniature ionization chamber or electronic instruments utilizing small Geiger-Müller tubes are available. Geiger-Müller tubes are now available in a variety of forms, ranging from those only giving warning of the presence of radiation to sophisticated types capable of giving dose rates, integrated doses and having preset alarm indication when selected dose rates or integrated doses are detected.

Some instruments are designed specifically for monitoring surfaces for radioactive contamination. It should be noted that there is no completely satisfactory instrument for monitoring directly for tritium contamination. The information from contamination monitors is usually displayed in counts per second. It is necessary to convert these readings into quantities of radioactive contamination per unit area. As the response of the instrument is dependent on the energy of the emissions from the contaminant, it is necessary to have some form of conversion table or graph. Most instrument suppliers provide this information in the manuals that are supplied with the instrument.

It is a statutory requirement that portable instruments are checked by a qualified person to ensure that they remain within specified calibration limits at least once in every 12 months and whenever they may have been repaired or been subject to treatment that may have resulted in damage to the instrument.

Film badges and *thermoluminescent dosimeters* are used primarily for monitoring cumulative radiation doses to personnel, but they may also be installed to monitor accumulated doses in occupied and unoccupied areas. The recorded cumula-

tive personal doses are to be held as the legally required dose records for radiation workers. The film badge when processed provides a visual record of the person's dose and by reason of various filters in the film badge holder can differentiate between beta radiation, gamma radiation, X-radiation and doses due to neutrons of thermal energies. It should be remembered that either of these dosimeters only records the dose to the dosimeter and therefore only to the part of the body on which it is worn. If it is suspected that extremities of the body may be exposed to short-range beta radiation then special extra dosimeters should be worn.

Biological effects of radiation

The biological effects of ionizing radiations can be considered at varying biological levels: to the body as a whole, to individual cells of the body and even down to biological material such as the chromosomes and genes.

The effect on the body as a whole can be either acute, as in the case of very high doses being received in a short time or long term. Short-term effects due to very high doses delivered in a short space of time (from seconds to hours) may give rise to: changes in the blood count, the number of some cells such as the leucocytes being reduced (at 0.25 Sv); sickness (at 1.0 Sv); failure of blood-forming organs (at 4.0 Sv); damage to the intestinal tract linings; and damage to the central nervous system (at 6.0–12.0 Sv). Any whole-body dose in excess of 6 Sv is almost certain to prove fatal. Visible effects to the skin are hair loss (4.0–5.0 Sv) and from erythema of the skin to serious skin burns (from 6.0 to > 10.0 Sv). The long-term effects may result in the production of cancers, in either the blood (leukaemia), bone or soft tissue.

The effect on individual cells may be to cause the cell to die, it may slow down its speed of reproduction or it may cause damage to the genes or chromosomes in the cell. If the DNA (deoxyribonucleic acid) in the sex cells in either the male sperm or the female ova is damaged then damage may be passed on to offspring.

Damage sustained by an irradiated individual is termed the 'somatic' effect and that passed on to

descendants is called the ‘hereditary’ effect. ‘Deterministic’ effects, or ‘non-stochastic’ effects as they were originally called, are those that will not occur until a certain threshold of dose is exceeded. ‘Stochastic’ effects are those for which there is a probability of them occurring, with no dose threshold apparent and an increasing probability of occurrence with increase in dose.

Control

Legislative control

Within a decade or so of the discovery of radioactivity and radiation, it became apparent that they presented a potential hazard to persons working with them and that some sort of control on exposure was required. National bodies eventually united to form an International X-ray and Radium Protection Committee in 1928, which eventually developed into the International Commission on Radiological Protection (ICRP). The ICRP has a number of technical committees advising on various aspects of radiological protection, their findings being published in the *Annals of the ICRP*. Their latest general recommendations, based on revised risk assessments calculated from the long-term effects on Hiroshima and Nagasaki atomic bomb survivors and other data accumulated since their previous publications in 1975, appear in ICRP Publication 60 (ICRP, 1991a). The recommendations of this publication are intended to prevent the occurrence of deterministic effects and limit the probability of stochastic effects, both to individuals and to their immediate and second-generation offspring, to an acceptable level. Table 22.3 summarizes the dose limits

recommended by ICRP 60, which should achieve the aims stated above.

The ICRP recommends no special occupational dose limits for women who are not pregnant, the limits being the same as for men. Once pregnancy has been declared, the ICRP recommend that the dose to the surface of a woman’s abdomen should be restricted to 2 mSv and 1 mSv to the foetus for the remainder of the pregnancy.

Established in earlier publications, the same overriding principle remains of keeping doses ‘as low as reasonably achievable’ (the ALARA principle), with economic and social factors being taken into account. As a method for achieving their recommendations, the ICRP commended the following practices.

- No practice should be adopted unless its introduction produces a positive net benefit.
- All exposures should be kept as low as reasonably achievable by application of constraints.
- None of their recommended dose limits should be exceeded.

National governments such as the UK Parliament, and organizations such as the European Commission, take note of these recommendations and incorporate them in their legislation.

Legislation to control exposures not only requires control of doses to those persons exposed in the course of their employment but also to members of the general public. It should be noted that both the ICRP recommendations and the statutory limits embodied in regulations exclude radiation doses incurred as a result of medical or dental practices and as a result of exposure to natural background radiation. Exceptions to the latter are that some form of control may be necessary for persons who may be exposed to:

Table 22.3 Summary of ICRP 60 dose limits* (from International Commission on Radiological Protection, 1991).

Radiation workers	50 mSv in any 1-year period but not more than 100 mSv in any 5-year period (implied average dose of 20 mSv per year)
General public	1 mSv per year, but in special circumstances a higher effective dose could be allowed in a single year, but the average dose must not exceed 1 mSv per year during the person’s lifetime

*There is no change to the dose limits for individual organs or for the eye lens.

- materials containing elevated levels of natural radionuclides;
- cosmic rays to aircrew and other frequent fliers in aircraft;
- radon and its daughters.

Administrative measures for controlling exposure to ionizing radiation

Before any work with ionizing radiation or radioactivity is started, a hazard assessment to determine the potential risk to the workforce is required. When significant amounts of radioactivity are to be used, there may be a requirement also to assess the potential hazard to the environment and the general public living in the neighbourhood of routine operations and discharges of radioactive waste, as well as the effects of an accident. The latter may require the development of an emergency plan in cooperation with the local emergency services such as the police and fire services.

Following the assessment, which will indicate the degree of hazard, the degree of control can be determined. The establishment of areas must be considered. These are usually defined by boundary demarcation and notices, but may be defined by clear description in documents and require differing levels of control, for which adequate records are necessary. Deciding which members of the workforce are to be designated as classified radiation workers and arranging for their medical sur-

veillance and personal dosimetry must be done at an early stage. Radiation protection supervisors (RPS) and radiation protection advisors (RPA) must be appointed and their training arranged. Training of the personnel who will be working with radiation and/or radioactive material must also be arranged. Radiation and contamination monitoring programmes are to be designed and the necessity for air sampling considered. In some circumstances it may be necessary to provide personal air samplers. Monitoring programmes of waste streams must also be considered, whether solid, liquid or gaseous. Local rules must be prepared for each laboratory or plant, describing the precautions to be taken, and made available to every member of the workforce. A copy of the rules is to be displayed at the workstation. If the main document is long or complicated, a summary of the rules may be displayed at the workstation, but every operator must still have access to the main document.

Practical measures for controlling exposure to ionizing radiation

Practical control of exposure to ionizing radiations requires both foresight and planning, as with other potential hazards. Methods to protect persons from any potential harm are to be designed and developed to deal with the hazard. Control of radiation and radioactive material can be conveniently dealt with by considering the hazard from external radiation (radiation from sources outside

Table 22.4 Summary of UK permitted annual dose limits (mSv) (from Ionising Radiations Regulations, 1989).

	<i>Whole body</i>	<i>Individual organs</i>	<i>Lens of eye</i>
Radiation workers	*20	500	150
Trainees under the age of 18	6	150	50
Employees who are not radiation workers	6	50	15
Other persons including members of the general public	**1	50	15

The dose limit to the abdomen of a woman of reproductive capacity is 13 mSv in any consecutive 13-week period.

*100 mSv in any consecutive 5 years implying a maximum of 20 mSv in any calendar year but subject to a maximum of 50 mSv in any single year.

**In certain circumstances, 5 mSv in a single year may be allowed.

The dose limit to the foetus once pregnancy has been declared is 1 mSv during the period of the pregnancy.

the body) and internal radiation (radiation from sources inside the body) separately although in practice concurrent precautions may be required.

Control of external exposure

External radiation can arise from sealed or unsealed radioactive sources, X-ray equipment, accelerators or any electrical device capable of accelerating electrons to greater than 5.0 MeV. There are three classical methods that can be applied to reduce exposure to external ionizing radiation: time, distance and shielding.

Time

As cumulative dose is a function of exposure time and dose rate (just as cumulative distance travelled is a function of time and speed), reducing the time spent in a field of radiation will reduce the total dose received. In other words, do not spend any more time in a radiation field than is required to perform the task in hand. Do not conduct discussions in a radiation field unless absolutely necessary; adjourn to an area of low background.

Distance

The greater the distance between oneself and a source of radiation, the lower the dose rate that will be encountered and consequently the total dose received will be reduced. It is not always advantageous to put the maximum distance possible between oneself and an exposed source. For example, manipulation of a source using 2-m-long tongs is much more difficult than when using 1-m-long tongs and therefore the amount of time necessary to complete the task may well offset the advantage gained from the increased distance. In physics, there is a law called the 'inverse square law', which states that if the distance between a point source of radiation and the object being irradiated is doubled, the dose rate at the target is reduced to one-quarter of the original dose rate.

For example, if the dose rate at point 'A', 1 m from a source of radiation, is 10 mSv h^{-1} , then the dose rate at point 'B', 2 m from the source, will be

reduced to 2.5 mSv h^{-1} , and at point 'C', 3 m from the source, it will be reduced to 1.1 mSv h^{-1} :

$$D_B = 10/2^2 = 10/2 \times 2 = 2.5 \text{ mSv h}^{-1}$$

and

$$D_C = 10/3^2 = 10/3 \times 3 = 1.1 \text{ mSv h}^{-1}$$

thus

$$D_B = D_A/d^2 \quad (22.9)$$

where D_B is the dose rate at point 'B', D_A is the dose rate at point 'A' (which must be a unit distance from the source, e.g. 1 m or 1 cm, etc.) and d is the distance from the source to the point of interest such as point 'B' or 'C'. The converse is true, so that on halving the distance between a source and the irradiated body, the dose rate is increased fourfold. It is for these reasons that it is always advised that radioactive material should not be manipulated by hand; the use of some form of handling device should always be used. As an example, if a beta source is manipulated with a gloved hand giving an estimated distance of 1.0 mm between the source and the hand, the dose rate at 100 mm will only be about one ten-thousandth of that encountered at 1.0 mm.

Shielding

The best method of ensuring that doses are kept to a minimum is to introduce shielding material between the radiation source and the person likely to be irradiated. The thickness of the shield required will depend on the energy of the radiation, the amount of radiation present and the required dose rate at the point of interest. The aim is to keep dose rates in the occupied zones to as low as reasonably practicable to ensure compliance with statutory and recommended dose limits. The more energetic the radiation and the higher the dose rate present then the thicker the shielding material required to provide the necessary degree of protection. The thickness of the shield required will also depend on the shield material. Thus for radiation of a specific energy and similar dose rate, only half the thickness of lead will be required to provide the

same degree of protection as that provided by steel of a certain thickness. Concrete (2.35 g cm^{-3}) is commonly used for shielding large sources. Shielding for neutron sources is usually complex, comprising a neutron moderator to reduce the energy of the incident neutrons, followed by a neutron capture material to stop the lower energy neutrons and, finally, a gamma shield to attenuate gamma radiation associated with the neutron source as well as gamma radiation produced in the neutron shields. When designing shielding for beta sources, consideration must be given to shielding against bremsstrahlung X-rays produced in the beta shield material or within the beta source itself as a result of self-absorption of beta particles in the source material.

Radiation monitoring

An adequate monitoring programme should be designed. This could include monitoring with hand-held portable instruments or installed monitors, which may be fitted with remote reading and alarm systems in addition to giving local indication of dose rate and/or integrated accumulated dose.

Control of internal radiation

Internal radiation arises from radioactive material deposited inside the body. There are four routes by which radioactive material can enter the body:

- 1 *inhalation* – inhaling contaminated air;
- 2 *ingestion* – taking radioactive material in through the mouth;
- 3 *injection* – radioactive material entering the body via wounds or medical conditions such as eczema causing skin lesions;
- 4 *absorption* – through the intact skin by radioactive material penetrating through the intact skin.

Absorption is a particular problem when working with tritium (hydrogen-3). Approximately 30% of any intake of tritium oxide (tritiated water) is by absorption through the skin, when exposure is due to airborne tritium. Where the skin is wetted by tritium-contaminated water, the intake by absorption is almost 100%.

Radioactive material will only enter the body if it escapes from the facility in which it is manipulated and contaminates surfaces, becomes airborne or contaminates food and drink. To prevent this occurring it is necessary to provide some form of containment. Examples of containment vary from fume cupboards, which rely on an adequate airflow into them to prevent dispersal into the workplace, up to sealed glove boxes working under negative pressure.

There should always be a system of ‘defence in depth’. The apparatus in which the work is performed should be in some form of primary containment, such as a fume cupboard or glove box. This primary containment should be in an area that has an appropriate classification, either a ‘controlled’ or ‘supervised’ area as defined in national legislation, and there should be a buffer area surrounding such areas incorporating personal monitoring and washing facilities.

It is necessary to ensure that an adequate monitoring programme is introduced. This may include air sampling as well as surface contamination monitoring. The issue of personal air samplers may be required. The results of the monitoring programme are to be compared with ‘derived limits’ for both air and surface contamination. Derived limits for inhalation and ingestion are found in ICRP Publications 30 and 61 (ICRP, 1981, 1991b).

A wide variety of monitoring techniques are available for assessing the amount of radioactive contamination on surfaces, ranging from direct monitoring to taking smear or wipe samples, collecting samples on adhesive tape and measuring the amount of material deposited on deposition trays. Many of the techniques available are described in the IAEA’s *Monitoring of Radioactive Contamination on Surfaces* (IAEA, 1970).

When unsealed radioactive sources are being manipulated, local rules should prohibit eating, drinking, applying cosmetics or the use of tobacco in any form in the designated area. The use of mouth-operated apparatus must also be banned. In areas in which there is a serious risk of contamination, personal belongings should be left in the changing room. Nothing should leave a contamin-

Table 22.5 Magnitude of doses of radiation due to natural background radiation and lifestyles.

Natural background (87%)	Cosmic radiation from outer space Gamma radiation from radioactive materials in the ground and buildings, radioactive gas in the air (radon gas) Radioactive materials in the body (potassium-40 in the skeleton)
Food and drink	Natural radioactive materials in food and drink (radium in Brazil nuts)
Medical exposure (11%)	Diagnostic X-ray examinations and therapeutic irradiations
Lifestyle (0.5%)	TV, air travel, luminous watches and instruments

ation controlled area unless it has been monitored and certified clear of significant contamination. Personal monitoring, washing and hygiene must be strictly enforced.

Transport of radioactive material

National and international regulations governing the transport of radioactive material exist, most of them being based on the IAEA's *Regulations for the Safe Transport of Radioactive Materials* (IAEA, 1985) and associated advisory material. In addition to these, individual carriers impose their own restrictions on the types and amounts of radioactive material that they will transport. Reference to these should be made before transporting radioactive material. Carriers and their regulations involved are: the Postal Services, Railway Authorities, Inland Waterways Authority, International Air Transport Association (IATA), Merchant Shipping (Dangerous Goods) Regulations (the Blue Book), etc. There is also the European Agreement concerning the International Carriage of Dangerous Goods by Road (ADR), which includes the carriage of radioactive materials.

Background sources of radiation

Radioactivity was discovered towards the end of the nineteenth century, but humans have been exposed to radiation from natural sources since their first appearance on the planet. It is interesting to consider the magnitude of doses due to natural background radiation and lifestyles (Table 22.5). The largest dose of radiation, about 87% of the total, received by a person living in the UK is due to

'natural background' radiation. Other radiation comes from food, lifestyle, radioactive material in the body, etc.

Further reading

- Annals of the International Commission on Radiological Protection* ICRP, No. 21, Nos 1–3; Publication 60, Recommendations of the ICRP on Radiological Protection, 1991
- Annals of the International Commission on Radiological Protection*, Vol. 6, 1981, ICRP Publication 30, Limits for Intakes of Radionuclides by Workers.
- Annals of the International Commission on Radiological Protection*, 21, No. 4, 1991, ICRP Publication 61, Limits for Intakes of Radionuclides by Workers. (Taking account of Dose Limits in ICRP Publication 60.)
- International Atomic Energy Agency (1980) *Monitoring of Radioactive Contamination on Surfaces*. Technical Report Series 120, Vienna 1970.
- Hughes, D. (1993) *The Control of Sources of Ionising Radiation*, HHSC Handbook No. 12.
- International Atomic Energy Agency (1985 *et seq.*) *Regulations for the Safe Transport of Radioactive Materials*, Safety Series No. 6
- International Atomic Energy Agency (1990) *Advisory Material for the IAEA Regulations for the Safe Transport of Radioactive Materials*, Safety Series No. 37, 3rd Edition, amended Vienna 1990
- European Agreement Concerning the International Carriage of Dangerous Goods by Road (ADR)* Department of Transport, HMSO, London.

National Statutory Regulations

Current relevant UK regulations are:
All of the following documents are available through The Stationary Office Ltd, P.O. Box 29, Norwich, NR3 1GN.
The Ionising Radiations Regulations 1999 Statutory Instrument 1999 No. 1333 and associated *Approved Code of Practice*.

The Radioactive Substances Act 1993, Chapter 12.

The Radioactive Material (Road Transport) Act 1991.

The Radioactive Material (Road Transport) Regulations
2002.

The Carriage of Dangerous Goods by Road (Driver Training)
Regulations 1996.

The Radiation (Emergency Preparedness and Public Information)
Regulations 2001.

The Reporting of Injuries and Dangerous Occurrences Regulations
1995.

Chapter 23

Biological agents

Julia M. Greig and Chris J. Ellis

Introduction
Agricultural and allied workers
Diarrhoeal diseases
 Brucellosis
 Q Fever
 Orf
 Cowpox
 Ringworm
 Erysipelothrix
 Lyme disease
 Hydatid disease
 Streptococcus suis infection
 Psittacosis (*Chlamydia psittaci*)
 Bovine tuberculosis
 Anthrax

 Leptospirosis
 Infections associated with occupational travel
 Malaria
 Travellers' diarrhoea
 Hepatitis A
Health-care workers
 Blood-borne viruses
 Hepatitis B virus
 Hepatitis C virus
 Human immunodeficiency virus (HIV)
 Varicella zoster (chicken pox)
 Tuberculosis
Microbiology laboratory workers
References

Introduction

When considering biological agents as causes of diseases associated with occupations, workers can be divided into four groups:

- 1 agricultural and allied workers;
- 2 occupational travel;
- 3 health-care workers;
- 4 microbiology laboratory workers.

Agricultural and allied workers

Farmers, vets, abattoir workers and butchers may be exposed to zoonoses, diseases transmitted from animals to man under natural conditions. The incidence of occupational disease in farmers is probably under-reported, but 20 000 zoonoses are reported annually in the UK (Health and Safety Commission, 2001). These are mainly food hygiene related.

Diarrhoeal diseases

There are a number of biological agents that can cause intestinal infection in man and animals.

Campylobacter is the commonest cause of sporadic gastroenteritis in the UK, with the number of reported infections in the general population rising annually (Department for Environment, Food and Rural Affairs, 2000). In recent years there has been a fall in the number of reported cases of *Salmonella* infection, although this organism remains the commonest cause of outbreaks of food-borne disease. *Escherichia coli* 0157 is another important zoonotic cause of gastroenteritis, with possible serious complications such as haemolytic uraemic syndrome. The protozoan parasite *Cryptosporidium* can cause water-borne outbreaks of prolonged diarrhoea.

Domestic animals frequently have high carriage rates of these organisms although remaining asymptomatic, making detection and elimination of infection difficult. Spread to farm workers can occur by eating contaminated foods, as for the population in general, or occupationally by direct contact with infected animals and contact with faecally contaminated environments. High rates of *Salmonella* and *Campylobacter* carriage occur in flocks of domestic fowl, whereas *E. coli* 0157 has its main reservoir in ruminants (Smith *et al.*,

2002). *Cryptosporidium* has a wide distribution in farm and wild animals. Farm workers must be provided with adequate washing facilities and must be encouraged to wash their hands frequently and thoroughly; they should not eat while working. Work is under way to control *Salmonella* infection in breeding flocks of domestic fowl with a statutory surveillance programme, vaccination and elimination of *Salmonella* from animal feeds (DEFRA, 2000).

Brucellosis

Brucellosis is a zoonosis of both public health and economic importance, especially in developing countries. In the UK, where *Brucella abortus* has been eliminated from cattle, the only cases reported in recent years were acquired abroad. However, in Northern Ireland, where *B. abortus* infection of cattle is ongoing, there is still a risk to public health and occupational cases (farmers and meat plant workers) have been reported (DEFRA, 2000).

Brucellae are small Gram-negative, aerobic coccobacilli. There are three species that are of public health significance. *Brucella melitensis* is the most pathogenic and invasive species. It has an animal reservoir in sheep and goats, and is found particularly in Mediterranean and Middle Eastern countries, where the annual incidence of human brucellosis is between 1 and 78 cases per 100 000 (Memish, 2001). *B. abortus* is found principally in cattle herds, whereas *Brucella suis* is found in domestic and feral swine.

Transmission from animals to humans is mainly via contaminated or untreated milk or milk products. Infection can also spread from direct contact with infected animals, carcasses and abortion products, either via the aerosol route or by direct inoculation through the conjunctiva or skin abrasions.

Brucellosis generally has an incubation period of between 1 and 3 weeks but it may be many months. The onset can be insidious or abrupt, with fevers, myalgia, fatigue and depression that may last for weeks to months. Physical examination is often unremarkable with lymphadenopathy, hepatomegaly or splenomegaly in only

10–30% of cases. Diagnosis is difficult both because the symptoms are non-specific and isolation of *Brucella* is difficult. Thus it is important to obtain an accurate history of occupational exposure, travel to enzootic areas or consumption of high-risk foods. A definitive diagnosis can only be made if the organism is isolated from blood or bone marrow. However, the rate of isolation from blood is only 15–70%, depending on the method used and the period of incubation. Cultures of bone marrow have a higher yield than blood. In many cases, a presumptive diagnosis is made on the basis of high or rising titres of specific antibody.

Treatment of brucellosis relieves symptoms, shortens the duration of illness and reduces the rate of complications. *Brucella* sp. is an intracellular organism and thus requires treatment with an antimicrobial agent with good intracellular penetration. To prevent disease relapse, treatment with a combination of two agents is required. Tetracyclines are among the most active drugs for treating brucellosis, and a combination of tetracycline (for 6 weeks) and streptomycin (for 3 weeks) has been shown to be the most effective treatment (Acocella *et al.*, 1989). Doxycycline and rifampicin provide a possible alternative combination (Montejo *et al.*, 1993). In children, there has been some success with cotrimoxazole and an aminoglycoside.

Prevention is by the elimination of disease in the animal population and the avoidance of raw milk and raw milk products. Eliminating disease in the animal population involves vaccination programmes and test and slaughter of infected animals. Heat treatment or pasteurization of milk also protects against brucellosis. As yet there is no human vaccine.

Q Fever

Q (query) fever is a rickettsial zoonosis caused by infection with *Coxiella burnetii*. Although in animals it rarely causes overt disease, infection in man occurs world-wide and is reported in the general population and in agricultural and allied workers.

Coxiella burnetii is a small coccobacillus with a Gram-negative cell wall and is an obligate intracellular parasite. It is resistant to heat, drying and

many common disinfectants and can thus survive for long periods in the natural environment. Cattle, sheep and goats are the main reservoirs of *C. burnetii* (Babudieri, 1959), although infection has been noted in a variety of other animals. The organism is probably maintained in ticks and other arthropods that infect domestic and other animals, and is shed in the milk, urine and faeces of infected animals. During parturition, large numbers of organisms are shed from the amniotic fluid and placenta. These can survive in the environment for several months and infect humans by inhalation of contaminated aerosols (Welsh *et al.*, 1958). Human infection can also result from the consumption of contaminated raw milk but human-to-human transmission has only very rarely been reported (Mann *et al.*, 1986).

Q fever has an incubation period of 14–39 days. Acute infection may be asymptomatic. The most common presentation is a self-limiting febrile illness of 2–14 days' duration. Acute infection may also present with atypical pneumonia or occasionally with rapidly progressive pneumonia, which may be complicated by hepatitis. Chronic Q fever, defined as infection that persists for more than 6 months, is uncommon but a much more serious disease, with approximately 60% mortality. It has a number of manifestations, including endocarditis, osteomyelitis, hepatitis and purpuric eruptions.

The diagnosis of Q fever is based on the results of serological testing, as most laboratories do not have the facilities to isolate *C. burnetii*. A complement fixation test is most frequently used and a fourfold rise in antibody titre between acute and convalescent samples is diagnostic of acute infection. Chronic Q fever is diagnosed in the presence of high titres to phase I antigen. Polymerase chain reaction (PCR)-based tests have also been used to identify DNA from *C. burnetii*.

Doxycycline is the treatment of choice in acute infection. Quinolones are also effective (Marrie, 2000). Chronic infection is more difficult to treat and requires two drugs. Doxycycline has been used in combination with quinolones, rifampicin or hydroxychloroquine. Treatment needs to be continued for at least 3–4 years and is lifelong in some cases (Levy *et al.*, 1991).

Prevention involves the appropriate disposal of birth products and aborted fetuses from sheep and goats, pasteurization of milk and education of those at risk of infection. A human vaccine has been developed and used successfully in Australia (Gilroy *et al.*, 2001), although those previously exposed to *C. burnetii* should avoid vaccination because of severe local reactions.

Orf

This is a parapoxvirus and is a common cause of skin lesions in sheep, goats and man. Laboratories report low numbers of cases, but these probably represent only a small proportion of the total, most cases being correctly diagnosed by farm workers. It is transmitted to humans by direct contact with animal lesions or infected hides or wool. Usually a single, granulomatous lesion develops with associated regional lymphadenopathy, 1 week after inoculation and is most commonly found on the hands or forearm. The lesion is painless unless it is complicated by secondary infection and typically has a red centre surrounded by a white ring and an erythematous halo. There is no specific treatment and the lesions resolve after a number of weeks. Immunity is lifelong, preventing further attacks.

Cowpox

This is a viral infection with historical significance. Edward Jenner inoculated cowpox material from a dairymaid 200 years ago and showed that the recipient became resistant to smallpox. Today, cowpox is a rare disease, found only in Europe and adjacent parts of the former Soviet Union. The reservoir hosts are rodents from which it spreads to cows, cats and humans. Although traditionally infection occurred following direct contact with the infected teats of dairy cows, in recent years infection has been seen more commonly in domestic cats, from which it can be transmitted to humans (Stolz *et al.*, 1996). The last recorded case of a cow with cowpox in the UK was 1978. Cowpox virus infection in humans usually produces localized painful pustular lesions at the site of inoculation into the skin, usually on the hands. Systemic symptoms are rare. In most cases the lesions

regress spontaneously but antivaccinia immunoglobulin may be given to more severe cases. Cowpox can be confirmed using electron microscopy of vesicle or scab contents (Baxby *et al.*, 1994). Prevention is by the identification and isolation of infected animals and proper attention to hand washing in people handling animals.

Ringworm

This is a common fungal infection of the skin characterized by round, crusty lesions. Most lesions arise from cattle (*Trichophyton verrucosum*) or from dogs (*Microsporum canis*). Horses, pigs and sheep may also harbour *T. verrucosum*. Transmission is by direct contact with skin lesions of infected animals, via a pre-existing skin abrasion or cut. In humans the lesions are most commonly annular, often starting as single or multiple red papules that spread, causing round or oval ring-like lesions. The centre of the ring is typically scaly, whereas the active borders are raised and reddish in colour. Treatment is with one of a number of topical antifungals of the azole group, such as clotrimazole. In extensive dermatophyte infection, an oral agent such as terbinafine may be necessary. Prevention is by the rapid identification and treatment of cattle infection.

Erysipelothrix

Erysipelothrix rhusiopathiae is a Gram-positive bacillus with a world-wide distribution. It is the aetiological agent of swine erysipelas and domestic swine are thought to provide the main reservoir of infection. However, it also causes disease in turkeys, chickens and ducks. The organism can persist for long periods in the environment and survive in marine locations. Transmission to humans from animals is by direct cutaneous contact and thus most infections are related to occupational exposure. Infection is particularly common in those who handle fish but is also seen in butchers, abattoir workers, veterinarians and occasionally in housewives.

Human infection can take one of three forms: a localized cutaneous infection known as erysipeloid, a diffuse cutaneous eruption with systemic

symptoms or, rarely, as bacteraemia, which is sometimes associated with endocarditis. The localized skin infection has a distinctive appearance with a slowly progressive, painful, purple swollen area at the site of inoculation, which spreads centrifugally. Diagnosis of erysipeloid can be difficult if it is not recognized clinically as the organism is slow growing and resides deep in the skin. Recently, two PCR assays have been described for the diagnosis of swine fever, one of which has been successfully applied to human samples (Brooke and Riley, 1999).

Treatment is with penicillin, which should be given parenterally to septicaemic patients and those who are relatively immunocompromised. Ciprofloxacin provides an alternative for penicillin-allergic patients.

Lyme disease

Lyme disease is caused by the spirochaete *Borrelia burgdorferi* in the USA and *Borrelia afzelii* and *Borrelia garinii* in Europe and Asia. Like many spirochaetal infections, Lyme is a disease of exacerbations and remissions with long periods of latency. In recent years the number of reported cases has risen, although only a small proportion of these were occupationally acquired, mainly in forestry and farm workers.

Lyme disease is the commonest vector-borne disease in the USA and is also well established in forested areas of Europe, including, in the UK, the New Forest, Thetford Forest and Scotland. Ticks of the *Ixodes ricinus* complex transmit Lyme disease. The spirochaete is maintained in nature in a variety of rodents and deer, from which it is passed to man following the blood meal of a tick.

The clinical features of Lyme disease can be divided into three stages. The disease often begins with flu-like symptoms with malaise, fatigue, headache and arthralgia, associated with a characteristic skin lesion: erythema migrans. In Europe this rash is commonly a localized erythematous ring with central clearing. In the USA, the rash tends to be more widespread, with a number of disseminated lesions. Stage two may follow days to weeks later, with involvement of joints, nervous system or heart. Arthritis is particularly common

in the USA and presents as intermittent attacks of arthritis, which, if untreated, can progress to a chronic arthritis. Acute neuroborreliosis may present as aseptic meningitis or unilateral or bilateral facial nerve palsy. Stage-three disease causes late or persistent infection of the nervous system, joints or skin.

Diagnosis of Lyme disease requires the characteristic clinical findings, a history of possible exposure and a positive antibody response on enzyme-linked immunosorbent assay (ELISA) and Western blotting. A definitive diagnosis by culture of *Borrelia* is difficult and in practice is only possible from erythema migrans lesions. PCR assays have been used successfully to detect *Borrelia* in joint fluid (Bunikis and Barbour, 2002).

Treatment is with doxycycline, 100 mg b.i.d., for 14–21 days. Amoxicillin can be used in children and pregnant women. If there is evidence of cardiac or central nervous system involvement then treatment should be with intravenous ceftriaxone.

Prevention involves the use of protective clothing in tick-infested areas and tick checks after leaving such areas (ticks probably require attachment for 24 h to transmit Lyme disease). Doxycycline has been used successfully to prevent Lyme disease following tick bites (Nadelman *et al.*, 2001). A vaccination against Lyme disease is available but it only has an efficacy of 76% after three injections. Antibody levels wane quickly and booster doses are required every 1–3 years. Thus vaccination should be reserved for those with continual exposure to tick bites in high-risk areas (Steere, 2001).

Hydatid disease

Hydatid disease is caused by the larval stage of cestodes (tapeworms) of the genus *Echinococcus*. *Echinococcus granulosus* causes cystic echinococcosis and is most frequently encountered; *Echinococcus multilocularis* causes alveolar echinococcosis and is not known to be present in the UK. *E. granulosus* has a world-wide distribution and occurs more frequently in rural grazing areas where dogs ingest organs from infected animals. In the UK, infection is relatively uncommon but most

commonly seen in mid-Wales, Herefordshire and the Western Isles (DEFRA, 2000).

The adult tapeworm lives in the small bowel of the definitive host (dogs) from where eggs are passed in the faeces. After ingestion by a suitable intermediate host (sheep, cattle, goats), an oncosphere is released from the egg and penetrates the gut wall. It migrates via the bloodstream to various organs, especially the liver and lungs, where a cyst develops and enlarges. The definitive host is infected when it ingests the infected organs of an intermediate host. In the UK, *E. granulosus* becomes a problem when dogs (definitive host) are in close contact with sheep (intermediate host). Man is an accidental host and becomes infected by ingesting eggs with the resulting release of oncospheres in the intestine and the development of cysts in various organs.

Symptomatic hydatid disease usually occurs many years after infection and results from an enlarging cyst. Hepatic involvement can present with abdominal pain, a mass or biliary tract obstruction. Cysts in the lung result in chest pain, cough and haemoptysis. Occasionally, cysts rupture, producing fever, urticaria, eosinophilia, anaphylactic shock and cyst dissemination. The diagnosis of hydatid disease is usually radiological with characteristic findings on ultrasonography, supported by positive serological tests.

Treatment is most commonly surgical with removal of the cyst and concurrent medical treatment with albendazole to prevent cyst dissemination and recurrence.

Prevention, with the regular screening and treatment of dogs in endemic areas, has been successful in New Zealand, Tasmania (Gemmell, 1990) and the UK. Early removal of carcasses and inspection of carcasses by abattoir workers is important. Vaccination of sheep has been used successfully in some areas (Lightowlers *et al.*, 1999).

Streptococcus suis infection

This infection is found in pigs, causing meningitis, polyarthritis, septicaemia and pneumonia, and is a rare but serious cause of human disease, with an average of two reported cases per annum in the UK (DEFRA, 2000). Most cases in humans occur in

those who handle pig meat. *S. suis* type 2 is most associated with disease in pigs and humans, although since 1996 serotype 14 has been emergent (Heath *et al.*, 1996). The disease in humans is acute pyogenic meningitis with deafness and ataxia. It can be successfully treated with penicillin G or vancomycin in those allergic to penicillin.

Psittacosis (*Chlamydia psittaci*)

Chlamydia are obligate intracellular parasites. Of the three *Chlamydia* species that infect man only *Chlamydia psittaci* is known to be zoonotic. It is common in birds (Crosse, 1990) and can infect cats and ruminants. Infection is therefore a hazard to those who handle pets, particularly birds, poultry farmers (turkey-associated psittacosis has a high attack rate), workers in abattoirs and processing plants, and veterinarians. In England and Wales in 2000, there were 177 laboratory-confirmed cases of *C. psittaci* infection (DEFRA, 2000).

Infection is spread by the respiratory route from infected birds, which may or may not show signs of disease. Infection in humans varies from asymptomatic to a non-specific, flu-like illness with fever and malaise or an atypical pneumonia, with non-productive cough, fever, headache and changes on chest radiography. Diagnosis is based on serological testing. Tetracyclines are the treatment of choice, although erythromycin is also effective.

Since the 1980s, exposure to *C. psittaci* has been reported as a risk to pregnant women assisting at lambing or who have contact with slaughtered sheep. Two cases were reported in 2000 (DEFRA, 2000). *C. psittaci* is well known as the commonest cause of abortion in sheep (where it is now referred to as *Chlamydophilia abortus*). Transmission to women in the third trimester of pregnancy can result in a severe septicaemic illness with fever, vomiting, headache and hypotension (Helm *et al.*, 1989). Treatment is with erythromycin, which should be started without waiting for diagnostic confirmation by serological testing. Early delivery is also beneficial as the placenta is heavily infected and may act as a reservoir of infection.

Prevention is by the exclusion of pregnant women from the lambing shed, annual vaccination of breeding sheep in which enzootic abortion has

been confirmed and the isolation of aborting ewes, with early removal and destruction of the aborted products and bedding.

Bovine tuberculosis

Mycobacterium bovis infection of cattle was common in the UK until the 1930s when about 40% of slaughtered animals had macroscopic signs of infection. Infection in man was acquired from drinking contaminated milk and usually resulted in tuberculous cervical lymphadenitis. With the introduction of pasteurization and a compulsory eradication programme in cattle (notification, compulsory testing, slaughter, valuation and compensation), there has been a significant reduction of cattle infection, such that most cattle herds in the UK are free from tuberculosis at present. However, the infection has not been completely eliminated because wild animals such as badgers and deer maintain this zoonosis. In 2000, 28 cases of *M. bovis* infection were confirmed in people in the UK (DEFRA, 2000). These were thought to be due to reactivation of latent infection or imported infection. The distribution of disease in humans does not reflect the prevalence of disease in cattle. *M. bovis* is usually sensitive to rifampicin, isoniazid and ethambutol, but not pyrazinamide. Bacille Calmette–Guérin (BCG) inoculation provides some protection against infection.

Anthrax

In the UK, anthrax is a rare and sporadic disease. The last notified case in England and Wales occurred in a woollen mill worker in 2000, but there have been no cases in Scotland and Northern Ireland since 1991 and 1993 respectively (DEFRA, 2000). Anthrax is caused by the Gram-positive bacteria *Bacillus anthracis*. Because of the long viability of its spores, *B. anthracis* still poses some risk to workers who process hides, animal hair, bone products and wool, and to veterinarians and agricultural workers who handle infected animals. There have been no cases of anthrax recorded in animals in the UK since 1997.

B. anthracis can be transmitted by cutaneous contact, by inhalation or via the gastro-intestinal

system. The spores gain entry through a skin abrasion and, after an incubation period of 2–7 days, cause itching followed by a papule, vesicles and a black eschar. There is usually surrounding oedema with associated lymphangitis and regional lymphadenopathy. Untreated, a sepsis syndrome may ensue. Pulmonary anthrax is a serious disease with an associated mortality approaching 90%. The initial symptoms are non-specific upper respiratory tract symptoms followed by rapidly worsening breathlessness, fever, respiratory failure and shock. Consumption of *B. anthracis* in contaminated meat may result in intestinal infection. This presents with nausea, vomiting and abdominal pain, followed by bloody diarrhoea and shock. There is a significant associated mortality of 25–60%.

Diagnosis of anthrax can be confirmed by the demonstration of the organism in blood, sputum or skin swabs. Ciprofloxacin is the treatment of choice. Penicillin and tetracyclines provide alternatives (Torres-Tortosa *et al.*, 2002).

Prevention is by the vaccination of those in high-risk occupations, such as those who handle imported hides and animal hair. The risk to agricultural workers and veterinarians is too low to warrant vaccination unless they are working in a country with a high incidence of anthrax.

Leptospirosis

Leptospirosis is a spirochaete with a world-wide distribution, caused by pathogenic serovars of the genus *Leptospira*. Different serovars are maintained in nature by different animal species. The organism survives in the kidneys and genital tract, and is excreted in urine and genital fluids. Man is an accidental host, infected either by direct contact with animal urine or by contact with an environment contaminated with animal urine. *Leptospira* invade through mucous membranes or via skin abrasions.

The serovars encountered most commonly in the UK are *Leptospira hardjo* and *Leptospira icterohaemorrhagiae*. *L. hardjo* is commonly found in cattle in the UK, in which it can cause a fall in milk production and infertility problems. *L. icterohaemorrhagiae*

rarely infects domestic animals, its principal animal reservoir being rats. Occupations most at risk are those in contact with water contaminated with rat urine. The use of protective clothing and the treatment of waste water has dramatically reduced leptospirosis in sewer workers and farmers and those who participate in water sports are now more commonly affected (Morgan *et al.*, 2002).

L. hardjo infection is frequently asymptomatic or may present as a flu-like illness with fever and headache. Rarely, lymphocytic meningitis may occur. Weil's disease, most commonly associated with *L. icterohaemorrhagiae* infection, is characterized by fevers associated with impaired renal and hepatic function, haemorrhage, shock and a high mortality. Treatment of severe infections is with parenteral penicillin G. Milder infections can be treated with oral amoxicillin or doxycycline.

Because of the large animal reservoir of leptospiral infection, disease control is difficult. Prevention has been successful in sewer workers with the use of protective clothing, but less so in dairy workers. *L. hardjo* infection in cattle has been reduced with the use of appropriate vaccination (Little *et al.*, 1992).

Infections associated with occupational travel

Those advising international travellers need to consider not only their immunization requirements and need for malaria prophylaxis but must also provide advice on healthy behaviour. Travellers should be informed that infection causes only 3% of all travel associated mortality, the greater risk being from road accidents and other trauma, that unprotected sexual intercourse in sub-Saharan Africa, Latin America, India, Thailand and China is extremely dangerous and that those who engage in sexual activity while abroad should use condoms. Travellers' diarrhoea is very common, affecting approximately 30% of travellers (Farthing, 1995). Simple advice regarding safe drinking water and food hygiene can prevent much morbidity.

Malaria

Malaria is an illness caused by one of four malaria parasites that infect humans: *Plasmodium falciparum*, *P. vivax*, *P. ovale* and *P. malariae*. *P. falciparum* is found principally in tropical regions and gives the greatest concern to non-immune travellers, with its potentially fatal outcome and resistance to many routinely used chemoprophylactic agents. In contrast, *P. vivax* usually arises in those returning from the Indian subcontinent and virtually never causes death. *P. ovale* and *P. malariae* are much less common and are rarely fatal. *P. vivax* and *P. ovale* both have dormant liver stages and can cause relapsing disease many years after leaving a malaria endemic area. In contrast, *P. falciparum* does not have a dormant form and causes disease from 2 weeks after entering a malaria endemic region to 3 months following return.

Malaria is transmitted from man to man by its vector, the female anopheline mosquito. The life cycle of *Plasmodium* spp. is dependent upon both asexual replication in man and sexual replication in the mosquito midgut. Thus the distribution of malaria follows the distribution of the anopheles mosquito. Travellers should avoid mosquito bites whenever possible by using clothing that covers arms and legs, insect repellants containing diethyl toluamide (DEET) and sleeping under bed nets impregnated with insect repellent (McClellan and Senthilselvan, 2002).

Falciparum malaria usually presents with non-specific flu-like symptoms including fevers, headache and myalgia. It is essential that a travel history is obtained and the possibility of malaria considered and investigated promptly. The diagnosis should be confirmed by examination of a peripheral blood film for malaria parasites. More recently, alternative techniques have been developed to aid the diagnosis of *P. falciparum*. These include an ELISA for a histidine-rich *P. falciparum* antigen (Richardson *et al.*, 2002) and an immunoassay for species-specific lactate dehydrogenase isoenzymes (Iqbal *et al.*, 2002). The first of these two tests is useful in the field and can be performed in 10 min using commercially available reagents. PCR-based techniques are also available (Richardson *et al.*, 2002).

Falciparum malaria should be treated with quinine, given intravenously in the presence of confusion, a seriously ill patient or if the patient is unable to retain oral medication. To reduce the risk of disease and recrudescence with partially resistant organisms, quinine treatment should be followed by a single dose of pyrimethamine with sulphadoxine (Fansidar) and oral doxycycline for 7 days. *P. vivax*, *P. ovale* and *P. malariae* can be treated with chloroquine. A course of primaquine is also required to eliminate persistent liver stages not removed by chloroquine. Those whose work takes them to remote areas away from medical assistance for prolonged periods should consider carrying quinine tablets for self-treatment of possible falciparum malaria.

As well as avoiding mosquito bites as described above, those travelling to high-risk areas should take a chemoprophylactic agent. There are a number of possible regimens and those advising travellers should be aware of national guidelines (Bradley and Bannister, 2003). The choice of agent depends on the travel destination and also on the traveller's medical history and personal preferences. Regimens presently available include chloroquine, chloroquine plus proguanil, mefloquine, doxycycline and a combination of atovaquone and proguanil (malarone) (Ling *et al.*, 2002). Chloroquine-containing regimens should not be used for those travelling to sub-Saharan Africa because of the high rates of chloroquine resistance there. Travellers must be aware that no regime can guarantee protection against malaria.

Travellers' diarrhoea

This is extremely common, affecting 20–50% of travellers (Farthing, 1995). The greatest risk is to those visiting countries in Africa, Latin America and Asia (Black, 1990). The majority of travellers' diarrhoea is infective in origin, the commonest pathogen being enterotoxigenic *E. coli* (Gorbach *et al.*, 1975), although a large number of other pathogens are frequently implicated. Most episodes are relatively mild and self-limiting and do not require medical attention. Travellers should know how to keep themselves well hydrated using a safe water source. Oral rehydration salts

are not normally necessary except for the elderly and young children. However, travellers may be confined to bed or have to alter their travel plans with resulting disruption to their business. Quinolone antibiotics, such as ciprofloxacin, have been used successfully to prevent travellers' diarrhoea (Radermaker *et al.*, 1989), but because of the risk of unwanted side-effects and concern over the emergence of drug-resistant bacteria, this is not generally recommended. However, prophylaxis may have a place for those on short-term trips when loss of a day may seriously affect the outcome of a trip. Another successful approach has been to use a single 500-mg dose of ciprofloxacin with the first episode of diarrhoea to shorten the duration of illness and incapacity (Salam *et al.*, 1994). Carrying antibiotics for self-treatment should probably be limited to those visiting remote areas and those with serious underlying medical conditions who will tolerate diarrhoea less well. All travellers should be advised regarding safe water and food hygiene.

Hepatitis A

Hepatitis A virus is acquired by the faecal–oral route and causes overt hepatitis in adults but often a subclinical infection in children. Following an incubation period of 2–6 weeks, there is a prodrome with flu-like symptoms, followed by jaundice. Symptoms normally resolve spontaneously after 2–4 weeks, although occasionally a prolonged cholestasis can complicate the picture. Treatment is supportive and the mortality associated with acute infection is extremely low. Hepatitis A has never been reported to cause chronic liver disease.

In developing countries, infection is usually acquired in childhood and most adults are immune. Business travellers at greatest risk are those who eat with local families and the risk is proportional to the length of time abroad. Generally, the risk to business travellers who stay in large hotels is small. Business travellers who travel frequently or who are likely to eat in unhygienic situations should be vaccinated against hepatitis A. Frequent travellers can be tested for antibodies to hepatitis A, which, if present, make vaccination unnecessary.

Health-care workers

Blood-borne viruses

Blood-borne viruses that are of the greatest risk to health-care workers (HCWs) are those that cause a chronic carrier state with persistent viral replication and viraemia in their human host. The most common that are encountered are HIV, hepatitis B virus (HBV) and hepatitis C virus (HCV). In general, risk of transmission of these viruses to HCWs is from exposure to blood, but also occasionally from other body fluids or body tissues. Transmission most commonly occurs after percutaneous exposure to a patient's blood after a 'needlestick' or 'sharps' injury. Use of a hollow needle during venepuncture has been associated with most cases of occupationally acquired blood-borne viruses (Mast *et al.*, 1993). HBV provides the greatest risk, with a transmission rate of 1 in 3 from a source patient who is 'e' antigen positive. When the source patient is infected with HCV, the transmission rate is 1 in 30 and for HIV the transmission rate is 1 in 300 (Expert Advisory Group on AIDS and Advisory Group on Hepatitis, 1998). Occasionally, infection has been documented from exposure of mucous membranes or non-intact skin to infected blood.

Occupational exposure to blood-borne viruses is unnecessarily common (Mast *et al.*, 1993). Most incidents result from failure to follow recommended guidelines, including the safe handling and disposal of sharps and wearing protective clothing. If an incident does occur, there should be a review to consider how recurrences may be prevented. All HCWs should know how to report and understand the importance of seeking advice following possible exposure to blood-borne viruses. All employers should have a policy for managing such exposures.

Immediately following any exposure, the wound should be washed well with soap and water and free bleeding of puncture wounds should be gently encouraged but wounds should not be sucked. Antiseptics and skin washes should not be used. Exposed mucous membranes, including conjunctivae, should be irrigated with plenty of water. The occupationally exposed HCW should seek urgent advice from a designated doctor, who should

perform a risk assessment and consider the risk of exposure to hepatitis B, hepatitis C and HIV. Specific post-exposure prophylaxis is discussed below.

Prevention of occupational blood-borne virus infection involves a number of general measures.

- 1 HCWs should wash their hands before and after any patient contact.
- 2 Gloves should be worn when contact with blood or other body fluid is anticipated.
- 3 Any broken skin should be adequately covered.
- 4 Sharps should be safely disposed of in a designated sharps box, close to their point of use.
- 5 Needles should not be resheathed unless there is a specific device available for this purpose.
- 6 Employers are required by law to carry out an assessment of current procedures to prevent or control exposure to substances hazardous to health.

Hepatitis B virus

HBV is a double-stranded DNA virus, infection with which can cause a chronic carrier state resulting in chronic active hepatitis, liver cirrhosis and ultimately hepatocellular carcinoma. In those who are chronic carriers, the virus can be isolated from nearly all body fluids, but blood, semen and vaginal fluids are mainly implicated in the spread of HBV infection. The three most common ways for the transmission of HBV infection to occur are:

- 1 through unprotected sexual intercourse;
- 2 through needle-sharing injecting drug users;
- 3 perinatally from an infected mother to her baby.

Approximately 10% of those infected as an adult and 90% of those who acquire the infection perinatally will develop a chronic carrier state.

Acute infection follows an incubation period of between 6 weeks and 6 months and may be asymptomatic or present with acute hepatitis that is occasionally complicated by fulminant liver failure. Hepatitis B surface antigen (HBsAg) appears in the blood between 2 weeks and 6 months after infection. Its persistence is associated with failure to clear HBV. If HBsAg remains positive at 6 months following infection then the patient is regarded as having developed a chronic carrier state. The appearance of antibody to HBsAg (anti-HBs) is gen-

erally associated with the disappearance of infectious virus from the circulation and immunity. Hepatitis Be antigen (HBeAg) is associated with ongoing viral replication and those who are HBeAg positive are highly infectious.

Effective prevention against HBV infection is possible with vaccination and this should be offered to all HCWs whose work involves handling needles or sharp instruments or brings them into contact with blood. Primary immunization involves three doses of vaccine at time 0, 1 and 6 months. The response to immunization should be checked 2–4 months after the last dose of vaccine. An antibody level of 100 iu l⁻¹ is considered to provide adequate protection and a booster dose of vaccine should be offered after 5 years (Salisbury and Begg, 1996). About 15% of those vaccinated will be non-responders with an antibody level of less than 10 iu l⁻¹. A small proportion of these will be infected with HBV and should be identified, so any risk to patients can be assessed. The rest of the non-responders should be offered repeat vaccination. If they do not develop a protective antibody response then it will be necessary to offer HBV immunoglobulin (within 48 h) in the event of exposure to infection. Active immunization should be given at the same time.

Hepatitis C virus

HCV is an RNA virus spread mainly by direct blood-to-blood contact. In the UK the greatest risk is to injecting drug misusers who share needles and syringes. It can also be spread by sexual contact and perinatal transmission, although these are less important modes of viral transmission. Seroprevalence to HCV is frequently asymptomatic although it may be marked by an acute hepatitis. Between 50% and 75% of those infected with HCV will develop chronic infection, with subsequent risk of chronic hepatitis, cirrhosis and hepatocellular carcinoma. Treatment with a combination of interferon and ribavirin is effective for some with chronic disease (Chander *et al.*, 2002).

The risk to HCWs exposed to blood from patients carrying HCV is less than for HBV but is still significant. At present there is no vaccine against

HCV, so general preventative measures as outlined above are essential.

Human immunodeficiency virus (HIV)

HIV is spread by inoculation of virus by sexual intercourse, by contaminated blood or blood products or vertically *in utero*, during birth or by breast milk. Infection with HIV occurs when the viral glycoprotein, gp120, binds to the cellular membrane CD4 molecule of T-helper lymphocytes (and other cells) and the virus enters the cell cytoplasm. HIV is an RNA virus that encodes for reverse transcriptase, enabling a DNA copy of viral RNA to be produced in an infected cell. This DNA provirus is incorporated into the host genome. Over time, if HIV infection remains untreated, there is progressive damage to the immune system and the infected individual may present with opportunistic infections.

Although HCWs have been reported who have acquired HIV infection at work, this is relatively unusual (PHLS, 1998) and the reported risk of infection following a needlestick injury from a known HIV-infected patient is 3 per 1000 injuries (Bell, 1997). The risk of occupationally acquired HIV is highest if the source patient has a high viral load, with a deep injury, if visible blood is on the device that caused the injury, and if the injury was from a needle that had been placed in a source patient's artery or vein.

If a HCW receives a significant injury from a known HIV-positive patient or someone considered likely to be HIV positive, then post-exposure prophylaxis (PEP) should be considered. There is evidence that zidovudine prophylaxis reduces the risk of occupationally acquired HIV infection by 80% (Cardo *et al.*, 1997). Although this is the only drug with proven benefit for PEP, a combination of three antiretroviral drugs is normally recommended because in the treatment of known HIV-infected individuals combination therapy is substantially more effective than monotherapy. The most commonly used regime is zidovudine, 250 mg b.i.d., lamivudine, 150 mg b.i.d. and nelfinavir, 1250 mg b.i.d. for 28 days. (Department of Health, 2004). Zidovudine and lamivudine are nucleoside analogue reverse transcriptase inhibi-

tors and nelfinavir is a protease inhibitor. However, expert advice should be sought, especially if the source patient is taking antiretroviral therapy. PEP prophylaxis should be started as soon as possible after the incident, ideally within 1 h, and PEP treatment packs should be readily available 24 h per day.

Blood should be taken from the HCW at the time of injury and at 3 and 6 months afterwards to check for seroconversion to blood-borne viruses. Whenever possible, blood should be obtained from the source patient to aid with risk assessment. This should not delay starting PEP when the risk is considered high. PEP can always be discontinued if HIV infection is later excluded.

Varicella zoster (chicken pox)

Varicella is a highly infectious viral disease that is transmitted by the respiratory route or by direct contact. Most infections occur in children under 10 years of age and result in a mild illness, unless the child is immunocompromised. Approximately 90% of the adult population are immune. However illness in non-immune adults can be more serious, particularly in pregnant women and smokers, and they are at risk of developing varicella pneumonia.

Health care workers who are not immune to varicella are both at risk themselves from contracting infection and also pose an infection risk to patients. Thus varicella vaccine is now recommended for all susceptible HCWs who have no contraindication to vaccination.

When assessing a HCW for vaccination, anyone with a definite history of VZV infection can be considered immune. If there is doubt about prior infection, serological testing should be used and those with VZV antibodies should be considered immune. Those who are non-immune should be offered vaccination unless they are pregnant, immunocompromised or have a history of hypersensitivity reactions to previous VZV immunization or to neomycin or gelatine. Women should be advised to avoid becoming pregnant for 3 months after vaccination, although follow-up of those inadvertently vaccinated during pregnancy showed no evidence of congenital varicella in liveborn infants.

HCWs should be warned to seek advice from occupational health specialists should they develop a rash in the month following vaccination. Local vaccine site rashes, that can be covered with a bandage or clothing, should not prevent HCWs from working. If they develop a more generalized rash, they should be excluded from work until the lesions have crusted.

Tuberculosis

Since 1987 the number of notified cases of tuberculosis (TB) in England and Wales has increased, particularly in urban populations and in ethnic minority groups from the Indian subcontinent and Africa (Rose, 1999). A significant number of cases of TB are associated with HIV (3% of cases in 1998). Drug resistance is an important issue, with about 6% of isolates showing isoniazid resistance and 1.3% of isolates being multidrug resistant (MDR), i.e. resistant to both rifampicin and isoniazid (Irish *et al.*, 1999).

Mycobacterium tuberculosis is spread from person to person via the aerosol route. Thus patients with non-pulmonary TB do not provide an infection risk to contacts. Patients whose sputum is smear positive (i.e. acid-alcohol-fast bacteria are visible on direct staining of sputum) provide the greatest infection risk. MDR-TB should be suspected in those who have had previous drug treatment for TB, have been in contact with a known case of MDR-TB or who fail to respond to conventional treatment.

HCWs are at a twofold increased risk from tuberculosis compared with the general population (Meredith *et al.*, 1996). Protection of HCWs involves both pre-employment measures, regular surveillance during employment and safe practices for patient care. Figure 23.1 summarizes the recommended pre-employment measures from The Joint Tuberculosis Committee of The British Thoracic Society. It depends upon recording a history or symptoms of tuberculosis, details of previous BCG vaccination and the presence of a BCG scar (Joint Tuberculosis Committee of the British Thoracic Society, 2000). Occasionally, chest radiography and tuberculin skin testing are indicated. HCWs need to be aware that BCG does not confer

complete protection against tuberculosis and they should know to report any symptoms suggestive of tuberculosis. During employment, completion of an annual questionnaire has been found useful as an additional screening measure. HIV-positive HCWs should not be exposed to infectious tuberculosis and they should not be given BCG vaccine.

Safe practices for patient care to prevent the spread of tuberculosis to HCWs and other hospitalized patients involve a number of measures. Patients with TB should be treated at home whenever possible. Patients admitted to hospital with suspected or confirmed pulmonary TB should be cared for in a single room, vented to the air outside and the door should be kept shut. Early identification of patients suffering from possible pulmonary TB is important to prevent nosocomial spread. Patients suffering from smear-positive pulmonary TB are usually considered non-infectious after 2 weeks' treatment, unless they are infected with MDR-TB. If MDR-TB is known or suspected then patients should be isolated in a negative-pressure ventilation room; not because MDR-TB is more virulent but because the consequences of infection are more serious. Staff and visitors should wear dust/mist masks and the number of staff caring for the patient should be limited. If the patient has to attend another hospital department then he/she should wear a dust/mist mask.

If aerosol-generating procedures such as bronchoscopy, nebulizers or induced-sputum examination are to be performed on a patient with possible pulmonary TB then these must not be performed on an open ward. Sputum induction should be avoided if at all possible. If such procedures are necessary, they must be performed in an appropriate room with adequate local exhaust ventilation.

Microbiology laboratory workers

The Control of Substances Hazardous to Health (COSHH) Regulations 1994 requires that an assessment be made of workers who are exposed to biological agents. Assessment should include

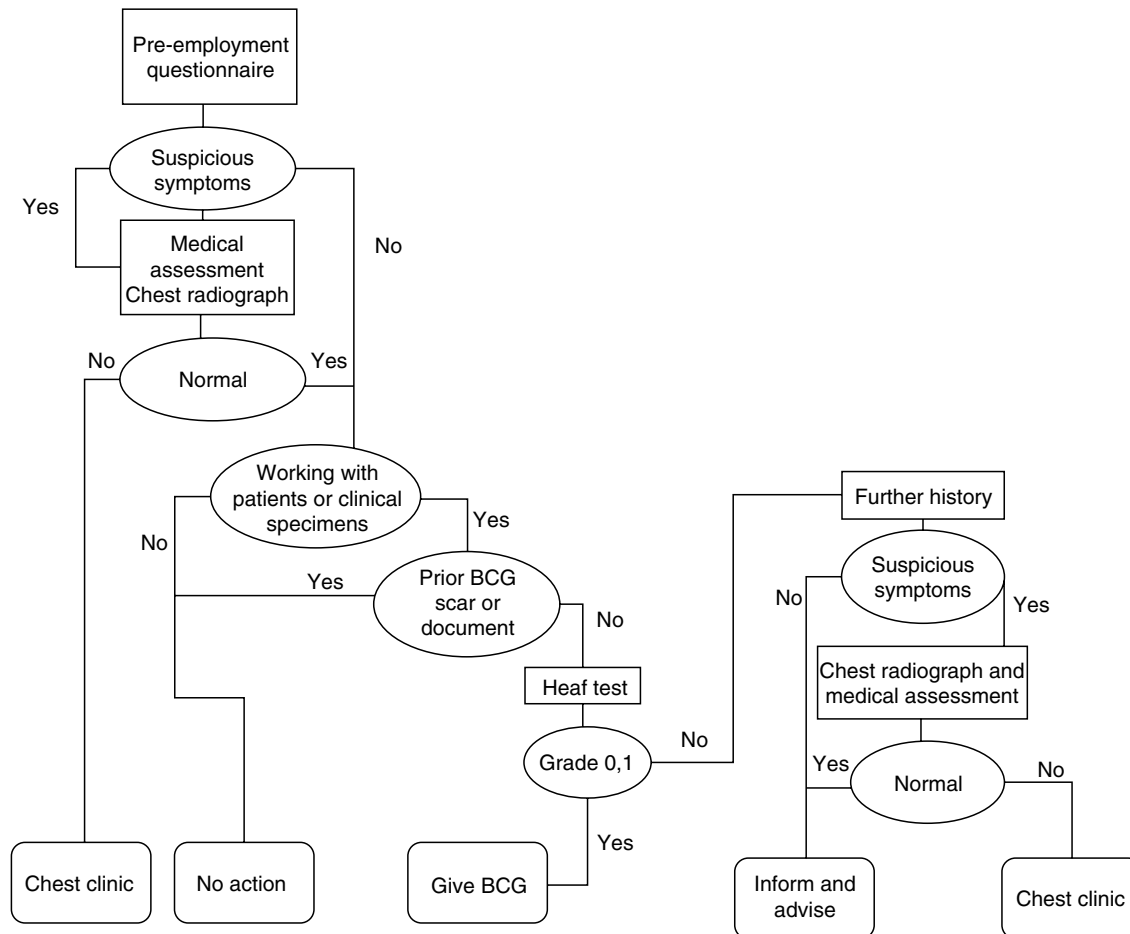


Figure 23.1 Screening of health-care workers for tuberculosis (from Joint Tuberculosis Committee of the British Thoracic Society, 2000).

measures to prevent or control exposure to infection. When there is infection risk, the use of preventative measures must be considered.

Medical laboratory workers have an increased risk of occupationally acquired infection, although this risk is declining. Four cases of pulmonary tuberculosis were reported among Public Health Laboratory Service staff in 1971 (97.2 per 100 000 person-years, six times the rate in the general population). In comparison, only one case of tuberculosis was reported in any British microbiology laboratory worker between 1988 and 1989. The incidence of hepatitis B in laboratory staff has shown a similar decline (Salisbury and Bagg, 1996).

When assessing the occupational risk of infection in laboratory workers, a number of factors need to be considered; what pathogens are likely to be involved, the local epidemiology of disease, the nature of material handled and the frequency of contact with infected material. The local laboratory facilities need to be looked at, as well as the availability of containment measures. When there is a vaccine to prevent infection, its safety and efficacy should be taken into account. However, vaccination should never be a substitute for good laboratory practice.

All laboratory workers should be vaccinated against hepatitis B and their vaccination history should be reviewed, checking for routine immun-

ization against tetanus, rubella and polio. BCG status and tuberculosis risk should be assessed as described above. Those who handle faeces should be offered typhoid vaccine. Cholera vaccine is not routinely indicated. The risk that faecal material will contain *Vibrio cholerae* is low and the infective dose is high, making the occupational risk negligible. There is little evidence of occupationally acquired hepatitis A and vaccination against this virus is usually only offered to those who handle viral cultures. Similarly, vaccination for influenza A, Japanese encephalitis, rabies, yellow fever and tick-borne encephalitis can be offered to those working with viral culture of these organisms.

The risk of laboratory-acquired meningococcal disease is low but not negligible. A recent survey identified five probable secondary cases in microbiology workers in England and Wales over a 15-year period (Boutet *et al.*, 2001). All cases had prepared suspensions of *Neisseria meningitidis* outside a safety cabinet, reinforcing the importance of safe laboratory practices. Meningococcal C conjugate vaccine (MenC) is now available and should be considered for those handling potentially infected specimens, particularly reference laboratory workers. However, there is still no vaccine against *N. meningitidis* group B.

References

- Acocella, G., Bertrand, A., Beytout, J., Durrande, J.B., Garcia Rodriguez, J.A., Kosmidis, J., Micoud, M., Rey, M., Rodriguez Zapata, M. and Roux, J. (1989). Comparison of three different regimens in the treatment of brucellosis: a multinational study. *Journal of Antimicrobial Chemotherapy*, **23**, 433–9.
- Babudieri, B. (1959). Q fever: a zoonosis. *Advances in Veterinary Science*, **5**, 81–181.
- Baxby, D., Bennett, M. and Getty B. (1994). Human cowpox 1969–93: a review based on 54 cases. *British Journal of Dermatology*, **131**, 598–607.
- Bell, D.M. (1997). Occupational risk of human immunodeficiency virus infection in healthcare workers: an overview. *American Journal of Medicine*, **102**, 9–15.
- Black, R.E. (1990). Epidemiology of travelers' diarrhoea and relative importance of various pathogens. *Review of Infectious Diseases*, **12** (Suppl.), 735–9S.
- Boutet, R., Stuart, J.M., Kaczmarek, E.B., Gray, S.J., Jones, D.M. and Andrews, N. (2001). Risk of laboratory-acquired meningococcal disease. *Journal of Hospital Infection*, **49**, 282–4.
- Bradley, D.J. and Bannister, B. (2003). Guidelines for malaria prevention in travellers from the United Kingdom for 2003. *Communicable Disease and Public Health*, **6**, 180–99.
- Brooke, C.J. and Riley, T.V. (1999) Erysipelothrix rhusiopathiae: bacteriology, epidemiology and clinical manifestations of an occupational pathogen. *Journal of Medical Microbiology*, **48**, 789–99.
- Bunikis, J. and Barbour, A.G. (2002). Laboratory testing for suspected Lyme disease. *Medical Clinics of North America*, **86**, 311–40.
- Cardo, D., Culver, D.H., Ciesielski, C.A. Heptonstall, J., Ippolito, G., Lot, F., McKibben, P.S. and Bell, D.M. (1997). A case control study of HIV seroconversion in health care workers after percutaneous exposure. *New England Journal of Medicine*, **337**, 1485–90.
- Chander, G., Sulkowski, M.S., Jenckes, M.W., Torbenson, M.S., Herlong, H.F., Bass, E.B. and Gebo, K.A. (2002). Treatment of chronic hepatitis C: a systematic review. *Hepatology*, **36** (Suppl.), S135–44.
- Crosse, B.A. (1990). Psittacosis: a clinical review. *Journal of Infection*, **21**, 251–9.
- Department for Environment, Food and Rural Affairs, Scottish Executive Environment and Rural Affairs Department, National Assembly for Wales Agriculture Development, Department for Agriculture and Rural Development, Northern Ireland, Department for Health, Food Standards Agency (2000). Zoonoses Report. DEFRA Publications, London.
- Department of Health (2004). *HIV Post-exposure Prophylaxis. Guidance from the UK Chief Medical Officer's Expert Advisory Group on AIDS*. Department of Health, London.
- Expert Advisory Group on AIDS and Advisory Group on Hepatitis (1998). *Guidance for Clinical Health Care Workers: Protection Against Infection with Blood-borne Viruses*. Recommendations of the Expert Advisory Group on AIDS and Advisory Group on Hepatitis. Department of Health, Wetherby.
- Farthing, M.J.G. (1995). Travellers' diarrhoea. In: *Travel-associated Disease*, (ed. G.C. Cook). Royal College of Physicians, London.
- Gemmell, M.A. (1990). Australasian contributions to an understanding of the epidemiology and control of hydatid disease caused by *Echinococcus granulosus*—past, present and future. *International Journal of Parasitology*, **20**, 431–56.
- Gilroy, N., Formica, N., Beers, M., Egan, A., Conaty, S. and Marmion, B. (2001). Abattoir-associated Q fever: a Q fever outbreak during a Q fever vaccination program. *Australia and New Zealand Journal of Public Health*, **25**, 362–7.
- Gorbach, S.L., Kean, B.H., Evans, D.G., Evans Jr, D.J. and Bessudo, D. (1975). Travellers' diarrhoea and toxigenic *Escherichia Coli*. *New England Journal of Medicine*, **292**, 933–6.

- Health and Safety Commission and Royal Agricultural Society of England (2001). Report of a conference on occupational health in agriculture organized by the Health and Safety Commission and the Royal Agricultural Society of England, Stoneleigh October 2001. Health and Safety Commission.
- Heath, P.J., Hunt, B.W., Duff, J.P. and Wilkinson, J.D. (1996). *Streptococcus suis* serotype 14 as a cause of pig disease in the UK. *Veterinary Record*, **139**, 450–1.
- Helm, C.W., Smart, G.E., Cumming, A.D., Lambie, A.T., Gray, J.A., MacAulay, A. and Smith, I.W. (1989). Sheep-acquired severe *Chlamydia psittaci* infection in pregnancy. *International Journal of Gynaecology and Obstetrics*, **28**, 369–72.
- Iqbal, J., Khalid, N. and Hira, P.R. (2002). Comparison of two commercial assays with expert microscopy for confirmation of symptomatically diagnosed malaria. *Journal of Clinical Microbiology*, **40**, 4675–8.
- Irish, C., Herbert, J., Bennett, D., Gilham, C., Drobniowski, F., Williams, R., Smith, E.G., Magee, J.G., Watt, B., Chadwick, M. and Watson, J.M. (1999). Database study of antibiotic resistant tuberculosis in the United Kingdom, 1994–6. *British Medical Journal*, **318**, 497–8.
- Joint Tuberculosis Committee of the British Thoracic Society (2000). Control and Prevention of tuberculosis in the United Kingdom: Code of Practice 2000. *Thorax*, **55**, 887–901.
- Levy, P.Y., Drancourt, M., Etienne, J., Auvergnat, J.C., Beytout, J., Sainty, J.M., Goldstein, F. and Raoult, D. (1991). Comparison of different antibiotic regimens for therapy of 32 cases of Q fever endocarditis. *Antimicrobial Agents and Chemotherapy*, **35**, 533–7.
- Lightowers, M.W., Jensen, O., Fernandez, E., Iriarte, J.A., Woollard, D.J., Gauci, C.G., Jenkins, D.J. and Heath, D.D. (1999). Vaccination trials in Australia and Argentina confirm the effectiveness of the EG95 hydatid vaccine in sheep. *International Journal of Parasitology*, **29**, 531–4.
- Ling, J., Baird, J.K., Fryauff, D.J., Sismadi, P., Bangs, M.J., Lacy, M., Barcus, M.J., Gramzinski, R., Maguire, J.D., Kumusumangsih, M., Miller, G.B., Jones, T.R., Chulay, J.D. and Hoffman, S.L., Naval Medical Research Unit 2 Clinical Trial Team (2002). Randomized, placebo-controlled trial of atovaquone/proguanil for the prevention of *Plasmodium falciparum* or *Plasmodium vivax* malaria among migrants to Papua, Indonesia. *Clinical Infectious Diseases*, **35**, 825–33.
- Little, T.W., Hathaway, S.C., Boughton, E.S. and Seawright, D. (1992). Development of a control strategy for *Leptospira hardjo* infection in a closed beef herd. *Veterinary Record*, **131**, 383–6.
- Local Collaborators, PHLS AIDS and STD Centre, Scottish Centre for Infection and Environmental Health. (1998). Occupational acquisition of HIV infection among health care workers in the United Kingdom: data to June 1997. *Communicable Disease and Public Health*, **1**, 103–7.
- Mann, J.S., Douglas, J.G., Inglis, J.M. and Leitch, A.G. (1986). Q fever: person to person transmission within a family. *Thorax*, **41**, 974–5.
- Marrie, T.J. (2000). *Coxiella burnetti* (Q fever). In: *Mandell, Douglas and Bennett's Principles and Practice of Infectious Diseases*, 5th edn (eds G.L. Mandell, J.E. Bennett, R. Dolin), pp. 2043–9. Churchill Livingstone, New York.
- Mast, S.T., Woolwine, J.D. and Gerberding, J.L. (1993). Efficacy of gloves in reducing blood volumes transferred during simulated needlestick injury. *Journal of Infectious Diseases*, **168**, 1589–92.
- McClellan, K.L. and Senthilselvan, A. (2002). Mosquito bed nets: implementation in rural villages in Zambia and the effect on subclinical parasitaemia and haemoglobin. *Tropical Doctor*, **32**, 139–42.
- Memish, Z. (2001). Brucellosis control in Saudi Arabia: prospects and challenges. *Journal of Chemotherapy*, **13** (Suppl.), 11–17.
- Meredith, S., Watson, J.M., Citron, K.M., Cockcroft, A. and Darbyshire, J.H. (1996). Are healthcare workers in England and Wales at increased risk of tuberculosis? *British Medical Journal*, **313**, 522–5.
- Montejo, J.M., Alberola, I., Glez-Zarate, P., Alvarez, A., Alonso, J., Canovas, A. and Aguirre, C. (1993). Open, randomized therapeutic trial of six antimicrobial regimens in the treatment of human brucellosis. *Clinical Infectious Diseases*, **16**, 671–6.
- Morgan, J., Bornstein, S.L., Karpati, A.M., Bruce, M., Bolin, C.A., Austin, C.C., Woods, C.W., Lingappa, J., Langkop, C., Davis, B., Graham, D.R., Proctor, M., Ashford, D.A., Bajani, M., Bragg, S.L., Shutt, K., Perkins, B.A. and Tapero, J.W., Leptospirosis Working Group (2002). Outbreak of leptospirosis among triathlon participants and community residents in Springfield, Illinois, 1998. *Clinical Infectious Diseases*, **34**, 1593–9A.
- Nadelman, R.B., Nowakowski, J., Fish, D., Falco, R.C., Freeman, K., McKenna, D., Welch, P., Marcus, R., Agüero-Rosenfeld, M.E., Dennis, D.T. and Wormser, G.P., Tick Bite Study Group. (2001). Prophylaxis with single-dose doxycycline for the prevention of Lyme disease after an *Ixodes scapularis* tick bite. *New England Journal of Medicine*, **345**, 79–84.
- Radermaker, C.M., Hoepelman, I.M., Wolfhagen, M.J., Beumer, H., Rozenberg-Arska, M. and Verhoef, J. (1989). Results of double-blind placebo-controlled study using ciprofloxacin for prevention of traveller's diarrhoea. *European Journal of Clinical Microbiology and Infectious Diseases*, **8**, 690–4.
- Richardson, D.C., Ciach, M., Zhong, K.J., Crandall, I. and Kain, K.C. (2002). Evaluation of the makromed dipstick assay versus PCR for diagnosis of *Plasmodium falciparum* malaria in returned travelers. *Journal of Clinical Microbiology*, **40**, 4528–30.
- Rose, A.M.C. (1999). 1998 TB Survey in England and Wales: final results. *Thorax*, **54** (Suppl.), A5.

- Salam, I., Katelaris, P., Leigh-Smith, S., Farthing, M.J. (1994). Randomised trial of single-dose ciprofloxacin for travellers' diarrhoea. *Lancet*, **344**, 1537–9.
- Salisbury, M.D. and Begg, N.T. (eds) (1996). *Immunisation against Infectious Diseases*. The Stationary Office, London.
- Smith, D.G., Naylor, S.W., Gally, D.L. (2002). Consequences of EHEC colonisation in humans and cattle. *International Journal Medical Microbiology*, **292**, 169–83.
- Steere, A.C. (2001). Lyme Disease. *New England Journal of Medicine*, **345**, 115–25.
- Stolz, W., Gotz, A., Thomas, P., Ruzicka, T., Suss, R., Landthaler, M., Mahnel, H. and Czerny, C.P. (1996). Characteristic but unfamiliar: the cowpox infection, transmitted by a domestic cat. *Dermatology*, **193**, 140–3.
- Torres-Tortosa, M., Caballero-Granado, F.J., Moreno, I. and Canueto, J. (2002). Antimicrobial therapy for anthrax. *European Journal of Clinical Microbiology and Infectious Diseases*, **21**, 696.
- Welsh, H.H., Lennette, E.H., Abinanti, F.R. and Win, J.F. (1958). Air-borne transmission of Q fever: the role of parturition in the generation of infective aerosols. *Annals of the New York Academy of Science*, **70**, 528–40.

Chapter 24

Psychological issues

Anne Spurgeon

- Introduction
- What is stress?
- Coping strategies
 - Personality and attitudes
 - Locus of control
 - Hardiness
 - Negative affectivity
 - Type A behaviour pattern
 - Attitudes
 - Psychosocial hazards
- Work organization
 - Working hours
 - Role
 - Interpersonal relationships
 - Career development
 - Change
 - Violence and trauma
 - Bullying and harassment
- Home–work interface

- The physical environment
- Effects of stress
 - Mental health
 - Physical health
 - Gastrointestinal disorders
 - Cardiovascular disorders
 - Immune system disturbance
 - Cancer
 - Musculoskeletal disorders
 - Diabetes
 - Behaviour
 - Organizational effects
 - Management of stress
 - Recognition
 - Assessment
 - Intervention and control
- References
- Suggested further reading

Introduction

Interest in the psychosocial aspects of the working environment has increased rapidly in recent years. Problems associated with ‘occupational stress’ are now recognized as a central issue for those concerned with health and safety in the workplace. Perhaps because of the inherent subjectivity involved in its investigation, however, this is a field that excites considerable controversy and presents a variety of challenges. In particular, debate has centred on the extent to which features of the workplace can be identified as sources of pressure, independently of the individual responses of the workers themselves. It has frequently been pointed out that what may be construed as stress by one worker may simply represent an interesting challenge to another. Even more problematically, what may seem like stress to an individual in one set of circumstances may be viewed quite differently by

the same person at another time. This shifting nature of stress and its multiple determinants is what distinguishes it from many of the other more traditional workplace hazards and makes its recognition, assessment and control particularly challenging.

The continuing debate about the role of individual perception in determining the outcome of organizational pressure is reflected in the differing approaches to the management of psychological issues adopted by different organizations at different times. Broadly, these approaches have tended to be either individually based, focusing on stress management training and rehabilitation of distressed workers, or organizationally based, emphasizing the need for change in the structure and function of the workplace itself. In general, the organizationally based approach has gained more prominence in recent years, perhaps reflecting an understandable desire to avoid victim blaming

in situations where stress-related problems have developed. It has also been suggested that providing stress management training represents the measure of last resort, in some ways analogous to providing personal protective equipment within a hierarchy of hygiene control. Thus, it is argued that first every effort should be made to remove the actual sources of stress.

Others, however, maintain that an approach that entirely neglects the input of the individual response to a given situation is somewhat unrealistic. Probably the most effective policies have been designed to take account of both elements, recognizing the need to identify sources of pressure in the workplace but also providing assistance to individual workers where necessary. This chapter will therefore discuss both the individual and organizational determinants of occupational stress, its potential consequences for health and well-being, and approaches to its assessment and management.

What is stress?

Although there are many definitions of stress, most include the concept that it results from an imbalance between the demands placed upon an individual and their ability to meet those demands. A number of psychological models have been proposed to describe the processes involved in the

development of stress. The most influential of these has been that of Karasek (1979) (Fig. 24.1). The central elements of this model are two factors termed *demands* and *control*. ‘Demands’ encompasses all of those factors in the working environment that have the potential to exert psychological pressure on the individual. ‘Control’ (or ‘decision latitude’) refers to the amount of freedom available to the worker to make decisions in meeting those demands. The extent of control is regarded as a key concept in stress development as well as in the maintenance of psychological health in general.

In Karasek’s model, work experiences are grouped according to the level of demand (high or low) and the level of control (high or low). The most stressful jobs are those which combine high demands with low control. Jobs that involve high demands and high control are defined as ‘active’, containing positive stress that leads to personal growth and development. By contrast, jobs that involve only low demands, combined with either high or low control constitute ‘passive’ work experiences that are likely to result in a reduction in overall activity and well-being. Thus in Karasek’s model, stress (termed by Karasek job strain) increases as job demands increase relative to decreasing decision latitude. Further personal development is likely to increase when the challenges of the job are matched by the individual’s skill and freedom to decide how and when to meet those challenges.

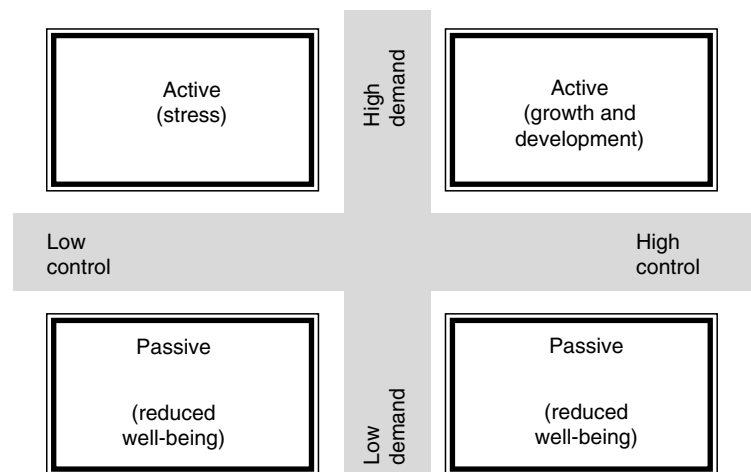


Figure 24.1 Model of stress development (adapted from Karasek, 1979).

Recently, a third factor, the presence or absence of social support, has been included as an effect modifier in the process (Sargent and Terry, 2000). This model remains one of the most widely accepted approaches to the assessment and management of stress at work and underpins much of the latest research in the field as well as providing a basis for the development of monitoring tools.

Coping strategies

One criticism of the Karasek model has been the emphasis on organizational factors and the assumption of uniformity of individual response. In this respect, the approach of Lazarus (1981) (termed the ‘transactional’ model) has been regarded as complementary to the demands–control model, in that it characterizes the experience of stress in terms of the individual’s appraisal of stress and subsequent adoption of preferred coping strategies in a given situation. Coping is defined as the individual’s ongoing efforts to meet demands that are perceived as exceeding their current resources. These efforts are broadly of two types, termed ‘problem-focused’ and ‘emotion-focused’, although each group contains a variety of subcategories (Pearlin and Schooler, 1978).

Problem-focused strategies consist of active attempts to solve the problem, for example by seeking information and testing different solutions or sometimes by curtailing contact with the source of stress. Emotion-focused strategies by contrast represent attempts to reduce emotional distress, either cognitively (for example by trying to see the positive side, reminding oneself that work is not everything) or behaviourally (for example by exercising, smoking or drinking alcohol). Within each category, the concept of control remains important; for both problem-focused and emotion-focused approaches, some strategies may be regarded as actively attempting to take charge of the problem, whereas others may be directed at escape or avoidance.

Although the concept of coping strategies is well accepted, there remains some dispute about whether these are flexible within individuals, in terms of adapting to the needs of the situation, or

whether they constitute inherent dispositions (in psychological terms ‘behavioural traits’) that lead individuals to favour particular approaches regardless of the circumstances. The distinction is important in terms of assessing the likely success of certain forms of stress management training.

Personality and attitudes

A number of other aspects of human behaviour are considered to be important in understanding the development of stress in response to work demands. Such factors may affect the way individuals initially perceive work demands (stressors) and/or the way in which they subsequently respond to them. In common with aspects of coping, however, there remains dispute about whether each constitutes a relatively fixed disposition or a potentially modifiable behaviour pattern. All of the characteristics described below have been included in research on stress and some have been incorporated into recently developed assessment tools.

Locus of control

This term describes people’s generalized expectancy regarding the responsibility for life events and has been applied to the working situation by Spector (1982). Individuals are regarded as having a tendency to attribute event outcomes either to their own actions (internal locus of control) or to fate, chance or the actions of powerful others (external locus of control). There is some evidence that those who tend towards an internal locus of control respond more positively to the experience of pressure and suffer less from the consequences of stress (Fusilier *et al.*, 1987).

Hardiness

Hardiness, a concept developed by Kobasa (1982), is a personality characteristic consisting of three elements, namely a belief in one’s ability to influence events (control), a sense of curiosity and meaningfulness in life (commitment) and an expectation that change is normal and stimulating (challenge). Individuals who score highly on

hardiness scales appear to experience less physical illness and psychological distress (Rush *et al.*, 1995).

Negative affectivity

Negative affectivity (Watson and Pennebaker, 1989) is characterized by a predisposition to experience negative emotions (distress, dissatisfaction, depression, anxiety) in all circumstances, regardless of the presence or absence of overt stressors. Although there is no evidence that negative affectivity is related to an increase in physical health problems, it does appear to be associated with increased reporting of non-specific health symptoms. In the work context this tends to manifest itself in high levels of job dissatisfaction and occupational stress.

Type A behaviour pattern

Prominent features of Type A behaviour are impatience, hostility, aggressiveness, competitiveness and a general sense of time urgency (Friedmann and Rosenman, 1974). In the workplace it is characterized by constantly working long hours and weekends to meet self-imposed unrealistic deadlines, constantly competing with others and expressing frustration and irritation with the work of subordinates. This behaviour pattern has been associated with an increased risk of coronary heart disease. Unsurprisingly, there is also evidence that a combination of Type A behaviour pattern and jobs containing low levels of control are associated with high levels of psychological disturbance (Rhodewalt *et al.*, 1991).

Attitudes

The term attitude refers to a consistency within individuals in their feelings towards a certain aspect of their environment (other people, situations, objects). As such, attitudes are strong motivators of behaviour. Although they are known to be strongly resistant to change, it is generally accepted that they are learned rather than fixed dispositions. Within the context of occupational stress they are important in terms of influencing

the response to the presence of perceived hazards. Thus the tendency to report physical symptoms may be linked to a range of factors unconnected with physical or chemical exposures.

These factors include individual personality factors as described earlier, situational factors such as isolation, which can lead to a preoccupation with health status, and attitudes or beliefs about illness and disease, its causes and likely prognosis (Spurgeon, 2002). Awareness of these factors helps in the understanding of a number of illnesses that have been ascribed to workplace conditions but which are often difficult to explain in terms of measured exposures. Perhaps the most obvious example is the constellation of symptoms known as *sick building syndrome*, which appears to result from a combination of physical and psychological exposures that occur to varying degrees in different circumstances (Crawford and Bolas, 1996). In addition, it has been demonstrated on a number of occasions that the presence of psychological stressors in the workplace is a significant predictor of musculoskeletal complaints. Here, psychological issues have often been shown to be of equal importance to ergonomic factors, and in some cases of greater importance (Skov *et al.*, 1996).

In summary, individually based factors (personality, behaviour patterns and attitudes) can influence the reporting of stress at work in a number of important ways. First, such factors help to determine the initial perception of stress in response to a given demand. Second, they act as modifiers in the relationship between perceived stress and subsequent consequences (strain and ill health) by determining the approach to coping that the individual adopts. Finally, in some circumstances when symptoms are difficult to explain, psychological factors can be strong determinants of the response to perceived physical and chemical hazards in the workplace.

Psychosocial hazards

Work demands as contained in Karasek's model constitute the stressors emanating from the work environment to which the worker is required to respond. In recent years these have often been referred to as psychosocial hazards, placing them

alongside physical and chemical hazards within a conventional risk assessment framework (Cox, 1993). A number of such hazards have been identified and these are described below.

Work organization

This includes such factors as workload, work pace and work schedule. In modern workplaces, these tend to be cited very frequently as sources of stress. An associated factor is what has been referred to as 'too little orderliness' (Schabracq *et al.*, 2001). Essentially, this refers to a situation in which there are numerous different competing demands such that many tasks remain half completed, awaiting further attention, or are completed unsatisfactorily with workers often attempting to carry out two or more tasks at the same time. The converse of this situation, which can be an equal source of stress, is the experience of routine, monotonous work that requires skills well below the capabilities of the worker.

Working hours

With the advent of the 24-h society, the requirement for shiftwork and unsocial hours is increasing. A substantial body of research has indicated the potential for adverse health effects associated with shiftwork, including persistent fatigue, anxiety and depression, as well as longer term effects such as increases in coronary heart disease (Spurgeon and Cooper, 2000). It is generally accepted that shiftwork, particularly when this involves night work or rapidly rotating shifts, is a physiological and psychological stressor. Although shiftwork is in many cases carried out by a 'survivor' population who have successfully adapted to this form of working time, certain groups remain vulnerable, notably those with pre-existing health complaints, those requiring regular medication and also older workers who appear to become less tolerant to shiftwork with increasing age. In addition to shiftworking, concerns have focused on the potential adverse effects of long working hours, which have also been linked to an increased risk of cardiovascular complaints and a deterior-

ation in mental health (Sikejima and Kagamimori, 1998).

Role

Included here are positions involving tasks that are too difficult as a result of inadequate training, producing errors, negative feedback and time spent correcting mistakes. Other examples include the experience of excessive responsibility, unclear requirements, role ambiguity and situations in which differing and conflicting expectations emanate from different parts of the organization (role conflict) (Beehr, 1976).

Interpersonal relationships

These may be characterized by poor communication, distrust and conflict, or social isolation. Alternatively, relationships may be too demanding in terms of social exchanges and create feelings of lack of privacy. Associated with this, emotional contagion of anxious or depressed feelings can occur spontaneously in some circumstances. The most dramatic examples of this have occurred in cases of 'mass hysteria'.

Career development

Most obviously a fear of redundancy represents a significant source of stress, which has been linked to an increased incidence of illness (Depolo and Sarchielli, 1987). Additional factors include 'glass ceilings' that do not allow particular groups of workers (e.g. women, ethnic minorities) to rise above a certain level in the organization, and policies that promote certain individuals beyond their level of competence ('Peter principle') or, alternatively, to positions where they will be isolated and ineffective.

Change

Restructuring and reorganization occur frequently in modern workplaces and have been identified as major sources of stress, particularly when they are carried out without adequate preparation or communication. The management of change has devel-

oped into a specialized area of study, recognizing its particular problems and requirements.

Violence and trauma

In a number of organizations the potential exists for intermittent and unpredictable acts of violence or verbal abuse against staff (Leather *et al.*, 1990). In addition to a variety of public service organizations (health service, social services, fire service and police) this also occurs in sectors such as retail, catering and transport. In the emergency services there is also frequent exposure to traumatic events that may not necessarily be associated with violence (Gersons and Carlier, 1992).

Bullying and harassment

Bullying and various forms of harassment (racial, sexual) have received increasing attention in recent years. Research on the prevalence and nature of bullying is limited at present and effective policies to address this problem are at an early stage of development (Zapf *et al.*, 1996). By contrast, many organizations now have systems in place to address issues of racial and sexual harassment (Fitzgerald and Shullman, 1993).

Home-work interface

Many workers experience considerable conflict in terms of reconciling the competing demands of home and working life (Barnett *et al.*, 1991). These difficulties occur particularly in those with care responsibilities for young children or for ageing/disabled relatives. In addition, stress that emanates primarily from home-related problems (e.g. bereavement or divorce) is likely to impact on working life and render workers more susceptible to organizational pressures.

The physical environment

Health problems associated with the physical environment may frequently be mediated by stress rather than constituting a direct effect of physical

exposures. Reference has already been made to musculoskeletal disorders and the difficulty of explaining *sick building syndrome* in conventional exposure terms. It has been observed that individuals may unconsciously attribute to physical and chemical exposure, symptoms that result from psychosocial hazards.

In addition, the potential for physical conditions such as noise and high temperature to increase fatigue and impair performance has been shown in a number of studies (Hockey, 1983). Where noise is concerned, the development of stress appears to be dependent on the nature of the noise and the nature of the task. This has been explained in terms of the *arousal curve* (Hebb, 1949) (Fig. 24.2).

Optimal performance is judged to occur when mental stimulation is around the top of the curve. Thus performance is likely to be impaired both at low levels of stimulation (dull, monotonous work), which is similar to Karasek's 'passive' condition, and at over-high levels of stimulation, where demands are excessive. In terms of noise levels, it can be seen that dull monotonous work may benefit from the addition of noise, as this increases stimulation towards the top of the curve. This goes some way to explaining the popularity of music as a background on routine production lines. Conversely, when work is already demanding, the introduction of further stimulation is likely to increase those demands beyond the top of the

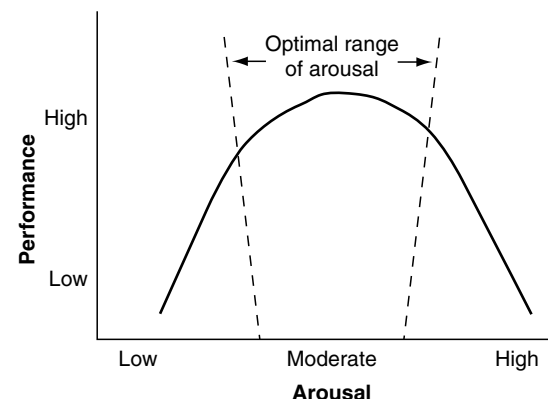


Figure 24.2 The arousal hypothesis: proposed relationship between arousal and performance.

curve and thus impair performance. This is particularly the case when noise contains meaning (e.g. conversation), rather than consisting of neutral 'white noise' such as may be generated by machinery. Noise with meaning creates a need for additional mental processing to screen out irrelevant information.

The psychological effects of temperature and lighting, aside from those extremes that could be regarded as beyond normal comfort zones, appear to be largely dependent on the extent of personal control. Thus stress tends to occur more often when the individual's opportunity to adjust heat and light to his or her own needs is limited.

Space is a complex psychological issue in the working environment and its effects should not be underestimated. In many organizations space and privacy are closely related to prestige and esteem, and physical assignment may be perceived as an important marker of both (Konar and Sundstrom, 1985). Expressions of dissatisfaction with physical working conditions may therefore represent a psychological issue rather than a physical one, a factor that may require consideration when investigating specific complaints.

Effects of stress

The effects of stress on individual workers are usually considered in terms of (1) effects on mental health, (2) effects on physical health and (3) effects on behaviour, although clearly these different aspects may frequently overlap or interact.

Mental health

A number of research studies have demonstrated a relationship between exposure to work-related stressors and mental health problems, typically in terms of an increase in symptoms of anxiety and depression. These problems may also be expressed as a range of non-specific physical symptoms such as headache, excessive fatigue, gastrointestinal problems and musculoskeletal pain. In addition, there may be cognitive difficulties such as disturbances of concentration and memory. A particular form of emotional exhaustion termed *burnout* has

been identified in those whose work involves providing support services to others, for example health-care and social services workers (Maslach and Jackson, 1981). In addition to emotional exhaustion, burnout is characterized by feelings of depersonalization, including negative and insensitive attitudes towards patients and clients and by feelings of inadequacy, discouragement and pessimism.

Exposure to traumatic events also appears to produce a particular pattern of psychological responses termed *post-traumatic stress disorder* (PTSD) (Gersons and Carlier, 1992). Typically, individuals suffering from PTSD show avoidance of any object, person or circumstance likely to remind them of the trauma, experience frequent nightmares and 'flashbacks' and exhibit heightened irritability and shock responses. Although workers in the emergency services clearly represent a group at particular risk of PTSD, it may also occur following violent incidents in many other sectors such as retail, catering, banking and transport.

Physical health

Stress has been implicated as a contributory factor in a number of physical conditions. The main areas of concern are summarized below.

Gastrointestinal disorders

Although it is now known that duodenal ulcers are associated with an infectious agent (*Helicobacter pylori*), epidemiological evidence shows that only a small percentage of infected individuals actually develop the condition, and that physiological changes associated with stress may act as a trigger (Melmed and Gelpin, 1996). Similarly, it has been suggested that conditions such as irritable bowel syndrome may be mediated by the action of hormones associated with the stress response (Burks, 1991).

Cardiovascular disorders

Epidemiological evidence indicates that prolonged psychological stress constitutes a risk factor for

coronary heart disease (Haan, 1988; Johnson *et al.*, 1989). Night work and long working hours have been identified as particular problems in this respect (Sparks and Cooper, 1997). Although the mechanisms underpinning these relationships are not fully understood, it appears that stress may increase the levels of cholesterol and lipoproteins in the blood, which constitute predisposing factors for cardiovascular disease.

Immune system disturbance

Hormonal changes in response to stress have been shown to result in changes to immune function, for example by reducing the functional efficiency of white blood cells (Biondi and Zannino, 1997). Impairment of the immune system increases susceptibility to infectious diseases and possibly to autoimmune disorders such as arthritis.

Cancer

The evidence linking the onset of cancer to the experience of stress is inconclusive, although several epidemiological studies have identified a small but statistically significant increased risk of cancer following stressful life events (Levenson and Bemis, 1991). Current opinion favours the view that the response to events in terms of the individual's particular coping style may moderate the relationship between stress and cancer.

Musculoskeletal disorders

A number of specific work-related stressors have been shown to be associated with the development of musculoskeletal disorders. These include monotonous work, high work load and time pressure (Buckle, 1997). In addition, low social support appears to predict increases in musculoskeletal symptoms (Leino and Hanninen, 1995). The most consistently demonstrated association is between low job control and the development of musculoskeletal problems (Houtman *et al.*, 1994). Furthermore, there would appear to be a strong interaction between physical and psychosocial factors, which is now generally accepted (Devereux *et al.*, 1999). However, the processes underlying

this relationship remain a matter of conjecture. It has been suggested that work pressure may increase muscle tension, which, coupled with poor posture and hurried movements, increases the likelihood of strain on muscles and joints.

Diabetes

Recent research has shown that acute stress tends to increase blood glucose levels, in both diabetic and non-diabetic subjects, although in non-diabetic individuals these levels usually return to normal on removal from the stress exposure (Hanson and Pichert, 1986; Wales, 1995). This has led to speculation that stress may act as a trigger for diabetes in those with a predisposition to develop the disease, although there is currently no evidence that stress *per se* is a causal factor.

Behaviour

The response to stress may include a number of health-threatening behaviours that represent forms of short-term coping. These include increases in cigarette smoking, alcohol consumption and drug abuse. High levels of job stress have been shown to be associated with heavier smoking, whereas workers with lower job stress appear to be more successful in giving up smoking (Green and Johnson, 1990). Where alcohol consumption is concerned, the relationship appears to be more complex in that job stress (notably high demands and low control) may lead to increases in alcohol abuse in certain vulnerable individuals, with personality factors playing an important role (Richman, 1992). To date, research on abuse of other substances has been rather limited, partly because the legal situation tends to preclude collection of valid self-report data.

Organizational effects

The organizational costs of stress are potentially of three types: (1) the costs of sickness absence; (2) the costs of recruitment and training when there is high staff turnover; and (3) the costs of poor performance in terms of productivity errors and accidents. All of these are difficult to measure.

For example, there are a number of reasons why figures relating to sickness absence may represent an underestimate, notably the reluctance of individuals or GPs to record stress or mental health problems as a reason for absence. However, figures derived from the 1990 Labour Force Survey indicated that 182 700 cases of work-related stress or depression (Davies and Teasdale, 1994) were reported at that time.

Management of stress

Current policies on work-related stress have developed within a conventional risk management framework (Cox, 1993). Although the differences between psychosocial and other types of workplace hazards create some difficulties in this respect, the approach would appear to be broadly applicable. Therefore, the management of stress essentially involves first a recognition of the problem, second, an assessment of the nature and scale of the problem and, third, the development of intervention strategies for control. Added to this is a requirement for on-going evaluation of the situation to monitor the effectiveness of any policies that have been implemented.

Recognition

Unlike various physical and chemical hazards, psychosocial problems can occur in any type of workplace. Therefore, they should ideally be part of any health and safety management policy. Identification of the presence of psychosocial hazards, however, may often result from the organizational consequences of stress, for example high levels of sickness absence, particularly frequent short-term absences, high levels of staff turnover and perhaps high levels of errors and accidents that are traced to time pressure and unrealistic deadlines.

Alternatively, there may be an unusually high rate of individual problems associated with aspects of mental health, which may be highlighted by occupational health services or reported anecdotally. More systematic assessment of this aspect may sometimes be carried out by a survey, using a validated mental health screening questionnaire

such as the General Health Questionnaire (Goldberg and Williams, 1988). This questionnaire provides an assessment of *caseness*, defined as vulnerability to developing mental health problems. Thus the percentage of the workforce scoring above a certain level (caseness) can be identified and the results benchmarked against those of similar organizations. It should be noted, however, that the introduction of mental health screening tools into an organization should be handled with extreme care, preceded by a clear explanation of the purpose, and accompanied by assurances about anonymity and about the use of group data only, with no reference to individuals.

Assessment

Assessment of psychosocial problems in the workplace involves two elements, namely (1) the level of pressure being experienced by the workers and (2) the particular source of that pressure. In recent years, a number of assessment tools have been developed for use in the workplace. The majority of these consist of self-report questionnaires and thus are based on the view that stress may only be identified in terms of the perceptions of the workers themselves. However, a smaller number have attempted to assess stressors objectively by providing a checklist of potential sources of pressure that can be identified in the course of a walk-through survey.

Some organizations have developed their own assessment tools in preference to those that are either reported in the published literature or commercially available. Such tools tend to be better tailored to the specific concerns of the organization. However, the disadvantage of this approach is that self-developed tools are unlikely to be valid or reliable, not having been subjected to extensive psychometric testing. Furthermore, they will produce data that are not readily comparable with those of organizations using standard instruments. As a compromise solution, many organizations have tended to combine the two approaches by using a published questionnaire together with a few additional organizationally specific items. Examples of commonly used instruments for the general assessment of stress are described below.

Pressure Management Indicator (PMI) (Williams and Cooper, 1998)

This is a revised and updated version of the original Occupational Stress Indicator (OSI) (Cooper *et al.*, 1988) that has been used very widely in industry and thus a large database of information is available for benchmarking purposes. The Indicator is a self-administered questionnaire consisting of a number of 'stand-alone' scales that assess sources and levels of pressure and job satisfaction in the organization, the physical and mental consequences of stress and aspects of personality and behaviour that may act as modifying factors.

Occupational Stress Measure (OSM) (Spurgeon and Barwell, 2000)

This recently published measure was originally developed for use in the National Health Service but has had increasing application in the private sector. It includes an initial rapid evaluation tool to determine the need for more in-depth assessment, and two self-report scales that measure perceived sources of pressure and individually experienced strains. It is intended as a diagnostic tool and is backed by an analysis and an advice/intervention service if required.

Occupational Stress Checklist (Kompier and Levi, 1994)

This is a checklist covering a range of potential sources of pressure in the organization, for example the way work is organized in terms of workload and deadlines, the length of time workers remain at their workstations without a break, and the availability of training and supervision. It is completed by the health and safety practitioner without input from the workers themselves. Although this approach has attracted criticism in some quarters, others have argued that, coupled with information on stress outcomes (mental health, sickness absence etc), it can represent a useful, rapidly administered diagnostic tool.

In addition to more general assessment tools, a number of diagnostic measures have been developed to assess the presence and/or consequences of specific sources of stress. Some of the more commonly employed are shown below:

- Standard Shiftwork Index (Barton *et al.*, 1995);
- Impact of Events Scale (PTSD) (Weiss and Marmar, 1996);
- Leymann Inventory of Psychological Terrorization (Bullying) (Leymann, 1996);
- Maslach Burnout Inventory (Maslach and Jackson, 1981).

Other questionnaires to assess specific problems, for example the incidence of violence, the sources of work/family conflict, the problems of older workers and many others are published in the wider literature on stress (see References).

Intervention and control

Detailed assessment of the potential sources of pressure in organizations has indicated that they are of three broad types: (1) non-intrinsic organizational factors; (2) intrinsic organizational factors; and (3) individually based factors. Essentially, non-intrinsic factors are sources of pressure that do not need to be there. They are not intrinsic to the work of the organization but arise as a result of factors such as poor management, lack of resources and lack of training. By contrast, intrinsic organizational factors are an integral part of a particular job, for example dealing with distressed people or with violent and traumatic incidents. In modern society the requirement for shiftwork would also fall into the category of factors that are an intrinsic part of some types of work. The third source of pressure is associated with the life circumstances of the individual worker or with certain characteristics of the worker that may make them more vulnerable to the effects of stress. Clearly, different types of intervention will be required in each case. These have usually been termed primary, secondary and tertiary interventions.

Primary intervention

Primary intervention is focused on the workplace and is concerned with instituting various forms of organizational change to address identified sources of pressure. Depending on which factors have been identified this might, for example, include particular forms of management training, provision of certain resources, reorganization of work sched-

ules or developing better means of communication between certain individuals or departments. When fear of violence or abuse has been identified, new forms of security arrangements may be considered. Alternatively, problems of home/work conflict may be addressed by what are termed 'family-friendly policies'.

Secondary intervention

This form of intervention, although still preventative in nature, focuses on the workers in terms of enhancing personal stress management techniques. In a general sense it might, for example, include stress awareness services, relaxation training, development of appropriate coping strategies or the provision of health promotional activities. More specifically, it may for some workers include training in self-protection and conflict management, or for shiftworkers include approaches to the management of sleep disruption and fatigue.

Tertiary intervention

In this area, the management of psychological hazards tends to diverge from that of traditional risk management, in that it acknowledges the likelihood of harm occurring in a proportion of the workforce, regardless of the intervention measures in place. The nature of psychosocial problems and their diverse individual and organizational causes means that they are likely to be experienced by a large percentage of any workforce at some time in their lives. Tertiary intervention essentially recognizes this and is concerned with the provision of services that include counselling and rehabilitation, either for stress-related problems in general or sometimes for specific behavioural consequences such as alcohol abuse.

Evaluation

A final important element of any policy is that of evaluation. This is usually carried out at intervals in terms of readministration of original assessment tools. It should be noted that the management of stress is a dynamic process that takes place against a background of constant organizational change. In this context procedures for ongoing rather than single instance monitoring and

evaluation constitute an essential element of any control strategy.

Legal considerations

Although there are no specific references to psychosocial hazards in UK health and safety law, the Management of Health and Safety at Work Regulations (1992) require employers to make an assessment of all risks to the health and safety of their employees and to institute measures to control those risks. Moreover, the Health and Safety at Work Act (1974) places a duty on employers to ensure the safety and health of their employees as far as is reasonably practicable. This includes mental as well as physical health.

Thus, there is a legal requirement for employers to ensure that workers do not become ill as a result of unacceptable levels of work-related stress. Several well-publicized civil cases have occurred in recent years, resulting in large financial settlements. In each case it was judged (1) that the employer should have been aware of the risk and (2) that the employee's ill health was caused or at least exacerbated by the employer's failure to amend working conditions that were known to be placing employees at risk. The number of such cases appears to be increasing (Earnshaw and Cooper, 1996) and it is clear therefore that, on legal as well as economic and humanitarian grounds, the assessment and management of psychosocial risks at work now occupies an increasingly important position in wider health and safety policy. This trend seems unlikely to diminish in the future.

References

- Barnett, R.C., Davidson, H. and Marshall, N.L. (1991). Physical symptoms and the interplay of work and family roles. *Health Psychology*, **10**, 94–101.
- Barton, J., Spelton, E., Totterdell, P., Smith, L., Folkard, S. and Costa, G. (1995). The Standard Shiftwork Index. A battery of questionnaires for assessing shiftwork-related problems. *Work and Stress*, **9**, 289–97.
- Beehr, T.A. (1976). Perceived situational modifiers of the relationship between subjective role ambiguity and role strain. *Journal of Applied Psychology*, **61**, 35–40.
- Biondi, M. and Zannino, L.G. (1997). Psychological stress, neuroimmunomodulation and susceptibility to infectious

- diseases in animals and man: a review. *Psychotherapy and Psychosomatics*, **66**, 3–26.
- Buckle, P. (1997). Upper limb disorders and work: the importance of physical and psychosocial factors. *Journal of Psychosomatic Research*, **43**, 17–25.
- Burks, T.F. (1991). Role of stress in the development of disorders of gastrointestinal motility. In *Stress: Neurobiology and Neuroendocrinology*, (eds M.R. Brown, G.F. Koob and C. Rivier), pp. 566–83. Dekker, New York.
- Cooper, C.L., Sloan, S.J. and Williams, S. (1988). *The Occupational Stress Indicator*. NFER-Nelson, Windsor, UK.
- Cox, T. (1993). *Stress and Research and Stress Management: Putting Theory to Work*. HSE Contract Report No. 61/1993. HSE Books, Sudbury.
- Crawford, J.O. and Bolas, S.M. (1996). Sick Building Syndrome, work factors and occupational stress. *Scandinavian Journal of Work, Environment and Health*, **22**, 243–50.
- Davies, N.V. and Teasdale, P. (1994). *The Costs to the British Economy of Work Accidents and Work-related Ill-health*. HSE Books, Sudbury.
- Depolo, M. and Sarchielli, G. (1987). Job insecurity, psychological well-being and social representation: a case of cost sharing. In *Proceedings of the West European Conference on the Psychology of Work and Organisation* (eds H.W. Scroiff and G. Debus). Elsevier, Amsterdam.
- Devereux, J.J., Buckle, P.W. and Vlachonikolis, G. (1999). Interactions between physical and psychosocial risk factors at work increase the risk of back disorders: an epidemiological approach. *Occupational and Environmental Medicine*, **56**, 343–53.
- Earnshaw, J. and Cooper, C.L. (1996). *Stress and Employer Liability*. Chartered Institute of Personnel and Development, London.
- Fitzgerald, L.F. and Shullman, S.L. (1993). Sexual harassment: A research analysis and agenda for the 1990s. *Journal of Vocational Behaviour*, **42**, 5–27.
- Friedmann, M. and Rosenman, R.H. (1974). *Type A Behaviour and Your Heart*. Wildwood House, London.
- Fusilier, M.R., Ganster, D.C. and Mayes, B.T. (1987). Effects of social support, role stress and locus of control on health. *Journal of Management*, **13**, 517–28.
- Gersons, B.P.R. and Carlier, I.V.E. (1992). Post-traumatic stress disorder: the history of a recent concept. *British Journal of Psychiatry*, **161**, 742–8.
- Goldberg, D. and Williams, P.A. (1988). *Users' Guide to the General Health Questionnaire*. Nelson, Windsor, UK.
- Green, K.L. and Johnson, J.V. (1990). The effects of psychosocial work organization on patterns of cigarette smoking among male chemical plant employees. *American Journal of Public Health*, **80**, 1368–71.
- Haan, M.N. (1988). Job strain and ischaemic heart disease: an epidemiologic study of metal workers. *Annals of Clinical Research*, **20**, 143–5.
- Hanson, S.L. and Pichert, J.W. (1986). Perceived stress and diabetes control in adolescents. *Health Psychology*, **5**, 439–52.
- Hebb, D.O. (1949). *The Organization of Behaviour: A Neuropsychological Theory*. Wiley, New York.
- Hockey, R. (ed.) (1983). *Stress and Fatigue in Human Performance*. Wiley, Chichester.
- Houtman, I.L.D., Bongers, P.M., Smulders, P.G.W. et al. (1994). Psychosocial stressors at work and musculoskeletal problems. *Scandinavian Journal of Work, Environment and Health*, **20**, 135–145.
- Health and Safety at Work Act*, 1974. HMSO, London.
- Management of Health and Safety at Work Regulations*, 1992. HMSO, London.
- Johnson, J.V., Hall, E.M. and Thesrell, T. (1989). Combined effects of job strain and social isolation on cardiovascular disease mortality in a random sample of the Swedish male working population. *Scandinavian Journal of Work, Environment and Health*, **15**, 271–9.
- Karasek, R.A. (1979). Job demands, job decision latitude and mental strain: implications for job re-design. *Administrative Science Quarterly*, **24**, 285–306.
- Kobasa, S.C. (1982). The hardy personality: towards a social psychology of stress and health. In *Social Psychology of Health and Illness* (eds G.S. Sanders and J. Suls), pp. 3–32. Hillsdale, Erlbaum, NJ.
- Kompier, M. and Levi, L. (1994). *Stress at Work: Causes, Effects and Prevention. A Guide for Small and Medium Sized Enterprises*. European Foundation for the Improvement of Living and Working Conditions, Dublin.
- Konar, E. and Sundstrom, E. (1985). Status demarcation in the office. In *Behavioural Issues in Office Design* (ed. J. Wineman). Van Nostrand, New York.
- Lazarus, R.S. (1981). The stress and coping paradigm. In *Models for Clinical Psychopathology* (ed. C. Eisdorfer), pp. 177–214. Spectrum, New York.
- Leather, P., Cox, T. and Farnsworth, B. (1990). Violence at work: an issue for the 1990s. *Work and Stress*, **4**, (1), 3–5.
- Leino, P.I. and Hanninen, V. (1995). Psychosocial factors at work in relation to back and limb disorders. *Scandinavian Journal of Work, Environment and Health*, **21**, 134–42.
- Levenson, J.L. and Bemis, C. (1991). The role of psychological factors in cancer onset and progression. *Psychosomatics*, **32**, 124–32.
- Leymann, H. (1996). The content and development of mobbing at work. *European Journal of Work and Organizational Psychology*, **5**, 165–84.
- Maslach, C. and Jackson, S.E. (1981). The measurement of experienced burnout. *Journal of Occupational Behaviour*, **2**, 99–113.
- Melmed, R.N. and Gelpin, Y. (1996). Duodenal ulcer: the helicobacterisation of a psychosomatic disease. *Israel Journal of Medical Sciences*, **32**, 211–16.
- Pearlin, L.I. and Schooler, C. (1978). The structure of coping. *Journal of Health and Social Behaviour*, **19** March, 2–21.
- Rhodewalt, F., Sansone, C., Hill, C.A., Chemers, M.M. and Wysocki, J. (1991). Stress and distress as a function of Jenkins Activity Survey-defined Type A behaviour and

- control over the work environment. *Basic and Applied Social Psychology*, **12**, 211–26.
- Richman, J.A. (1992). Occupational stress, psychological vulnerability and alcohol-related problems over time in future physicians. *Alcoholism: Clinical and Experimental Research*, **16**, 166–71.
- Rush, M.C., Schoel, W.A. and Barnard, S.M. (1995). Psychological resiliency in the public sector: 'hardiness' and pressure for change. *Journal of Vocational Behaviour*, **46**, 17–39.
- Sargent, L.D. and Terry, D.J. (2000). The moderating role of social support in Karasek's job strain model. *Work and Stress*, **14**, 245–61.
- Schabracq, M., Cooper, C.L., Travers, C. and van Maanen, D. (2001). *Occupational Health Psychology: the Challenge of Workplace Stress*, pp. 55–60. BPS Books, Leicester.
- Sikejima, S. and Kagamimori, S. (1998). Working hours as a risk factor for acute myocardial infarction in Japan: case-control study. *British Medical Journal*, **317**, 775–80.
- Skov, T., Borg, V. and Orhede, E. (1996). Psychosocial and physical risk factors for musculoskeletal disorders of the neck, shoulders and lower back in salespeople. *Occupational and Environmental Medicine*, **53**, 351–6.
- Sparks, K. and Cooper, C.L. (1997). The effects of hours of work on health: a meta-analytic review. *Journal of Occupational and Organisational Psychology*, **70**, 391–408.
- Spector, P.E. (1982). Behaviour in organisations as a function of employee's locus of control. *Psychological Bulletin*, **91**, 482–97.
- Spurgeon, A. (2002). Models of unexplained symptoms associated with occupational and environmental exposures. *Environmental Health Perspectives*, **110** (Suppl. 4), 601–5.
- Spurgeon, P. and Barwell, F. (2000). *Tackling the Causes of Workplace Stress. A Guide to Using the Organisational Stress Measure in the NHS*. Health Development Agency, London.
- Spurgeon, A. and Cooper, C.L. (2000). Working Time, Health and Performance. In *International Review of Industrial and Organizational Psychology* (eds C.L. Cooper and I.T. Robertson), pp. 189–222. John Wiley & Sons, Chichester.
- Wales, J.K. (1995). Does psychological stress cause diabetes? *Diabetic Medicine*, **12**, 109–12.
- Watson, D. and Pennebaker, J.W. (1989). Health complaints, stress and distress: exploring the central role of negative affectivity. *Psychological Review*, **96**, 234–54.
- Weiss, D.S. and Marmar, C.R. (1996). The Impact of Events Scale – Revised. In *Assessing Psychological Trauma and PTSD* (eds J. Wilson and T.M. Keane), pp. 399–411. Guildford.
- Williams, S. and Cooper C.L. (1998). Measuring occupational stress: development of the Pressure Management Indicator. *Journal of Occupational Health Psychology*, **3**, 306–21.
- Zapf, D., Knorz, C. and Kulla, M. (1996). On the relationship between mobbing factors and job content, social work environment and health outcomes. *European Journal of Work and Organisational Psychology*, **5**, 215–37.

Suggested further reading

- Addley, K. (ed.) (1997). *Occupational Stress. A Practical Approach*. Butterworth Heinemann, Oxford.
- Cooper, C.L. and Payne, R. (eds) (1990). *Causes, Coping and Consequences of Stress at Work*. John Wiley & Sons, Chichester.
- Grimshaw, J. (1999). *Employment and Health: Psychosocial stress in the workplace*. British Library Publications, Science Technology and Business, London.
- HSE (2001). *Tackling Work-related Stress. A Guide for Employees*. HSE Books, Sudbury.
- HSE (2001). *Tackling Work-related Stress. A Manager's Guide to Improving and Maintaining Employee Health and Well-being*. HSE Books, Sudbury.
- Sundstrom, E. (1986). *Work Places. The Psychology of the Physical Environment in Offices and Factories*. Cambridge University Press, Cambridge.
- Zalaquett, C.P. and Wood, R.J. (eds) (1997). *Evaluating Stress. A Book of Resources*. Scarecrow Press, London.

Chapter 25

The development of ergonomics as a scientific discipline

Joanne Crawford

- Definitions of and domains within ergonomics
- Ergonomics and occupational hygiene
- Methodologies used by ergonomists
- Current issues in physical ergonomics
- Problem recognition
 - Risk assessment
 - Ill health and injury data
- Problem evaluation
 - Identification of numbers involved
 - Workplace evaluation
 - Work evaluation techniques

- Control measures
 - Workplace design and human variability
 - Workplace layout
 - Tool design
 - Work design/work pacing
 - Workplace changes
- Summary
- References
- Web resources
- Further reading

Ergonomics is the scientific discipline that involves the application of theory, principles and methodologies to optimize human well-being and performance within a workplace or work system. Historically, in the UK, ergonomics was developed during the Second World War by a group of scientists, including anatomists, physiologists, engineers and psychologists, working together in the armed services (Pheasant, 1991). The main objective of this group was to maintain and improve the efficiency of the fighting man. However, it was appreciated that a multidisciplinary approach could also be used in areas outside the military, including the industrial work environment.

The Human Research Society of the UK was formed in 1949; the aim of the society was to promote interdisciplinary collaboration between scientific disciplines including engineering, physiology, psychology and anatomy (Robertson, 2000). It was appreciated that the name of this society was not really appropriate, and one of the first tasks was to develop a name appropriate to the subject area. The Ergonomics Research Society was the choice of this new society, the word *ergo-*

nomics coming from the Greek *ergos* (work) and *nomos* (natural law) (Pheasant, 1991). The Ergonomics Research Society became the Ergonomics Society in 1977, as it was found that, owing to the nature of ergonomics, members of the society were not only carrying out research but also applying ergonomics in the workplace (Robertson, 2000).

Over approximately the same time period in the USA, a similar discipline was evolving, which is now known as human factors. Today the terms ‘ergonomics’ and ‘human factors’ are synonymous as in the UK we have the Ergonomics Society and, in North America, the Human Factors and Ergonomics Society. At the international level, the International Ergonomics Association (IEA) was founded in 1959. The IEA has representation from countries throughout the world.

With regard to research output, possibly the best-known and respected journal is *Ergonomics*, first published in 1957 and taken on as the IEA official publication in 1961 (Robertson, 2000). Other publications include *Applied Ergonomics*, *Behaviour and Information Technology*, *Work and Stress* and *Ergonomics Abstracts*.

Definitions of and domains within ergonomics

A variety of definitions of ergonomics have been suggested since the inception of the discipline. These have included 'Ergonomics is the science of matching the job to the worker and the product to the user' (Pheasant, 1991). However, in a broader sense this should perhaps be 'a study of human abilities and characteristics which affect the design of equipment systems and jobs' (Clark and Corlett, 1984). Much time has been spent trying to define the discipline of ergonomics but, rather as Wilson (1995) suggests, concentration should be on the philosophy of ergonomics rather than the definition of it.

The philosophical concepts behind ergonomics are that an approach is taken to ensure that the human in the work system is considered as part of the design process using user-centred design or design for all concepts. By not considering the end user, poor fit will occur, which can result in poor fit physically, resulting in physical discomfort or damage to the individual or psychological overload of the individual worker.

The IEA has suggested domains of specialization within the discipline of ergonomics. *Physical ergonomics* is mainly concerned with human dimensions, anatomy and physiology. Topics include posture analysis, manual handling, work-related musculoskeletal disorders, work physiology and workplace layout (IEA, 2002).

Cognitive ergonomics is a domain concerned with mental process, including perception, memory, information processing and motor response and how they affect the interaction between the human and the work system. Themes included in this domain are mental workload, decision-making, skilled performance and human-computer interaction (IEA, 2002).

The final domain suggested by IEA is that of *organizational ergonomics*, including the optimizing of complex socio-technical systems, organizational structure, policies and processes. Subjects included in this domain include teamwork, participatory design, cooperative work, telework and quality management (IEA, 2002).

However, it should be appreciated that the three domains are not mutually exclusive and are constantly evolving with new areas of research and application.

Ergonomics and occupational hygiene

There are several areas in which ergonomics and occupational hygiene link together, most specifically, the area of *personal protective equipment*, in which the hygienist needs to ensure protection of the working population and the ergonomist needs to ensure the fit and compatibility of protective equipment. Other areas of overlap include the measurement of the thermal environment, whereby both professions are able to measure and monitor the environment but give different types of advice on either ventilation systems or work systems to protect the human.

The aim of this chapter is to provide a useful introduction to some of the methodologies used in physical ergonomics. It includes a list of further texts and publications that will allow the interested reader an opportunity to research further.

Methodologies used by ergonomists

Over the last 50 years, a number of different methodologies have been developed within ergonomics to either collect data from the human population or apply data to the workplace or product design. A comprehensive textbook with regard to methodologies used within ergonomics can be found in Wilson and Corlett (1995). However, a non-exhaustive list includes the following types of methodologies.

Objective or direct measures include the direct measurement of human response to particular stimuli, for example anthropometric data (human dimensions), heart rate, core temperature, maximal grip strengths and eye movement measures, whereas subjective measures include interviews, questionnaires, rating scales that can give an indication of perceptions of comfort and product preference or stress. Measurement of

behaviour when observation is made of individuals at work or within specific environments is a further tool used by ergonomists. An example of this method includes observing how individuals interact with products or, if evaluating seating, the amount of fidgeting can be measured. Finally, the use of computer or mathematical modelling has become an important method in the design of work spaces. With regard to computer modelling, software such as SAMMIE or MANNEQUIN is used. This type of software contains databases of information on body size or strength that can be used to build a computer model of a workplace. Figure 25.1 shows an example of a design from the SAMMIE program. However, as with any modelling tool, it does have its limitations and may oversimplify the design, which may still need testing in a physical mock-up.

When carrying out an ergonomics investigation there are a number of factors that need to be considered. It is apparent that when designing work systems or equipment for humans it is an essential part of the process to involve humans in the design or the evaluation of the design. From a scientific viewpoint, it is necessary to ensure that the humans involved in ergonomics investigations are representative of the end user. This includes consideration of the multicultural nature of modern society, gender, age, size and ability of all users. The number of users involved in an ergonomics

investigation should be based on the principles of experimental design to ensure that data collection is as unbiased as possible and that there can be a specific level of confidence in the results obtained.

As well as being people based, often the ergonomics investigation is also dependent on the use of job and task descriptions. This has resulted in a whole area of ergonomics called *task analysis*, which is based on the systematic recording of data collected from a work system and their representation into a formal structure. The process adopted when using *task analysis* is shown in Fig. 25.2. Data regarding the particular work task or tasks are collected via techniques including listing of specific work tasks, grouping of tasks together, the frequency or sequence of particular tasks and the work system objectives. Those involved in carrying out or supervising particular work tasks using techniques such as interviews, verbal protocol analysis and observation usually obtain these data.

One of the commonest techniques used within *task analysis* is *hierarchical task analysis* (HTA). This technique was developed by Duncan and Annett in 1967 (cited in Stammers and Shephard, 1995). As a method of task description it entails the identification of a system and its goals that could be, for example, the production schedule in a manufacturing environment. Detailed information is obtained on particular work tasks that

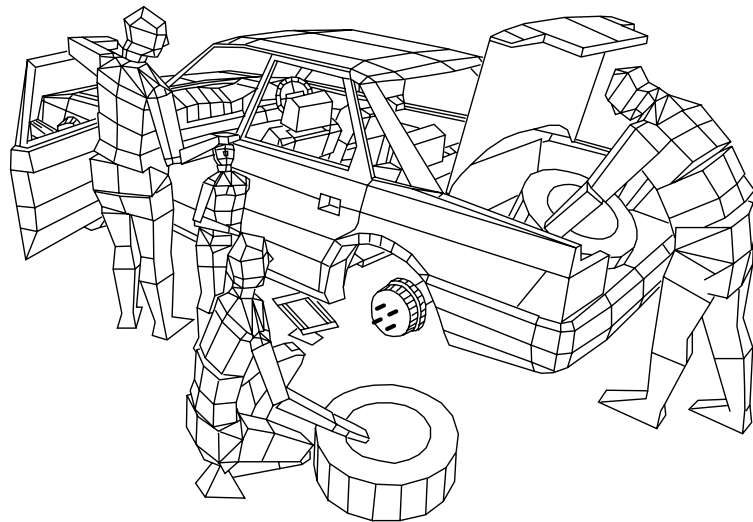


Figure 25.1 A complex car model. A hierarchical data structure enables functional as well as geometric relationships to be modelled, thus all moving parts of the model can be made to function.

working days lost due to ill health problems of the upper limbs and neck in 1995 and at an estimated cost of £200 million (HSE, 2002b). The types of injuries associated with WRULDs include medically recognized disorders such as carpal tunnel syndrome, beat hand, beat elbow and tenosynovitis. However, there are other non-specific pain syndromes that can also impact on the ability of an individual to be able to carry out either their job or general life activities. A number of workplace factors have been implicated in the aetiology of WRULDs and these are described in Table 25.1.

Both manual handling and upper limb disorders contribute a major cost to the UK economy. They are both found in a large spread of industries, not just those traditionally involved in heavy manual labour. However, with regard to WRULDs, some of the higher risk groups include workers in the meat and poultry industry, cleaning and domestic staff, and secretarial and clerical workers (HSE, 2002b). It is evident that it is essential to reduce these types of injuries owing to the number of individual workers affected and the cost of this to industry.

Table 25.1 Risk factors associated with the development of WRULDs.

Type of factor	Risk factor
Task factors	Highly repetitive work
	High levels of force required
	Awkward postures adopted
	Static loading of limbs
	Duration of task/shift
Environment factors	Working environment
	Cold
	Exposure to vibration
	Poor lighting causing adoption of poor posture
Individual factors	Poor organization, increasing psychosocial risks
	Health status, including diseases such as diabetes, arthritis, etc.
	Non-occupational factors, including sports, hobbies

From HSE (2002a) and Putz-Anderson (1988).

Problem recognition

Risk assessment

There may be several ways in which workers experiencing musculoskeletal problems can be identified. Risks may have been identified through the process of *risk assessment* for specific workplace problems such as the Display Screen Equipment (DSE) Regulations (HSE, 2002c) or the Manual Handling Operations Regulations (HSE, 2004). Both of these regulations require the risk assessment of workplaces for specific hazards.

The DSE Regulations (HSE, 2002c) require the risk assessment of any workplace where an individual worker uses a display screen or VDU habitually as part of their working day. The principle behind the introduction of this legislation was the fact that the use of computerized equipment was linked with a number of possible health problems, including the development of WRULDs and visual fatigue. Thus, a schedule for risk assessment was developed for the factors that must be risk assessed under the DSE Regulations (HSE, 2002c), including workstation analysis of the work equipment, furniture, work environment, the work routine and the specific needs of any individual workers. The DSE Regulations (HSE, 2002c) also give specific requirements for certain pieces of equipment, including seating, screens, keyboards, the work environment and the software used.

The Manual Handling Operations Regulations (HSE, 2004) also include a schedule for risk assessment of workplaces where there is a risk of injury to individual workers from any manual handling tasks in the workplace. Duties of the employer with regard to manual handling are to ensure that hazardous manual handling is avoided if it has been identified in the workplace. If the particular work tasks cannot be avoided then a risk assessment must be carried out and risk reduction measures taken. The guidance on manual handling operations gives an example of a risk assessment and factors that need to be assessed include the work task, the load being handled, the work environment and the individual worker.

The legislative requirement for risk assessment for particular workplace hazards should allow the

identification of individuals at risk from poor workplace design and work practices. However, the two sets of regulations described are for specific workplaces or work tasks and not for the general work environment where other types of hazards may exist.

Ill health and injury data

Other sources of information regarding workplace problems can include the use of ill health and workplace injury data. Where occupational health support is in place, this can be a source of data for the ergonomist to identify the existence of problem areas in the workplace or particular work tasks that are associated with ill health or injury reporting.

Other clues may also be apparent to highlight areas of risk within the workplace. These include identifying jobs that individuals are reluctant to carry out, work areas where individuals complain of discomfort, changes that the workers themselves have made, including adaptations to workplaces or equipment, and the more obvious bandages and or splints being worn in the workplace (HSE, 2002b).

Problem evaluation

Identification of numbers involved

When it is recognized that there is an ergonomics problem in the workplace, a number of tools can be applied to assess the number of individuals suffering from physical symptoms and the types of work tasks involved. Although there may be only two or three people obviously complaining of problems within the workplace, this may indicate a much larger workplace problem that needs to be addressed.

To identify the extent of a physical workplace problem, a first step may be the use of body maps to identify the body site at which the pain or discomfort is occurring. Figure 25.3 shows an example of a body map that can be presented to individual workers at specific times throughout the working day. The body map identifies the site of pain and individual workers can also be asked to

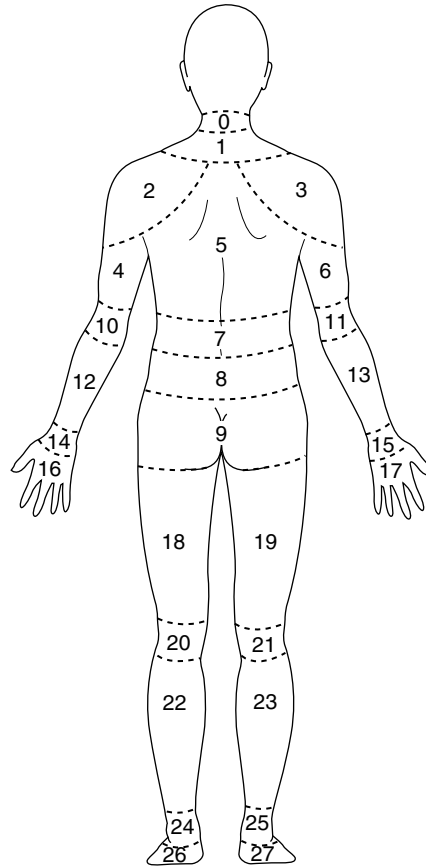


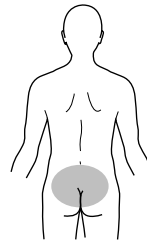
Figure 25.3 The body map for evaluating body part discomfort, either by rating or by ranking.

rate the extent of the pain on a scale. Corlett (1995) recommends that the body map is used at periods throughout the work shift to identify the problem areas and the degree of discomfort. By using the body maps at specified time periods throughout the working day, data can be plotted against time to find out whether rest breaks or changes in work routine are adequate to improve the pain and discomfort level reported. Although a simplistic technique to use, the data obtained via the use of body maps can identify the work tasks and work routines that are causing discomfort problems to the workforce.

A number of questionnaires have also been developed to identify the extent of musculoskeletal problems within the workplace. One of those is the

Questionnaire about low back trouble

The date of inquiry	____/____/____ year month day	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Sex	Female 2 Male	<input type="checkbox"/>
What year were you born	_____	<input type="checkbox"/>
How many years and months have you been doing your present type of work?	____ years * ____ months	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
On average how many hours a week do you work?	_____ hours a week	<input type="checkbox"/>
How much do you weigh?	_____ kg	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
How tall are you?	_____ cm	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Are you right-handed or left-handed?	1. <input type="checkbox"/> right-handed 2. <input type="checkbox"/> left-handed	<input type="checkbox"/>



LOW BACK

How to answer the questionnaire. In this picture you can see the appropriate position of the part of the body referred to in the questionnaire. By low back trouble is meant ache, pain or discomfort in the shaded area whether or not it extends from there to one or both legs (sciatica). Please answer by putting a cross in the appropriate box – one cross for each question. You may be in doubt as to how to answer but please do your best anyway.

<p>1. Have you ever had low back trouble (ache, pain or discomfort)?</p> <p>1 <input type="checkbox"/> NO 2 <input type="checkbox"/> YES</p>	<input type="checkbox"/>
<p>If you answer No to question 1, do not answer questions 2 – 6.</p>	
<p>2. Have you ever been hospitalized because of low back trouble?</p> <p>1 <input type="checkbox"/> NO 2 <input type="checkbox"/> YES</p>	<input type="checkbox"/>
<p>3. Have you had to change jobs or duties because of low back trouble?</p> <p>1 <input type="checkbox"/> NO 2 <input type="checkbox"/> YES</p>	<input type="checkbox"/>
<p>4. What is the total length of time that you have had low back trouble during the last 12 months?</p> <p>1 <input type="checkbox"/> 0 days 2 <input type="checkbox"/> 1 – 7 days 3 <input type="checkbox"/> 8 – 30 days 4 <input type="checkbox"/> More than 30 days, but not every day 5 <input type="checkbox"/> Every day</p>	<input type="checkbox"/>
<p>If you answer 0 days to question 4 do not answer the questions 5 – 8.</p>	
<p>5. Has low back trouble caused you to reduce your activity during the last 12 months?</p> <p>a. Work activity (at home or away from home)? 1 <input type="checkbox"/> NO 2 <input type="checkbox"/> YES b. Leisure activity? 1 <input type="checkbox"/> NO 2 <input type="checkbox"/> YES</p>	<input type="checkbox"/>
<p>6. What is the total length of time low back trouble has prevented you from doing your normal work (at home or away from home) during the last 12 months?</p> <p>1 <input type="checkbox"/> 0 days 2 <input type="checkbox"/> 1 – 7 days 3 <input type="checkbox"/> 8 – 30 days 4 <input type="checkbox"/> More than 30 days</p>	<input type="checkbox"/>
<p>7. Have you been seen by a doctor, physio therapist, chiropractor or other such person because of low back trouble during the last 12 months?</p> <p>1 <input type="checkbox"/> NO 2 <input type="checkbox"/> YES</p>	<input type="checkbox"/>
<p>8. Have you had low back trouble at any time during the last 7 days?</p> <p>1 <input type="checkbox"/> NO 2 <input type="checkbox"/> YES</p>	<input type="checkbox"/>

Figure 25.4 Low back trouble questionnaire.

Nordic Musculoskeletal Questionnaire (NMQ) developed by Kuorinka and co-workers (1987). This questionnaire has been tested for reliability and found to be within acceptable limits. The aim of this questionnaire is to obtain data on the types of symptoms occurring within the workforce and the duration of the symptoms. Figure 25.4, for example, is the section in the NMQ for information on low back symptoms.

Although the NMQ was developed in the late 1980s, this and adaptations of it are still considered very useful in workplaces at present. Questions specific to particular workplaces can be added to the questionnaire to make it a valid method of collecting symptom data in any workplace.

As both these techniques involve collecting data from a working population, it must be stressed that it is vital to ensure that as many members of a working population as possible complete the data collection to ensure that a valid sample is obtained. It is essential that non-response bias is avoided in data collection and that within the work force individuals both with and without symptoms are surveyed.

Workplace evaluation

As a first step in the evaluation of the workplace, it is important to identify if the workplace has been designed to fit the end user to allow adequate reach to tools and equipment, to check the fit of the workplace to the user and ensure that the user can see all that is required. A number of design fallacies have been suggested by Pheasant (1998) and are presented in Table 25.2. At this point, it is important to highlight that if the workplace has been designed for an individual of average dimensions, it is unlikely that it will fit many within the working population. However, this topic will be discussed further in the control measures to reduce physical ergonomic risks.

Work evaluation techniques

As mentioned in the section on risk assessment, specific schedules of assessment have been developed for the issues of manual handling and computer use in the workplace. Both the guidance notes contain helpful information on the reduction

Table 25.2 Pheasant's five fundamental fallacies.

The design is satisfactory for me – it will therefore be satisfactory for everybody else
The design is satisfactory for the average person – it will therefore be satisfactory for everybody else
The variability of human beings is so great that it cannot possibly be catered for in any design – but since people are wonderfully adaptable, it doesn't matter anyway
Ergonomics is expensive and since products are actually purchased on appearance and styling, ergonomic considerations may conveniently be ignored
Ergonomics is an excellent idea – I always design things with ergonomics in mind – but I do it intuitively and rely on my common sense so I don't need tables of data or empirical studies

From Pheasant (1998).

of risks found (HSE, 2002a, 2004). In 2002, the Health and Safety Executive developed a risk filter to identify musculoskeletal problems within the workplace and particular work tasks that need further assessment (HSE, 2002b). Figure 25.5 shows the *risk filter* that can be used to identify if further assessment is required for particular work tasks. The guidance on upper limb disorders does also include a more in-depth risk assessment to evaluate workplace risks.

From Table 25.1, the workplace risks highlighted include task factors, environment factors and individual factors. From the task factors it is essential to identify levels of repetition within work tasks. There has been difficulty in actually stating what high repetition and low repetition are, but general agreement has been reached in stating that a low repetitive task is one in which it takes 30 s or more to carry out one complete cycle of the work task and no more than 50% of the cycle time is spent on the same type of movements or subtask; high repetitive tasks are those with a complete cycle time of less than 30 s and more than 50% of the cycle time is spent on the same type of movements or subtask (Putz-Anderson, 1988).

Force requirements for particular work tasks are also difficult to quantify, as they are dependent on the tools in use, the strength of the individual worker and the postures that have to be adopted when applying force. Although direct measures can be made through the use of electromyography (EMG), this is not always transferable to the working environment from the laboratory. Instead, aspects of force have been assessed via the examination of work tasks, including the weights

of tools and equipment, whether fast movements are required and the postures adopted when applying force. In specific posture analysis methodologies, posture and force are considered together.

A number of different posture analysis methodologies have been developed to analyse the working postures adopted by individual workers. These include direct observation of the work and worker, and notation of the postures adopted.

One of the most recent developments in terms of posture analysis is the *quick exposure checklist* (QEC) developed by Li and Buckle (1999). The aim of the QEC is to develop a usable tool for non-experts that could be used to quickly assess exposure to musculoskeletal risk before and after any interventions have been made. One assessor uses the tool and there is a section for the individual worker to complete with regard to exposure to risk factors. At present, the QEC is still going through a process of development and testing, but does look positive for use by non-ergonomists and other non-specialists in the working environment.

Other postural analysis methods used include the Ovaka Working Analysis System (OWAS) developed by the Finnish Institute of Occupational Health (1992). The OWAS method involves observation of working postures and the recording of postures based on the position of the back, legs and arms, and the use of force during individual work tasks. These data are collated and transcribed into a risk matrix whereby action categories are then calculated. OWAS provides four action categories from 'no corrective measures' to 'corrective measures immediately'. The advantage of the OWAS

RISK FILTER			
Task: _____			
Assessor: _____			
Date: _____		Location/work area: _____	
<p>IF YOU ANSWER YES TO ANY OF THE STEPS, YOU SHOULD THEN MAKE A FULL RISK ASSESSMENT OF THE TASK. REMEMBER TO CONSIDER EACH OF THE BODY PARTS OF THE UPPER LIMBS (FINGERS, HANDS, WRISTS, ARMS, SHOULDERS AND NECK). ANSWER ALL QUESTIONS</p>			
Step 1: Signs and symptoms			
Are there any: <ul style="list-style-type: none"> • Medically diagnosed cases of ULD in this work? • Complaints of aches or pains? • Improvised changes to work equipment, furniture or tools? 	Are any of these present?	YES <input type="checkbox"/> NO <input type="checkbox"/>	Move on to Step 2
Step 2: Repetition			
Are there any repetitive elements such as: <ul style="list-style-type: none"> • Repeating the same motions every few seconds? • A sequence of movements repeated more than twice per minute? • More than 50% of the cycle time involved in performing the same sequence of motions? 	For more than 2 hours total per shift?	YES <input type="checkbox"/> NO <input type="checkbox"/>	Move on to Step 3
Step 3: Working postures			
Are there any working postures such as: <ul style="list-style-type: none"> • Large range of joint movement such as side to side or up and down? • Awkward or extreme joint positions? • Joints held in fixed positions? • Stretching to reach items or controls? • Twisting or rotating items or controls? • Working overhead? 	For more than 2 hours total per shift?	YES <input type="checkbox"/> NO <input type="checkbox"/>	Move on to Step 4
Step 4: Force			
Are there any forces applied such as: <ul style="list-style-type: none"> • Pushing, pulling, moving things (including with the fingers or thumbs)? • Grasping/gripping? • Pinch grips, i.e. holding or grasping objects between thumb and finger? • Steadying or supporting items or work pieces? • Shock and/or impact being transmitted to the body from tools or equipment? • Objects creating localized pressure on any part of the upper limb? 	Sustained or repeated application of force for more than 2 hours total per shift?	YES <input type="checkbox"/> NO <input type="checkbox"/>	Move on to Step 5
Step 5: Vibration			
<ul style="list-style-type: none"> • Do workers use any powered hand-held or hand-guided tools or equipment or do they hand-feed work pieces to vibrating equipment? 	Regularly (i.e. at some point during most shifts)?	YES <input type="checkbox"/> NO <input type="checkbox"/>	
<p><i>If you answer yes to any of the steps, you should make a full risk assessment of the task</i></p>			

Figure 25.5 Risk filter.

technique is that one person can observe a number of different tasks throughout the working period. However, skill and practice are required with this technique and it is recommended that users are trained and tested frequently, and that all observations are backed up by the use of video cameras.

Although OWAS is useful as a whole-body posture analysis technique, a more in-depth posture analysis tool was developed to identify the postures associated with WRULDs. The Rapid Upper Limb Assessment (RULA) tool was developed by McAtamney and Corlett (1993) and again involves direct observation by the assessor to evaluate the postures adopted during working tasks. Figure 25.6 shows the limbs that are assessed when using the RULA technique. As can be seen from the figure, the limbs are split into groups A and B and assessed. In addition, a muscle use and force score is added into the RULA analysis and a grand score table is used to calculate the action level of the postures analysed. The action levels of the RULA analysis again are from 1 to 4 and are presented in Table 25.3.

The use of posture analysis methods such as those described above allows assessors to be able to quantify the risk to workers from poor posture at work. By using analysis tools such as OWAS and RULA, work tasks that are identified as high risk will be recognized under the action categories and a priority of task changes given. However, all the methods described for posture analysis are subjective and any assessors must be trained in their use and their competency assessed. For direct measures of posture assessment see Wilson and Corlett (1995).

Control measures

When physical ergonomics risks have been identified, control measures need to be considered to reduce the level of risk in the workplace. Based on good ergonomics practice, it is essential to take a participatory approach whereby workers affected by any assessment or interventions are consulted throughout the process of any workplace changes. Ergonomists appreciate that the individuals who know the job well are those carry-

ing it out on a daily basis and as such are a mine of useful information. When control measures or intervention are also designed, it is important to include the workforce in discussions, as the success of any intervention is dependent on the cooperation of the worker.

When taking the ergonomics approach to problem solving, all parts of the work system and the end user are considered which enables the most effective way of dealing with physical ergonomics problems at work. This again is consistent with the basic philosophy of designing for the end user.

Workplace design and human variability

When we think about designing workplaces or products, we are struck immediately by the need to design to fit the vast majority of individuals. Failure to do this results in a mismatch between the user and the workplace or product. To help develop usable equipment, we use anthropometric data and apply these to the design of our workplace or product. When asked about our design criteria, the comment is often that we design to fit the 'average' person. However, by designing to fit the 'average' you will have a workplace or product that will fit very few people and be too big for those below average size and too small for those above average size. When human dimensions such as stature or height are examined, with a large sample of measurement, the data are normally distributed. What this means is that with a dataset and a large sample, percentiles at different points on the normal distribution can be calculated, for example the 5th percentile and the 95th percentile that will allow us to design to fit the middle 90% of the population or the majority of the population. Table 25.4 is an anthropometric dataset for British adults.

The use of anthropometric data tables will allow the development of workstations and equipment that will fit 90% of the population. The data given in such tables can also be manipulated to increase the population for whom you are designing. For example, with personal protective equipment it is vital to protect 100% of the workforce; therefore equipment must be designed to fit and protect everyone.

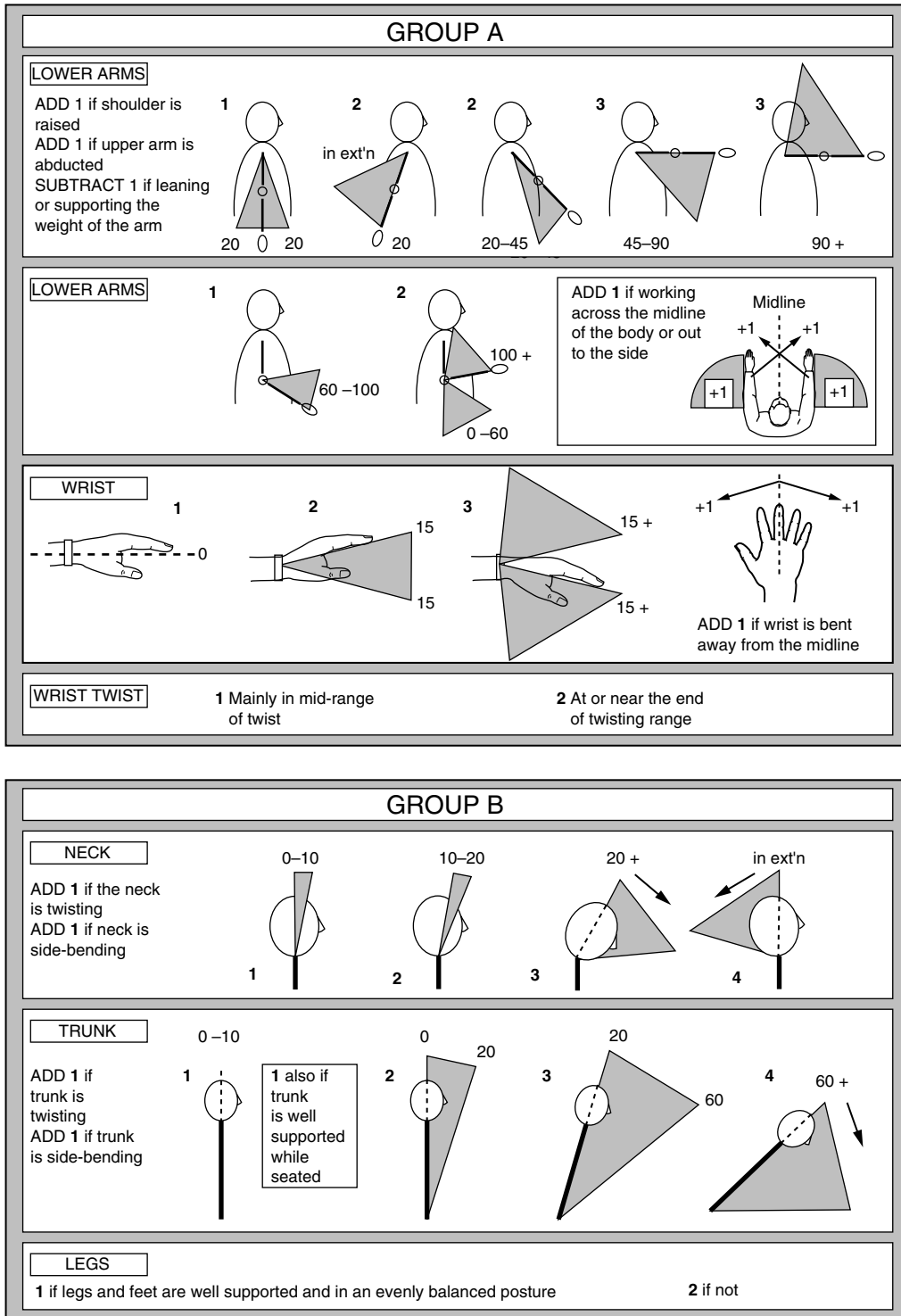


Figure 25.6 Items for assessment when using the RULA method.

Table 25.3 Action levels for the RULA analysis.

<i>Action level</i>	<i>Outcome</i>
Action level 1	A score of one or two indicates that the posture is acceptable if it is not maintained or repeated for long periods
Action level 2	A score of three or four indicates further investigation is needed and changes may be required
Action level 3	A score of five or six indicates investigation and changes are required soon
Action level 4	A score of seven or more indicates investigation and changes are required immediately

From McAtamny and Corlett (1993).

Table 25.4 Anthropometric data for British men and women aged 19–65 years (mm).

	<i>Men</i>			<i>Women</i>		
	<i>5th percentile</i>	<i>50th percentile</i>	<i>95th percentile</i>	<i>5th percentile</i>	<i>50th percentile</i>	<i>95th percentile</i>
Standing height	1625	1740	1855	1505	1610	1710
Eye height	1515	1630	1745	1405	1505	1610
Shoulder height	1315	1425	1535	1215	1310	1405
Elbow height	1005	1090	1180	930	1005	1085
Hip height	840	920	1000	740	810	885
Knuckle height	690	755	825	660	720	780
Fingertip height	590	655	720	560	625	685
Sitting height	850	910	965	795	850	910
Sitting eye height	735	790	845	685	740	795
Sitting shoulder height	540	595	645	505	555	610
Sitting elbow height	195	245	295	185	235	280
Thigh thickness	135	160	185	125	155	180
Buttock–knee length	540	595	645	520	570	620
Buttock–popliteal length	440	495	550	435	480	530
Knee height	490	545	595	455	500	540
Popliteal height	395	440	490	355	400	445
Shoulder breadth (bideltoid)	420	465	510	355	395	435
Shoulder breadth (biacromial)	365	400	430	325	355	385
Hip breadth	310	360	405	310	370	435
Chest (bust) depth	215	250	285	210	250	295
Abdominal depth	220	270	325	205	255	305
Shoulder–elbow length	330	365	395	300	330	360
Elbow–fingertip length	440	475	510	400	430	460
Upper limb length	720	780	840	655	705	760
Shoulder–grip length	610	665	715	555	600	650
Head length	180	195	205	165	180	190
Head breadth	145	155	165	135	145	150
Hand length	175	190	205	160	175	190
Hand breadth	80	85	95	70	75	85
Foot length	240	265	285	215	235	255
Foot breadth	85	95	110	80	90	100
Span	1655	1790	1925	1490	1605	1725
Elbow span	865	945	1020	780	850	920
Vertical grip reach standing	1925	2060	2190	1790	1905	2020
Vertical grip reach (sitting)	1145	1245	1340	1060	1150	1235
Forward grip reach	720	780	835	650	705	755

From Pheasant (1998).

Defining your user population is the first stage in applying data. There are many datasets available but first you must ensure that you know the population for whom you are designing. Consideration needs to be made of the age of the group, the sex of the group (is it only a male workforce or only female?) and the cultural or ethnic background of the group. When we consider data from different countries, the population in the USA is both bigger and heavier than in the UK.

When datasets are chosen, the ages of the data to be used must be examined. With regard to population change in the UK and other countries, there has been a growth in the population in terms of stature and weight. Therefore, data that we use must be as up-to-date as possible to ensure they are representative of the present population.

When using anthropometric data, it is usual to accommodate 90% of the population. We do this by designing workplaces to fit between the 5th percentile and 95th percentile from the data sheets. Anthropometric criteria fall into three categories: clearance, reach and posture, whereby reach distances to locate controls or equipment are usually designed to fit the smallest or 5th percentile user; clearances under desks and doors are usually designed to fit the largest or 95th percentile user, and posture with which there is often a trade-off between the expense of adjustable workstations when possible or an overall compromise within the design, which may reduce the percentage of the population that will fit the workplace.

By using relevant anthropometric datasets, workplace or equipment design can be improved upon to ensure that the workplace fits the end user. There will always be cases when individuals are either too small or too large to fit the workplace; however, in such cases the purchase of individually fitted equipment may be necessary.

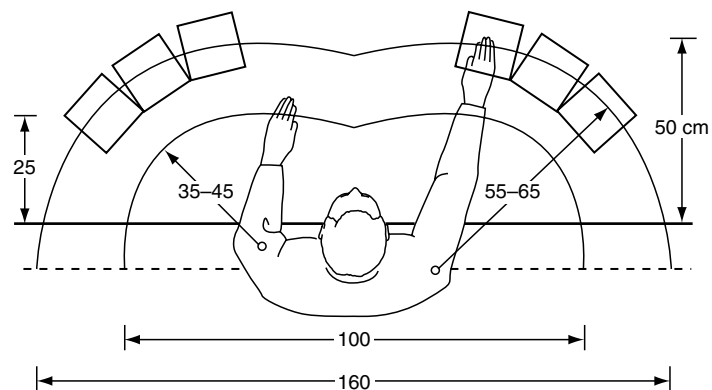
In terms of practical application of anthropometric data, its use can enhance the working environment by ensuring that individual workers can adopt good posture and reach all the relevant tools in the workplace. This is relevant for both manual handling and upper limb risks; workplaces can be designed to reduce the heights from/to which loads are handled and to ensure that overreaching does not occur.

For a thorough description of the application of anthropometric data, the reader is guided to Pheasant (1998).

Workplace layout

Ensuring that the workplace dimensions are designed to fit the user is one of the first components in terms of control measures; closely linked to this is the assessment of the workplace layout and ensuring that the individual worker can reach relevant tools and components without risk. Figure 25.7 shows a workplace envelope describing the horizontal arc of grasp on a work surface. The dimensions used in workplace envelopes are based on the 5th percentile reach distances of the working population and thus allow everyone with

Figure 25.7 Horizontal arc of grasp and working area at tabletop height. The grasping distance takes account of the distance from shoulder to hand; the working distance only elbow to hand. The values include the 5th percentile and so apply to men and women of less than average size.



reach distances over the 5th percentile dimension to reach the tools or components in front of them. There is also a vertical arc of reach above the work surface in which components or tools must be placed to allow the majority of the population to reach them.

The principles of workplace layout were described by Sanders and McCormick (1992). Four basic principles for the layout of components and equipment are described and include: the *importance principle*, which requires the placing of vital components at the most convenient or nearest location; the *frequency-of-use principle*, which requires that the most frequently used components or equipment are located closest to the user; the *function principle*, which requires that displays and controls with similar functions are grouped as a group; and, finally, the *sequence-of-use principle*, which requires that components should be arranged in the order that they are needed according to the sequence procedures being undertaken (for example, building a product) (Sanders and McCormick, 1992). It is apparent that each of the principles can allow a different workplace layout to be designed; however, often a trade-off has to be made between the workplace principles. For example, the use of emergency stop buttons is vital to the safe running of equipment and thus based on the importance principle that they should be closest to the user. However, this may result in positioning the emergency stop in a place where it can be accidentally activated. Hence, a trade-off has to be made in the positioning of the control.

In the development of workplace layouts, it is important to consider not only the layout of controls and equipment but also to ensure that the workplace dimensions fit the working population and the work activities are identified through the use of *task analysis*.

Tool design

In conjunction with the anthropometric data and workplace layout, tool design must also be considered as a control measure in the reduction of risk for musculoskeletal health. Often when we examine basic tools such as shovels or pliers, their design has not changed, although the tasks they are used

for in different working environments varies. When examining workplace tools, a number of risks can be identified by observation alone. These include: static loading, when the individual worker has to support the tool while using it; continuous pressure on the soft tissue of the hand; exposure to cold or vibration; and awkward hand positions due to the handle design and handles that create pinch points or stretching of the hand to grip handles (Putz-Anderson, 1988).

From the ergonomics viewpoint, a number of tool design principles can be considered. The principles are often based on hand biomechanics to ensure that a good posture can be maintained when using tooling. The principles include being able to maintain a straight wrist when using tools, so that the wrist is not bent towards the arm or in the direction of the thumb or little finger. This will allow the best presentation of the tool and can be achieved by bending handles on tools to ensure that the wrist can remain straight when in use.

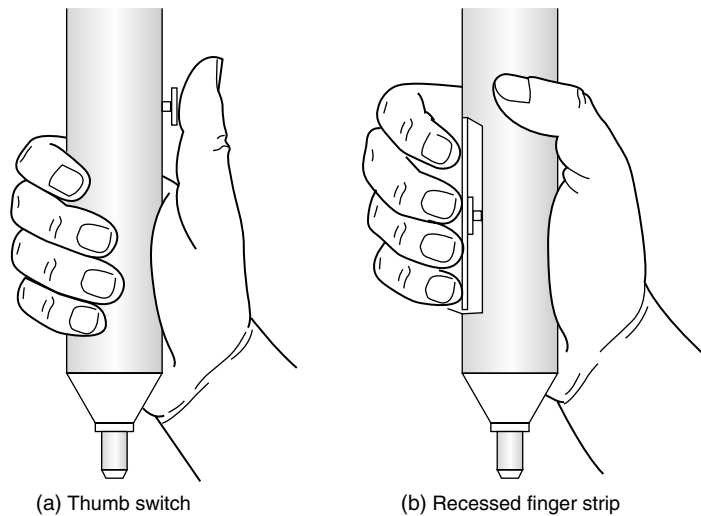
Tissue compression, especially of the soft tissue in the palm and fingers, should also be avoided. This can be achieved by ensuring that tool handles are not too short, ending in the palm of the hand where compression will occur.

Repetitive finger action from tools with triggers should also be avoided. Figure 25.8 shows a redesign of a control panel from a single thumb switch to a switch in which all the fingers can be used. Many tools are designed with a single finger trigger, which, if used repetitively, can result in fatigue and damage over time. Instead, tools in which power is used should be chosen to ensure that controls are designed for use by more than one finger.

If it is identified that the workforce is having to hold and support tools and through this creating a static load, some thought must go into the use of tool supports or suspending the tools above the workplace. This has become more common in practice and does reduce the need to hold tooling throughout the work cycle. However, in using suspension, care should be taken to ensure that all workers can reach the tools without stretching.

Owing to changes in technology and production, the types of tools in use in the workplace are constantly changing. For example, powered torque wrenches are now quite common in manufacturing

Figure 25.8 Thumb-operated and finger-strip-operated pneumatic tool. Thumb operation results in overextension of the thumb. Finger-strip control allows all the fingers to share the load and the thumb to grip and guide the tool.



environments. Care must be taken when examining the use of tools to ensure that their use is not creating new posture problems if the workplace has not been designed for them and individuals may be working at a height that is too high or too low. The whole work system, including workplace dimensions, layout and tooling, must be considered altogether, not each factor individually.

With regard to workplace changes that can be made when a problem is found, practical changes are not always expensive. Putz-Anderson (1988) suggests a threefold approach when tooling is a problem. First, altering the tools, as mentioned previously, changing the presentation of the handles to ensure that a straight wrist can be achieved. Second, changing the presentation of the component to the worker to ensure that it is accessible to work on without the adoption of poor posture. Finally, moving the position of the worker in relation to the part.

Work design/work pacing

It is not only the physical working environment that needs to be considered when making changes to reduce the risk of musculoskeletal injury. Work organization issues also need to be examined, including work rest scheduling and the level of overtime carried out by workers. When overtime is common, the duration and exposure to poor work environments obviously increases and thus

increases the level of risk. Job rotation has been suggested as a means of reducing exposure to high-risk jobs. This can be achieved when there are a variety of tasks available to do. However, in much of modern industry, the same muscle groups are used even when the work tasks are different and it is difficult to achieve a change in muscle groups. Instead, better workplace design should be used to try and achieve the best fit possible for the majority of the workforce.

Training and education of the workforce should also be made a priority. Any workplace changes or work process changes are likely to involve the setting up a system of work based on best practice. It is evident that the workforce must be trained in the use of best practice and the reasons for the change explained. Again, this is based on the participatory approach, for which it is essential to achieve best practice in the workplace and worker involvement.

Workplace changes

When making changes to the workplace, as mentioned, a participatory approach is essential. Workers should be involved in the development and design of new workplaces; they are the end users of the workplace and having their input at mock-up stages of workplace design is essential to ensure a fit between worker and workplace.

Summary

This chapter has aimed to introduce ergonomics to occupational hygienists and briefly cover the history of ergonomics, aims and design philosophy. The topic areas within ergonomics are very broad and this chapter has covered the area of physical ergonomics and dealing with musculoskeletal problems identified at work. The reader is guided to further texts that will give more in-depth information with regard to some of the methodologies discussed.

References

- Clark, T.S. and Corlett E.N. (1984). *The Ergonomics of Workspaces and Machines: A Design Manual*. Taylor & Francis, London.
- Finnish Institute of Occupational Health (1992). *OWAS, A Method for the Evaluation of Postural Load During Work*. Publication Office, Helsinki.
- HSE (2002a). *Getting to Grips with Manual Handling, INDG143 (rev 1) 4/02*. HSE Books, Sudbury, Suffolk.
- HSE (2002b). *Upper limb disorders in the workplace, HSG60rev*. HSE Books Sudbury, Suffolk.
- HSE (2002c). Work with display screen equipment, Health and Safety (Display Screen Equipment) Regulations 1992 as amended by the Health and Safety (Miscellaneous Amendments) Regulations 2002, *Guidance on the Regulations, L26*, HSE Books, Sudbury, Suffolk.
- HSE, 2004, *Manual Handling, Manual Handling Operations Regulations 1992 (as amended) Guidance on the Regulations L23*. HSE Books, Sudbury, Suffolk.
- Kuorinka, I., Jonsson, B., Kilbom, A., Vinterberg, H., Biering Sorensen, F., Andersson, G. and Jorgensen, K. (1987). Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. *Applied Ergonomics*, **18**, 3, 233–7.
- Li, G. and Buckle, P. (1999). *Evaluating Change in Exposure to Risk for Musculoskeletal Disorders – A Practical Tool*. HSE Contract Report 251/1999. HSE Books, Sudbury, Suffolk.
- McAtamney, L. and Corlett, E.N. (1993). RULA: a survey method for the investigation of work related upper limb disorders. *Applied Ergonomics*, **24**, 91–9.
- Pheasant, S. (1991). *Ergonomics, Work and Health*. Taylor & Francis, London.
- Pheasant, S. (1998). *Bodyspace, Anthropometry, Ergonomics and the Design of Work*, 2nd edn. Taylor & Francis, London.
- Putz-Anderson, V. (1988). *Cumulative Trauma Disorders: A Manual for Musculoskeletal Diseases of the Upper Limbs*. Taylor & Francis, London.
- Robertson, S.A. (2000). United Kingdom: The ergonomics society. In *International Encyclopaedia of Ergonomics and Human Factors* (ed. W. Karwowski), p. 178. Taylor & Francis, London.
- Sanders, M.S. and McCormick, E.J. (1992). *Human Factors in Engineering and Design*, 7th edn. McGraw-Hill International, London.
- Stammers, R.B. and Shepherd, A. (1995). Task analysis. In *Evaluation of Human Work: a Practical Ergonomics Methodology* (eds J.R. Wilson and E.N. Corlett). Taylor & Francis, London.
- Wilson, J.R. (1995). A framework and a context for ergonomics methodology. In *Evaluation of Human Work: a Practical Ergonomics Methodology* (eds J.R. Wilson and E.N. Corlett), Taylor & Francis, London.
- Wilson, J.R. and Corlett E.N. (eds) (1995). *Evaluation of Human Work: a Practical Ergonomics Methodology*. Taylor & Francis, London.

Web resources

- Human Factors and Ergonomics Society (2002). <http://www.hfes.org>
- IEA (2002). <http://www.iea.cc>
- The Ergonomics Society (2002). <http://www.ergonomics.org.uk>

Further reading

- Dul, J. and Weerdmeester, B. (2001). *Ergonomics for Beginners: a Quick Reference Guide*. Taylor & Francis, London.
- Kirwan, B. and Ainsworth, L.K. (eds) (1992). *A Guide to Task Analysis: the Task Analysis Working Group*. Taylor & Francis, London.
- Sanders, M.S. and McCormick, E.J. (1992). *Human Factors in Engineering and Design*, 7th edn. McGraw-Hill International, London.
- Pheasant, S. (1991). *Ergonomics, Work and Health*. Taylor & Francis, London.
- Pheasant, S. (1998). *Bodyspace, Anthropometry, Ergonomics and the Design of Work*, 2nd edn. Taylor & Francis, London.
- Wilson, J.R. and Corlett E.N. (eds) (1995). *Evaluation of Human Work: a Practical Ergonomics Methodology*. Taylor & Francis, London.

Chapter 26

Dermal exposure assessment

John W. Cherrie

Introduction
Dermal uptake
A conceptual model of exposure
Measuring exposure

Modelling exposure
Evaluating risks
References

Introduction

Occupational hygienists have always recognized the possibility of workers being exposed to chemicals by routes other than inhalation. The skin cancer hazards from exposure to soot, tar, mineral oils and other sources of polycyclic aromatic hydrocarbons have been known for many decades. The introduction of tetraethyl lead in the early 1920s resulted in many workers suffering from severe encephalopathy from this organic lead compound that readily dissolves in fat and is absorbed through the skin. More recently, attention has focused on dermal exposure from pesticides and biocides. In most instances, once the hazard was recognized there were considerable efforts made to reduce the risks in professions exposed to these agents. This has been done by changing the processes to minimize contact with the skin or, as was the case with mineral oils, to attempt to eliminate the carcinogenic compounds. However, in most instances personal protective equipment has been the main approach used to control exposure.

Despite having a good understanding of the hazards arising from skin contact, the techniques for risk assessment have been less well developed. The attempts at control have been based on common sense without rigorous quantification of dermal exposure. Sampling and analytical methods have been extensively developed and subsequently standardized for inhalation risks, whereas this has not happened to the same extent for dermal exposure measurement. Perhaps this is not surpris-

ing as it was generally inhalation exposure that determined the risk to those exposed and for many years it was believed that the intact skin was impervious to most chemicals. However, better control of the contaminants in the air and increasing use of products containing less volatile substances has focused attention on dermal exposure. In addition, we have a clearer understanding of the importance of the dermal route for chemical exposure.

In 1962 the American Conference of Governmental Industrial Hygienists introduced the concept of the skin notation to 'indicate that the liquid compound can penetrate the unbroken skin and cause systemic effects' (La Nier, 1984). This indication of the potential hazard arising from dermal exposure has subsequently been adopted into most national exposure limit systems. It has provided an indicator of situations when dermal exposure can be important. In such cases it has often been suggested that biological monitoring could provide a possible investigative approach. However, increasingly there are tools that can be used to directly evaluate dermal exposure.

This chapter outlines the processes by which chemicals are taken up through the skin, provides a conceptual framework to help analyse how people come into contact with hazardous substances and, finally, outlines the practical methods available to measure and assess the risks from dermal exposure. The text focuses on systemic uptake of chemicals through the skin rather than local effects such as dermatitis or skin cancer.

However, the principles of assessment and measurement apply equally to the latter situations.

Dermal uptake

The structure of the skin is described in Chapter 4. Hazardous substances that deposit on the skin land in the 'skin contamination layer', which comprises the wet oily layer covering the stratum corneum. These contaminants may then enter the body if they are dissolved in the skin contamination layer and they have physical and chemical properties that enable them to diffuse through the stratum corneum. Solid particles cannot pass through the unbroken skin and they must first dissolve in the skin contamination layer before they can be taken up through the skin. Diffusion through the skin is driven by a concentration gradient between the skin contamination layer and the peripheral blood supply. At steady state, this process is described by Fick's law such that the flux of contaminant (J) in, for example, $\text{mg cm}^{-2} \text{h}^{-1}$:

$$J = k_p dC \quad (26.1)$$

where k_p is the permeability constant and dC the concentration gradient between the skin contamination layer and the peripheral blood supply, which we can approximate to the concentration in the skin contamination layer if we assume that the concentration in the blood is negligible by comparison.

From the time that dermal exposure first begins there will be a delay until steady-state diffusion is achieved. This lag time, analogous to a breakthrough time for skin protective equipment, may range from seconds up to several hours. Of course, after exposure ceases there will be a corresponding period when uptake continues because there is still a concentration gradient between the skin contamination layer and the peripheral blood supply.

The preceding analysis suggests that the key variables that should describe uptake of chemicals via the skin are the concentration of the substance in the skin contamination layer (C_{Sk}), the area of skin exposed (A_{Sk}) and the duration of exposure (t_{Sk}). In addition, the mass of the hazardous substance in the skin contaminant layer (M_{Sk}) may

also be important. This is because Equation 26.1 assumes that there is an unlimited supply of chemical for diffusion. If there is less material available in the skin contaminant layer than could be taken up then the total absorbed will be the mass in the contaminant layer.

Cherrie and Robertson (1995) suggested a simple dermal exposure metric (E_{Sk}) based on this approach. They suggested the definition:

$$E_{\text{Sk}} = \int_{t=0}^T \oint C_{\text{Sk}} \, ds \, dt \quad (26.2)$$

where s is the surface area and T the time of the end of exposure.

With this approach the mass uptake through the skin (U_{Sk}) is then just:

$$U_{\text{Sk}} = k_p E_{\text{Sk}} \quad (26.3)$$

If we can assume that the concentration is constant over the exposed area throughout the exposure period then Equation 26.2 simplifies to:

$$E_{\text{Sk}} = C_{\text{Sk}} s t_{\text{Sk}} \quad (26.4)$$

Based on this, we can make simple predictions of the quantity of a substance that might be taken up from a given exposure scenario. For example, if it is assumed that a worker in a rubber plant mixes pure toluene with rubber compound eight times per day, on each occasion for 10 min. During this time he occasionally splashes the skin on his hands and forearms; based on our observation of the task let us assume that 5 cm^2 of skin is covered with liquid toluene for 5 min each time he mixes. From human volunteer experiments carried out by Kežić *et al.* (2001) the flux through the skin from such exposure (i.e. $J = k_p dC$) would be approximately $5 \mu\text{g cm}^{-2} \text{min}^{-1}$. Over the whole day we expect that the dermal toluene uptake would be 1.0 mg, i.e. $0.005 \times 5 \times 5 \times 8$. To put this into perspective we can estimate the inhalation uptake (U_{Inh}) from:

$$U_{\text{Inh}} = C_{\text{Inh}} t_{\text{Inh}} V_a \quad (26.5)$$

where C_{Inh} is the inhalation exposure level measured in the breathing zone, t_{Inh} is the duration of the inhalation exposure and V_a is the alveolar ventilation rate.

If the inhalation exposure level was 50 mg m^{-3} , the exposure lasted for 10 min on each occasion and the alveolar ventilation rate was $0.007 \text{ m}^3 \text{ min}^{-1}$ then the inhalation uptake over the eight mixing tasks was 28 mg. Therefore, we know that in this scenario the dermal exposure only contributes about 3% of the total uptake of toluene into the body.

Of course, the approach we have described is a simplification and the actual dermal uptake will depend on several other factors associated with the individual and the type of exposure. For example, if the skin is occluded, i.e. covered by, for example, a glove after the person is exposed, then the uptake will be greater. Similarly, when someone is exposed to a mixture of solvents then the net effect may be to increase or decrease the permeation rate for some of the components. Damage to the skin and the degree of hydration are also important determinants of personal risk for dermal uptake. However, ignoring such personal factors is analogous to the situation for inhalation exposure where we measure the concentration of the substance in the person's breathing zone without reference to the physiological factors that will determine what fraction of that inhaled substance will actually be absorbed in the body.

Absorption of vapours through the skin is generally of minor importance. For example, Brooke *et al.* (1998) exposed volunteers to 50 ppm of toluene in t-shirt and shorts and compared the uptake from dermal exposure alone and to that from inhalation and dermal exposure combined using a range of biological monitoring techniques. Estimated whole body dermal exposure contributed about 2% of the total exposure in this case. They suggest that the permeability coefficient for vapour exposure may be estimated from the following equation:

$$\log(k_{p,\text{vap}}) = 3.434 - 0.852 \log VP - 0.335 \log K_{ow} \quad (26.6)$$

where VP is the vapour pressure (Pa) and K_{ow} the octanol–water partition coefficient.

For some compounds the predicted $k_{p,\text{vap}}$ may be relatively large, for example the predicted values are 12 and 14 cm h^{-1} for the glycol ethers 2-methoxyethanol and 2-ethoxyethanol respect-

ively. In such situations dermal uptake from vapours is likely to be important. Also, in situations when workers may enter atmospheres with high inhalation exposure levels wearing breathing apparatus, the dermal exposure component may again be an important contribution to total exposure.

Sometimes the contaminant may be dissolved in water, for example in the case of toluene in the waste water from a chemical process where toluene-containing products are produced. In such cases the saturated concentration of the contaminant in water provides an upper limit on the concentration driving uptake by diffusion. A number of regression equations have been developed to estimate either the permeability constant ($k_{p,\text{water}}$) or the flux in this situation. These are generally based on *in vitro* skin permeation studies. The following equation was published by McKone and Howd (1992):

$$k_{p,\text{water}} = MW^{-0.6} \times \left[0.33 + \frac{\delta_{\text{derm}}}{2.4 \times 10^{-6} + 3 \times 10^{-5} K_{ow}^{0.8}} \right] \quad (26.7)$$

where δ_{derm} is the thickness of the stratum corneum (often assumed to be about 0.0025 cm) and MW is the molecular weight of the substance.

A conceptual model of exposure

A theoretical model has been developed to describe the way people may become exposed to chemicals (Schneider *et al.*, 1999). This is a source–receptor model, in which there are a number of environmental compartments and potential pathways for hazardous substances to pass from the source to the skin of the exposed person. This model provides a basis to analyse how exposure may arise.

Sources may emit into the work environment via four pathways that link the source of exposure from the process or activity being undertaken to the person (Fig. 26.1). These emissions are:

- to the air compartment (ϵ_{Air});
- to surfaces in the environment (ϵ_{Su});

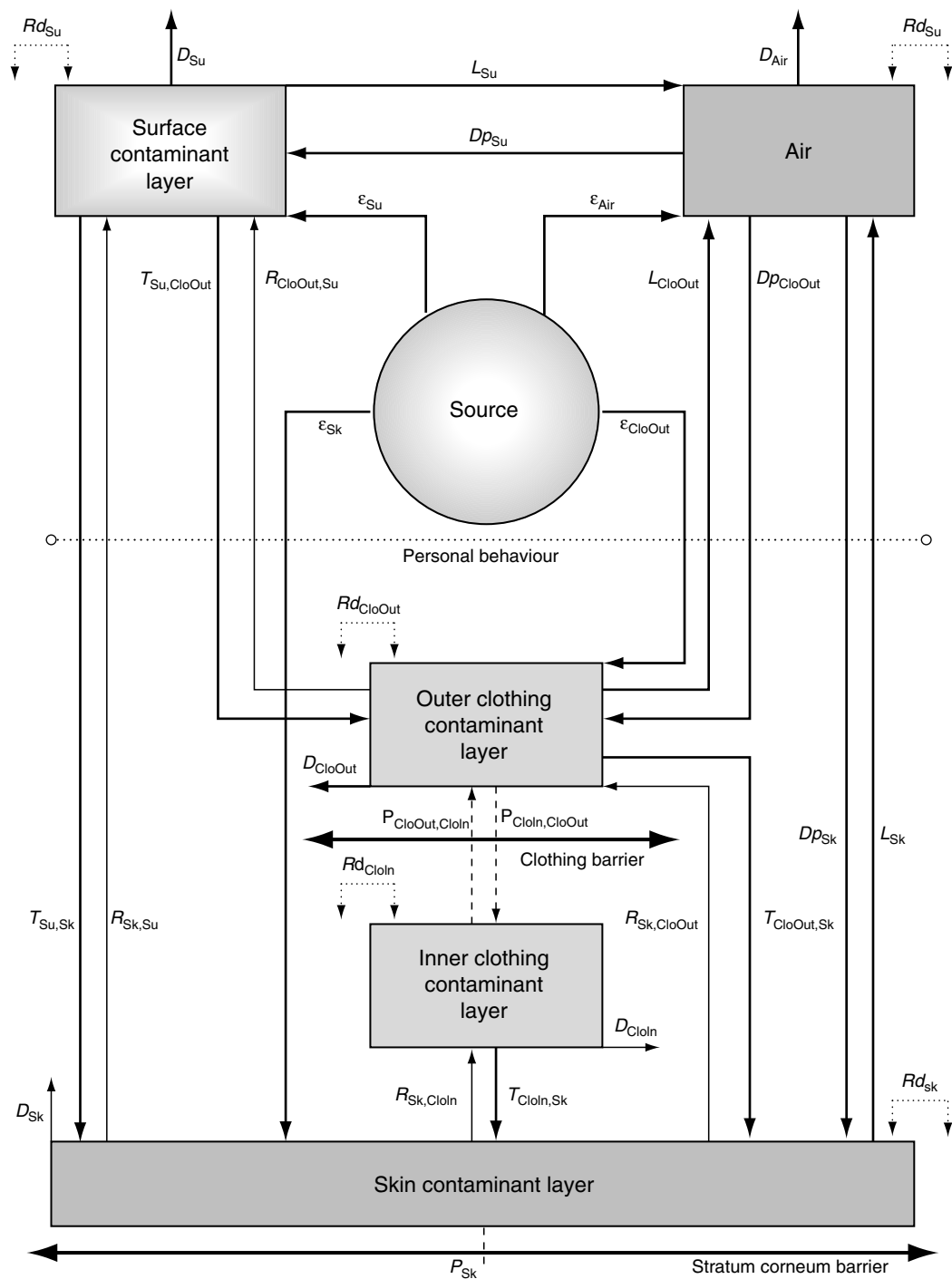


Figure 26.1 The conceptual model of dermal exposure.

- to the outside of the worker's clothing ($\epsilon_{\text{Clo,Out}}$); or
- the worker's unprotected skin (ϵ_{Sk}).

Emission to air may be by evaporation, spraying, grinding or some other activity. Emission to the other compartments can occur by splashing, spilling, spraying or other mechanisms and, although these emissions may pass through the air, they are not resident in the air long enough to be considered part of the air compartment.

Compartments are defined by a number of characteristics: mass and concentration of contaminant, the area of the compartment covered by the material or, in the case of the air compartment, its volume. The contaminant may 'flow' in or out of each compartment, giving rise to increases or decreases in mass therein. For example, material may be *lost* (L) from the surface contamination layer to the air, to the outer clothing layer or the skin contamination layer. In addition, there may be *deposition* (D_p) from the air to surfaces or *transfer* (T) from the surface contamination layer to the skin contamination layer or the outer clothing layer.

Transfer or removal from surfaces to the skin contamination layer may be mediated by contact between the hands or some other body part and the surface. The direction of the contaminant transfer will depend on whether there is more available contamination on the skin or the surface, the wetness of the hand, the properties of the surface and many other factors.

In this scheme clothing, including gloves, is recognized as an important protective measure for people handling hazardous substances. It is described by two compartments with a barrier layer that restricts the passage of substances to the skin. A substance can either permeate from the outer clothing layer (a diffusion process) or penetrate (bulk flow) to the inner layer, or bypass the clothing and be transferred directly to the skin contamination layer. This transfer from the outer clothing layer, or the reverse, could possibly occur as the worker's hands touch the outer surfaces of clothing or via some other activity.

In each compartment there are two other processes that may operate: *decontamination* (D) and *redistribution* (Rd). Decontamination, for ex-

ample by washing one's hands, results in loss of some of the contaminant from the system. Redistribution is the process of modifying the pattern of contamination within a compartment, for example by transferring some contaminant from one area of the body to another, perhaps by touching the face with dirty hands. Redistribution does not change the mass in a compartment but may alter the area covered or concentration.

One important purpose of the conceptual model is to help identify which compartments and mass transfer processes are likely to be important in a specific work situation. Once these are identified then we can see what contextual descriptive information should be obtained to properly describe the exposure. For example, if we plan to measure dermal exposure of a liquid in a situation where the surface contamination layer is judged to be important then we should at the same time collect contextual information about, for example:

- the likelihood of evaporation from the surfaces (L_{Su});
- the frequency and method of cleaning the surfaces (D_{Su});
- the number of times the worker contacts a contaminated surface with their hands ($T_{\text{Su,Sk}}$); and
- how the hand contacts the surfaces, e.g. fingertips, palm etc. ($T_{\text{Su,Sk}}$).

Vermeulen *et al.* (2000) highlight the potential value of the conceptual model from their studies of exposure in the rubber industry. They had obtained measurements of inhalation and dermal exposure, plus the data on the contamination of surfaces that workers were likely to contact. They calculated the correlation coefficients between the various compartment measures to identify which were most closely related to the skin contamination. It was clear that for some jobs (e.g. mixing) the air compartment was most important in determining dermal exposure, whereas for others (e.g. curing) it was the surface contamination layer that was most closely associated with dermal exposure. They also showed that there was often a correlation between the air compartment and surface contamination layer, suggesting that reducing emissions to the air might also result in less surface contamination and hence less dermal exposure as well as lower inhalation exposure. Understanding

which route or routes of exposure are important can help target control measures in the most appropriate way.

Measuring exposure

The available methods for measuring dermal exposure and or surface contamination all have advantages and disadvantages; none provides information on all of the key factors that fully describe exposure. There is very little standardization of methods for agents, other than pesticides, for which much of the measurement of exposure is for regulatory risk assessments. The available methods can be grouped into four broad categories, as shown in Table 26.1.

Patch sampling is widely used to estimate exposure to pesticides and other low-volatility liquids. They have mostly been used to measure pesticides,

although they have also been used for polycyclic aromatic hydrocarbons, dusts and using activated carbon cloth to measure volatile organic liquids. The patches generally comprise a 10-cm square of cotton or some other absorbent cloth with an impervious backing. However, as they are not available commercially there is no standardization in their construction. To obtain a sample the patches are affixed to various body locations before exposure, e.g. chest, forearms, etc., and they are then removed and analysed for the mass of contaminant after exposure. Figure 26.2 shows a typical patch sampling arrangement. Patches are normally attached to the outside of any clothing worn to assess *potential exposure*, although one or two inner patches are also often included. Samples collected from inside clothing are often said to assess *actual dermal exposure*. The information from the analysis of the patches may then be used to estimate the mass on the whole of the body by multiplying the

Table 26.1 Methods for measuring dermal exposure.

Method type	Description	Advantages	Disadvantages	Measurement of the key parameters				Reference
				M_{Sk}	A_{Sk}	t	C_{Sk}	
Patches	Small square cotton or other cloth patches attached to the body	Standard method for pesticides	Only low-volatility substances, only small proportion of body sampled	✗	✗	✓	✗	Soutar <i>et al.</i> (2000)
Suit sampling and gloves	Workers wear lightweight cotton overalls with hood	Whole body sample	Only low-volatility substances, practical difficulties in sampling and analysis	✗	✗	✓	✗	Soutar <i>et al.</i> (2000)
Washing or wiping	Defined areas of skin washed with a solvent or wiped with a moist cloth	Low cost, easy to use	Only for low-volatility substances that do not penetrate through the skin quickly	✓	✗	✓	✗	Brouwer <i>et al.</i> (2000)
Fluorescence	Fluorescent compound added to the source and then the intensity of fluorescence on the worker's body measured	Accurate assessment of area exposed, whole body sample	Requires specialist equipment, must add fluorescent agent to source	✓	✓	✓	✗	Cherrie <i>et al.</i> (2000)

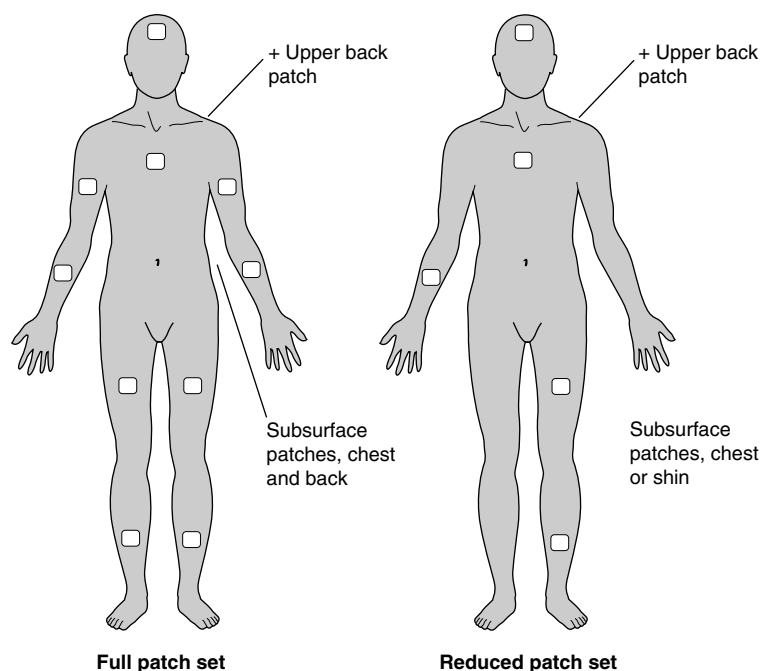


Figure 26.2 Typical patch sampling arrangements.

area of the body part corresponding with each patch (Table 26.2) with the mass of contaminant on that patch and summing overall.

In principle, absorbent patches should collect and retain all of the contaminant that lands on

Table 26.2 Area of various body parts for dermal exposure assessment – 50th percentile male, height 175 cm, weight 78 kg.

Body part	Area (cm ²)
Head	1075
Neck	230
Back	1536
Thighs	3456
Calves	2592
Feet	1229
Shoulder	1306
Chest	1536
Hips	1747
Hands	1075
Forearms	1286
Upper arms	1862
Whole body surface	19200

From HSE (1999).

them. However, this is not the same as the mass in the skin contaminant layer because liquid that spills or splashes onto the skin may run off. Patches cannot be used to obtain information about the concentration of the contaminant in the skin contamination layer or about the area of skin exposed, other than as a crude indicator from the results of the analysis of separate patches. Care must be taken to ensure that patches do not become overloaded or saturated, which may be a particular problem if activated charcoal cloth is used for volatile agents.

Progress is being made in developing a biologically relevant dermal sampler along the lines suggested by Cherrie and Robertson (1995). This patch sampler is designed to measure dermal exposure to toluene and other volatile substances and comprises an impervious backing, adsorbent layer and semipermeable membrane. Preliminary tests show that it responds to the concentration of the hazardous substance in contact with it rather than the mass of contaminant (Lindsay *et al.*, 2003).

The obvious advantage of suit sampling over patches is the almost complete coverage of the body. The suits that are used are typically made

from either cotton or cotton and polyester. A hood or hat can be used to collect a sample from the head and cotton gloves may be worn to obtain samples from the hands. After sampling the suit is removed and carefully cut into separate sections (e.g. arms, legs, etc.) for analysis. The much larger area of cloth presents practical problems for the analyst as it requires relatively large volumes of solvent for extraction, which in turn requires greater sensitivity in the analysis. Suit sampling is generally more time-consuming than patch sampling.

Hand washing and wiping provide a more direct measure of the contaminant in the skin contamination layer. Hand washing involves some mechanical scrubbing as well as liquid washing over the skin, whereas hand rinsing only involves liquid to skin contact. A wide range of liquids have been used for washing, including distilled or deionized water, with or without surfactant added, commercial liquid soaps and organic solvents such as ethanol or 2-propanol. Rinsing is often undertaken with a bag of liquid into which the hand is placed. This is sealed at the wrist and the hand is vigorously shaken for a predetermined number of times or for a fixed period. Wiping is also carried out using a variety of materials, from dry cotton fabric to filter paper soaked in ethanol to commercially available moist hand wipes. Most investigators use a flexible template to define the area to be sampled and specify the wiping action, pressure and the number of times that the area is sequentially wiped. Repeated wiping of the surface up to three or four times may increase the contaminant recovery efficiency.

Most washing and wiping protocols recommend collecting samples at times when workers would normally wash their hands, e.g. before meal breaks, at the end of the day and at other appropriate times. It is often assumed that the mass of contaminant collected on sequential samples should be added together; however, this may not be appropriate where the skin or the patch has a maximum capacity to retain material that is similar to the exposure. Combining samples may then give an overestimate of exposure.

Quantitative assessment of dermal exposure using fluorescence has been pioneered by a small number of research groups. These techniques use a

commercially available fluorescent compound such as Tinopal or Calcofluor diluted to about 0.001% by weight in the source of the contamination. After a period of work the subject's skin is photographed under ultraviolet light with a video camera that is linked to a computer analysis system. The gray level of the digitized image is then compared with calibration information to estimate the mass of fluorescent chemical on each image pixel and hence an estimate of the mass of contaminant. However, one must interpret this carefully as the fluorescent tracer is only a surrogate for the contaminant and may over- or underestimate the mass of the contaminant chemical in the skin contamination layer, depending on whether the tracer is more or less volatile than the chemical contaminant and/or adheres more or less strongly to the skin. The fluorescent tracer technique also gives an estimate of the area of skin exposed. Considerable improvements have been made in the ultraviolet illumination systems and software used by these systems to improve the accuracy of the exposure estimates, although the cost of the systems and the complexity of the assessment have meant that fluorescent imaging has remained a specialist research tool.

Quality assurance procedures are an important part of any measurement. For patches and suit samples this should include an evaluation of the recovery efficiency from the sample media, stability of the analyte over time, along with field and laboratory blanks. In the case of washing and wipe samples it is also necessary to determine the efficiency of removal of the contaminant from the skin. Recovery efficiency from patches and for removal from the skin with washing and wiping has sometimes been reported as low and variable, which underlines the importance of such evaluations. There is some suggestion that recovery efficiency is dependent on the loading, with poorer recovery from lower mass loadings.

Modelling exposure

As we have seen there are no measurement methods that can give a complete picture of the important dermal exposure parameters identified

in the conceptual model. One approach to provide some alternative picture would be to model exposure. At present there are a small number of models that can be applied in such circumstances, although this is a rapidly changing area with new models being developed. The best known model is EASE, the Estimation and Assessment of Substance Exposure model was developed by the UK Health and Safety Executive for European regulatory risk assessments. It is a general purpose exposure model implemented as a computerized expert system, which has a separate module for dermal exposure. EASE focuses on exposure to the hand and forearm from direct contact with contaminated surfaces, without any protective gloves being worn. It relies on an assessment of the extent of contacts with surfaces, i.e. contacts per day, as the main determinant of exposure, which is expressed as a mass density per day (mg cm^{-2} per day). Unfortunately, this model does not provide the other parameters needed to estimate the uptake of the contaminant.

However, perhaps a more useful approach is to subjectively estimate the area of skin exposed, the concentration of the substance on the skin, the duration of exposure and the likely mass of the contaminant on the skin. This can be done for specific tasks to make the evaluation easier. For example, workers involved in dipping (i.e. totally immersing) sheep in dilute organophosphate pesticides may undertake four tasks: handling dry sheep; making up the dilute pesticide dip; dipping sheep; and handling wet sheep post dipping. It can be assumed that there is no exposure to people involved in dry handling. However, those involved in dipping become thoroughly wet with dilute dip, with the lower body wettest. It would not be unreasonable to assume that all of feet and hands, 80% of legs, 40% of arms and upper body and 20% of head were exposed (about $12\,400\text{ cm}^2$).

Exposure to the concentrated pesticide occurs from occasional splashes and by contact with residues on the container. It might be that about 50 cm^2 of the hands would be contaminated in this way. Wet helpers might have their hands, part of the arms and part of the legs exposed (about 2500 cm^2). The pesticide is diluted with 500 parts of water to form the dip; this task takes place at regular intervals throughout the dipping (6 h) and it has been assumed that it will result in continual exposure. In this scenario the wet helper only works for 3 h. The estimated exposure parameters for this scenario are shown in Table 26.3. By assuming that the pesticide forms a layer on the skin of about $5\text{ }\mu\text{m}$ thick, the mass on the skin can be estimated. The uptake is calculated assuming a flux of the concentrated pesticide through the skin of $5\text{ }\mu\text{g cm}^{-2}\text{ h}$.

It can be seen that only about 5% of the mass of pesticide on the skin would be taken up into the body. Although the dipping task results in the most widespread contamination, it is the concentrate handling that contributes the greatest exposure and uptake. Those involved with wet handling have exposure that is about one-tenth of the dipping task.

Evaluating risks

A hierarchical step approach is recommended when investigating situations where dermal exposure may be important. Initially, the potential for exposure should be identified from details of the chemicals used and the available toxicity information to assess whether the substance may pass through the skin or has the potential to cause skin disease. At the end of the evaluation if none of the chemicals presents a risk from dermal uptake then this route may be eliminated.

Table 26.3 Example of organophosphate exposure from sheep dipping.

Job	$C_{sk}(\text{mg ml}^{-1})$	$A_{sk}(\text{cm}^2)$	$t_{sk}(\text{h})$	$M_{sk}(\text{mg})$	$E_{sk}(\text{cm}^2\text{ h}^{-1})$	$U_{sk}(\text{mg})$
Dipping	2	12 400	6	12.4	149	0.7
Handling concentrate	1000	50	6	25	300	1.5
Wet helper	2	2500	3	2.5	15	0.1

Part of this evaluation may involve scrutiny of whether any of the substances being used has been assigned a skin notation by the appropriate national regulatory authority. However, caution should be exercised in interpreting such information as the assignment is often not made on any systematic basis so the meaning of the skin notation is unclear. In the UK, the criteria for assigning a skin notation is now that exposure should make a substantial contribution to body burden and that arriving at a conclusion about risk based solely on an inhalation exposure assessment might be misleading, although not all earlier designations may fully conform to this definition. For example, Kežić *et al.* (2001) provide data on dermal uptake from brief exposure to five liquid solvents, of which three had a skin notation. The uptake for the two solvents that did not have a skin notation was similar to one of the three that did, whereas the other two had uptake of about 5–10 times higher.

The second stage of an investigation should include detailed observations of the work activity to assess the factors that will determine exposure and to evaluate which environmental compartments and transfer routes in the conceptual model are likely to be important. For exposure, the focus should be on the duration of exposure, the area of skin exposed, the concentration of the relevant substances in the products being handled (as a surrogate for the concentration on the skin) and the mass of the substance that may be in the skin contaminant layer. It is advisable to overestimate rather than underestimate these parameters so that decisions about management of any risks will be protective. Using the approaches described above it may be appropriate to estimate the dermal uptake, which for systemic risks may then be compared with inhalation uptake. It is important to remember that for some substances and some exposure circumstances dermal uptake from vapour will be important.

The third phase of an investigation will be necessary if appropriate risk management decisions cannot be made on the basis of estimated uptake and exposure measurements. The most appropriate practical method should be selected from those available, with the objective of refining the second

stage assessment. The considerations necessary for devising an appropriate sampling strategy are similar to those for inhalation exposure and, in fact, the within- and between-worker components of dermal exposure variation are remarkably similar to those found from inhalation exposure measurements. Workers are often selected for monitoring either in some stratified random scheme or by focusing on ‘worst case’ situations. In either case there are some advantages in deciding to have repeat measurements on the same workers on different days because the between-worker variance component often dominates. In planning any dermal exposure monitoring, consideration should be given to increasing the number of people sampled at the expense of fewer sampling locations on each person. If the prime source of variability arises from inter-individual differences in the work rather than spatial variations in exposure over the body then this will be a more efficient strategy. These decisions can be based on the results from a small pilot study. It seems likely that dermal exposure measurements and surface contamination levels are log-normally distributed.

Some authors have suggested that just as we have exposure limits for inhalation exposure we should have corresponding limits for dermal exposure, i.e. DOELs, although no national regulatory authority has introduced such limits. This approach ignores the fact that for systemic toxins it is the combined exposure from all routes that will determine the actual risk; high inhalation exposure with low dermal exposure may present the same risk as low inhalation exposure and high dermal exposure. Placing a limit on both routes would be unnecessarily restrictive as each would have to be set sufficiently low to eliminate risks from both routes. There is a valid argument for DEOLs to protect against local skin toxicity such as contact dermatitis, although in general there is insufficient scientific information available to support such limits.

Dermal exposure assessment forms an important component of risk assessment for most hazardous substances. There are many situations where inhalation exposure may be adequately controlled, whereas dermal exposure is not. Without careful

consideration of the dermal exposure and appropriate risk management in such situations some people will be exposed to unacceptable risks.

References

- Brooke, I., Cocker, J., Delic, J.I., Payne, M., Jones, K., Gregg, N.C. and Dyne, D. (1998). Dermal uptake of solvents from the volatile phase: an experimental study in humans. *Annals of Occupational Hygiene*, **42**, 531–40.
- Brouwer, D. H., Boeniger, M.F. and van Hemmen, J. (2000). Hand wash and manual skin wipes. *Annals of Occupational Hygiene*, **44**, 501–10.
- Cherrie, J.W. and Robertson, A. (1995). Biologically relevant assessment of dermal exposure. *Annals of Occupational Hygiene*, **39**, 387–92.
- Cherrie, J.W., Brouwer, D.H., Roff, M., Vermeulen, R. and Kromhout, H. (2000). Use of qualitative and quantitative fluorescence techniques to assess dermal exposure. *Annals of Occupational Hygiene*, **44**, 519–22.
- HSE (1999). *Dermal Exposure to Non-agricultural Pesticides. Exposure Assessment Document*. (EH74/3). HSE Books, Sudbury.
- Kežić, S., Monster, A.C., van de Gevel, I.A., Kruse, J. and Verberk, M.M. (2001). Dermal absorption of neat liquid solvents in brief exposures in volunteers. *American Industrial Hygiene Association Journal*, **62**, 12–18.
- LaNier, M.E. (1984). *Threshold Limit Values – Discussion and Thirty-five Year Index with Recommendations*. ACGIH, Cincinnati, OH.
- Lindsay, F., Cherrie, J.W. and Robertson, A. (2003). Development of a method to assess biologically relevant dermal exposure. HSE Research Report RR117. Available at <http://www.hse.gov.uk/research/rrhtm/rr117.htm>
- McKone, T.E. and Howd, R.A. (1992). Estimating dermal uptake of non-ionic organic chemicals from water and soil: I. Unified fugacity-based models for risk assessment. *Risk Analysis*, **12**, 543–57.
- Schneider, T., Vermeulen, R., Brouwer, D., Cherrie, J.W., Kromhout, H. and Fough, C.L. (1999) Conceptual model for the assessment of dermal exposure. *Occupational and Environmental Medicine*, **56**, 765–73.
- Soutar, A., Semple, S., Aitken, R.J. and Robertson, A. (2000). Use of patches and whole body sampling for the assessment of dermal exposure. *Annals of Occupational Hygiene*, **44**, 511–18.
- Vermeulen, R., Heideman, J., Bos, R.P. and Kromhout, H. (2000). Identification of dermal exposure pathways in the rubber manufacturing industry. *Annals of Occupational Hygiene*, **44**, 533–42.

Part 5

Allied and emerging issues

Chapter 27

Occupational accident prevention

Richard T. Booth and Anthony J. Boyle

Introduction	Establishing objectives
Chapter objectives	Defining system boundaries
Chapter content	Identifying continuing hazards
Evolution of safety management principles and law	Identifying failure hazards
Legislative controls of safety	Identifying consequences
Workplace safety management	Assessing risk
Amalgamation of the pathways	Deciding whether the risk is tolerable or broadly acceptable
Reactive safety management	Hazard and operability studies (HAZOPS)
Unsafe acts and conditions and underlying causal factors	HAZOP table headings
Safety management systems	Failure modes and effects analysis
Measurement of safety performance	Investigating accidents
Proactive and reactive monitoring	Managing errors and violations
Leading and lagging performance indicators	Classification of human errors
Limitations of lagging performance indicators	Error reduction strategies
Key performance indicators	Behavioural safety
Risk assessment	Safety culture
Risk assessment techniques	References
Overall procedure	

Introduction

An accident may be defined simply as an undesired event that leads to harm. Accidents associated with work activities occur in diverse circumstances and settings. A small number of serious accidents involve fires and explosions in chemical plants and transportation disasters. But the vast majority of accidents are associated with hazards that form an everyday part of working life, for example, the potential for contact with moving machinery, falls, cuts from material being handled, slipping on or striking against objects and injuries from hand tools. Road accidents experienced by people at work represent the major source of occupational fatalities. Notwithstanding the diversity of hazards and accident experience, there is now broad agreement that the underlying causal factors of all of these accidents, and the principles of accident prevention – the principles of safety management –

apply equally to the control of all these hazards, and indeed to the control of health hazards at work (Booth, 2003). For these reasons, it is inappropriate to consider the specific controls that might be applied to each hazard, rather attention should be given to the underlying causal factors that impact on most or all kinds of hazardous events, namely undesired events that may lead either to an accident or a ‘near miss’.

Chapter objectives

The objectives of this chapter are to describe the contemporary strategy for accident prevention founded on safety management systems and risk assessment, to outline how the strategy evolved and to describe the limitations of the ‘traditional’, but still current, reactive approach to accident prevention. The chapter covers in detail the elements of safety management systems and the factors that

determine their effectiveness, notably human behaviour and the concepts of safety culture. It further presents a systematic approach to predicting accidents (risk assessment) and outlines the key requirements for investigating accidents.

Chapter content

The chapter addresses the following accident prevention issues:

- the evolution of safety management and law;
- the shortcomings of reactive safety management;
- safety management, or strictly occupational health and safety (OH&S) management, systems;
- measurement of safety performance;
- predicting accidents via risk assessment;
- investigating accidents; and
- managing human errors and violations, including the concept of safety culture and the ‘behavioural safety’ approach.

Evolution of safety management principles and law

The regulation – in its broadest sense – of occupational safety in Britain evolved on two largely independent pathways. The first was the development of *legislative controls of safety*, gradually embracing an increasing range of hazards and of industries and employers. The second pathway was the development of *workplace safety management* (Booth, 2000).

Legislative controls of safety

The distinctive feature of industry and hazard-based safety legislation from the mid-nineteenth century to, and in fact beyond, 1974 was that the regulators identified the hazards and implicitly assessed the risks (largely as a reaction to accident experience) and prescribed ‘blanket’ (one size fits all) control standards. Safety was driven by legislation and a knowledge of the law was the foundation of, as well as the superstructure for, primarily ‘hardware’, workplace controls. These requirements were sometimes perceived as a legal-

istic chore both by employers and employees, leading to a poor safety culture and to violations of safety rules. Despite these criticisms, legislative requirements certainly made a substantial contribution to workplace safety that continues to this day, for example machinery guarding, maintenance of boilers and lifting equipment and supervision of young persons.

Workplace safety management

In contrast, safety management at company level became prominent during and after the Great War, promoted by what later became the Royal Society for the Prevention of Accidents. Here, industrial accident prevention was seen as being achieved by committees, workers’ participation, the employment of safety officers, joint accident investigations and what would now be termed the promotion of a positive safety culture. Companies that adopted this approach with enthusiasm were able thereby to reduce their accident rates significantly (HM Chief Inspector of Factories and Workshops, 1929). The limitations of this approach were the tendency to distance safety management from operational management generally, to adopt a reactive approach to prevention and to see accident causation in very simple terms, as will be discussed in the next section.

Amalgamation of the pathways

The amalgamation of the two pathways, or perhaps ‘traditions’, of safety within a framework of law followed the Robens’ Report (Committee on Safety and Health at Work, 1972). The Committee recognized that much safety law was outdated, and too detailed, prescriptive and legalistic. Robens proposed the wholesale repeal of the then existing law and its replacement with ‘goal-setting’ legislation. The Robens’ philosophy was partly realized by the Health and Safety at Work etc. Act 1974, but only fully implemented in the early 1990s as a result of European Union safety management directives. Whereas the old law drove the safety system, the contemporary doctrine is that the law should *underpin* good practice. Good practice embraces not just the development of safety

management systems (see Safety management systems), but also the promotion of a positive safety culture where everyone involved believes that the hazards are sufficiently serious to merit attention, and that the preventive measures are proportionate, workable and effective (see Managing errors and violations).

The last 30 years have thus seen a paradigm shift of great significance, including the recognition that safety management systems should form an integral part of company management and that systematic risk assessment – the prediction of accidents – should be the essential basis for proactive prevention (see Risk assessment). A continuing challenge is that many people and organizations remain wedded to the discredited reactive approach and its emphasis on uni-causality as will be discussed further now.

Reactive safety management

As has been stated in the previous section, a weakness of accident prevention generally (both in terms of legislative development and company ac-

tivity) was that the process largely involved taking preventative action only after an accident had occurred. The ‘traditional’ approach to accident investigation is shown in Fig. 27.1 (HSC, 1993). Accidents were deemed to be ‘caused’ either by an error by the injured person or by inadequate physical standards, and the cause chosen depended largely on the preconceptions of the investigator. It will be further noted that the choice of cause led directly to the choice of the preventative methods. Most practical accident prevention involved the preparation of a safety rule designed to prevent a recurrence of the unsafe act or a physical safeguard to remedy the unsafe condition, most proximate to the accident. Much of the corpus of traditional UK health and safety legislation contains rules and physical standards derived as has been described.

These rules and safeguards devised in the aftermath of disagreeable accidents may be overzealous (as perceived some years after the accident), and conflict with the needs of both employers and employees to get the job done. Both parties may tacitly conspire to evade the safety rules or to defeat the physical safeguards. Moreover, the measures taken to prevent one specific accident

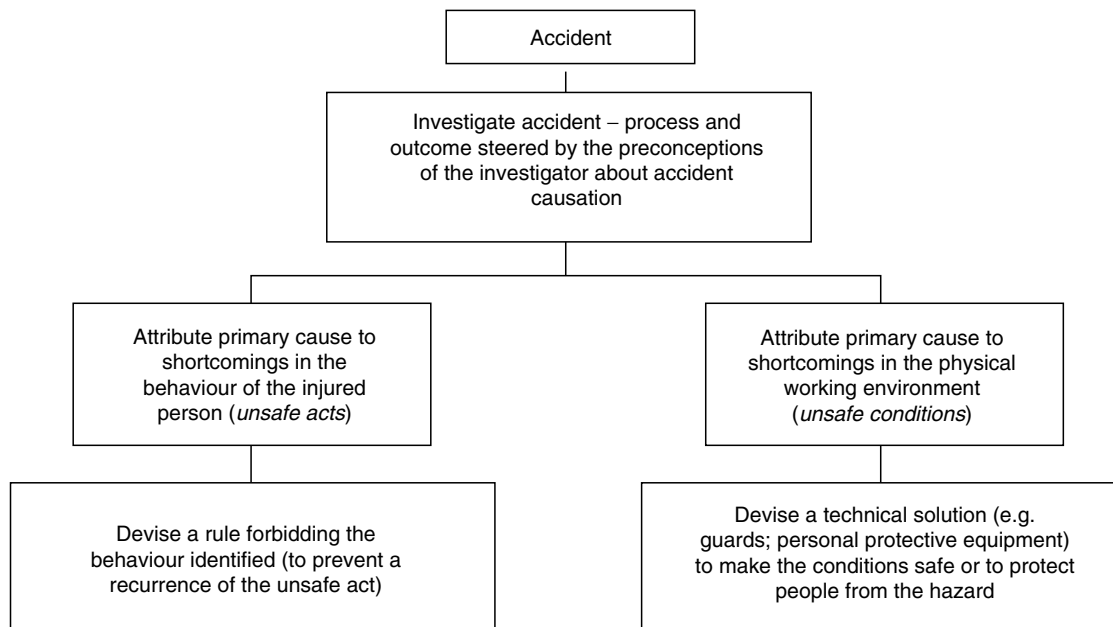


Figure 27.1 Traditional safety management (after HSC, 1993).

may conflict with the measures adopted to prevent a different accident, and with production-oriented rules. In summary, this reactive approach may succeed in preventing *repetitions* of accidents, but only while memories of the disagreeable accident consequences remain fresh in people's minds.

Whatever the achievements of the safety management approaches mentioned in the previous section, they are clearly inadequate to cope with major hazards in rapidly developing technologies. Here, preventive measures may be rendered obsolete by each technical advance and the occurrence of the first accident may itself be intolerable. But the weakness of the approaches is not confined to high-risk and rapidly advancing technologies. Rule books (and legislation) drawn up in this way are likely to become both incomprehensible and contradictory in time. Two company rules may exist for any situation: the 'rule' to get the job done in time and a more demanding rule that may be invoked when things go wrong (HSC, 1993).

Unsafe acts and conditions and underlying causal factors

The relative contribution of unsafe conditions and unsafe acts in accident causation has been a topic of prolonged, heated – and often sterile – debate. Heinrich (1969) in a seminal safety management text first published in 1931 reported that unsafe acts were the 'cause' in 88% of 12 000 accident insurance claims studied: 10% were attributed to unsafe conditions, and 2% were considered unpreventable. In contrast, government inspectors in Britain have argued that unsafe conditions play a much larger part (Department of Employment, 1974). The causation debate, clouded by political overtones and a desire to apportion blame, has often missed three crucial and inter-related issues:

- The concept of a single accident cause is a bizarre simplification of a complex multicausal process. Moreover the term 'unsafe act' embraces a wide range of human failures, including both undesired errors and intentionally risky behaviour, *violations* (see Managing errors and violations).
- The distinction between the contribution of unsafe conditions and unsafe acts in *causation* has

masked the more important distinction between the relative contribution of conditions and behaviour in *prevention*, and the need for prevention plans to promote both safe conditions *and* safe behaviour (Heinrich, 1969).

- The argument has focused almost exclusively on the errors made by the people who have had the accidents, not, for example, the managers and engineers whose fallible decisions (remote in time and place from the location of an accident) may have created a climate and a physical environment where errors by people directly at risk are made more likely or more serious. Unsafe acts create conditions in which further unsafe acts may lead to accidents. The remote errors by managers have been described by Reason (1990) as *latent* or *decision* failures, and the errors by people directly at risk as *active* failures. A latent error might typically be an organizational failure to develop an effective safety management system.

Safety management systems

Figure 27.2 presents a typical model of a safety management system, or strictly an OH&S management system (HSE, 1997).

Policy embraces the preparation of a mission statement that should outline an organization's commitment to OH&S and its overall approach to OH&S management.

Organizing covers issues such as allocation of responsibilities within a defined management structure, and the arrangements for:

- cooperation, so that everyone in the organization participates in the development of the overall system;
- effective communications within and also beyond the organization; and
- ensuring that all personnel are competent to perform their duties.

Planning and implementing has two elements, first the development of safety plans generally, and second the development of risk controls founded on risk assessment.

Measuring performance involves reactive monitoring, primarily the investigation of accidents (and cases of ill health), and proactive monitoring,

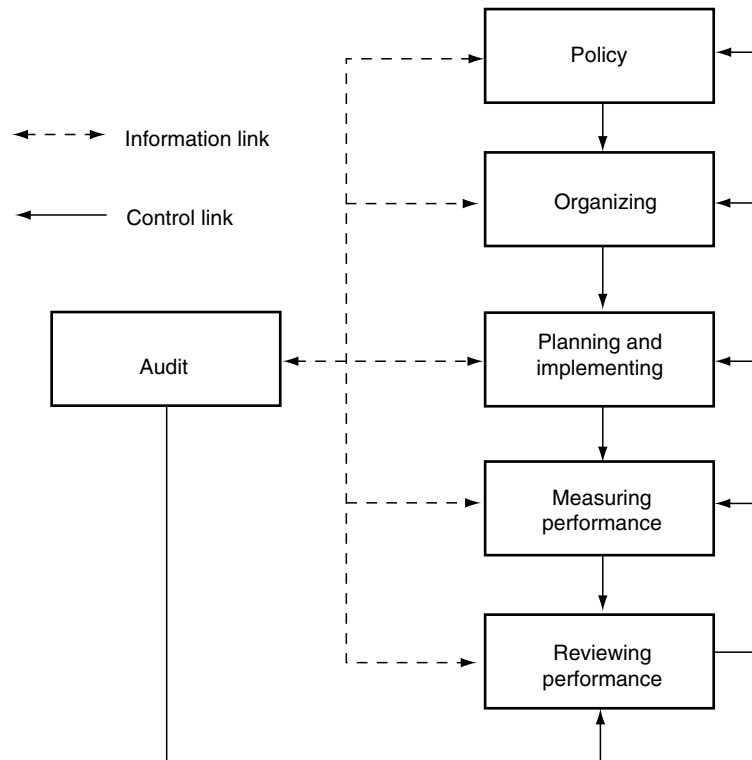


Figure 27.2 Occupational Health and Safety (OH&S) management system (the 'HSG65' model) (after HSE, 1997).

primarily timely checks to determine the level of compliance with risk controls (see Measurement of safety performance). The key distinctions between these two types of monitoring are described in the next section.

Audit embraces proactive measurement of part, or all, the OH&S management system by an independent auditor.

Review is the process by which action is taken to effect improvements in the system on the basis of the findings of performance measurements made within the organization, and the findings of independent audits.

Models such as that shown in Fig. 27.2 provide the generic superstructure within which organizations seek to manage safety. The deficiency of this and other models, for example BSI (1999), is that they do not show directly what is involved in managing safety at the level of the workplace, and the activities that go on there (Boyle, 2002). In particular, they do not show the linkage between risk assessment and control, and proactive and reactive

monitoring in the context of work activities and accident prevention.

Figure 27.3 shows a model that focuses on the tactical aspects of workplace safety management, derived from Boyle (2002). The starting point is a 'work activity', for example a maintenance task on a machine. The first step is to gather the necessary information (including standards and law) about the machine and the task and to examine the organization's compliance performance, *leading* indicators, and data from within the organization (and beyond) about machinery maintenance accident experience, *lagging* performance indicators (see Measurement of safety performance). Next, risk assessment is carried out to determine the likelihood of an accident and the nature and the extent of the (foreseeable) consequences. The risk controls embrace both 'workplace precautions' (e.g. a permit to work) and the ancillary activities 'control systems' that underpin the permit system (e.g. training, supervision and record-keeping). Proactive monitoring checks and

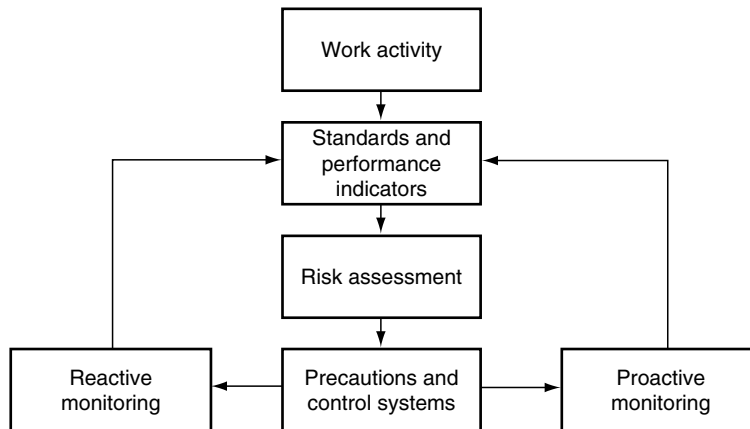


Figure 27.3 Work activity accident prevention model derived from Boyle (2002).

promotes compliance with the precautions and the control systems, and reactive monitoring provides (albeit belated) evidence that the overall risk management system is effective. The leading and lagging data derived from both monitoring methods provide continuing evidence (via the feedback loops) of the success or failure of the overall system to prevent accidents.

Measurement of safety performance

This section is derived from BSI (2004).

Proactive and reactive monitoring

The full meanings of the two terms are:

- *proactive monitoring* consists of timely routine and periodic *checks*:
 - that OH&S plans have been implemented;
 - to determine the level of conformance with OH&S management systems;
 - that seek evidence of harm that has not otherwise come to the attention of the organization via reactive monitoring;
- *reactive monitoring* embraces structured *responses* to OH&S management system failures, including hazardous events and cases of ill health.

In a fully effective OH&S management system, proactive monitoring provides reassurance that the system is operating as intended, for example confirmation that personnel have received relevant

training and that safe systems of work are being followed. In a less than fully effective system, proactive monitoring provides timely evidence of problems that need to be remedied, for example work being carried out without a risk assessment or that all accidents have been reported.

Reactive monitoring is exclusively concerned with belated, but systematic, responses to shortcomings in the OH&S system, and the investigation of hazardous events and the causes of ill health. These problems may have been brought to the attention of the organization, for example by a statutory inspector or by complaints by employees or members of the public or, in the case of hazardous events, by the people who have experienced or witnessed near misses or harm.

An exclusive reliance on reactive monitoring, as was discussed in Reactive safety management above, may lead to complacency, regulatory agency action, complaints by the public that could adversely affect corporate image and, above all, that an organization's OH&S management system is likely to lie dormant until harm occurs.

But it is wrong to assume that full compliance with risk controls, confirmed by proactive monitoring, means that risks are in fact fully controlled. For example, a safe system of work may not cater for all eventualities. Thus, reactive monitoring should be combined with proactive monitoring. Reactive monitoring of hazardous events may reveal that although risk controls were fully imple-

mented, confirmed by proactive monitoring, they were nonetheless ineffective in preventing harm.

Proactive monitoring of OH&S performance in many cases (e.g. when routine checks are carried out by supervisors) will lead to immediate corrective action and the information about the findings may not be formally recorded. But organizations should, where practicable, record the findings of proactive monitoring and *always* document the findings of reactive monitoring.

Leading and lagging performance indicators

The meanings of the two terms are:

- *Leading performance indicators* are data on compliance or non-compliance with the performance requirements of OH&S plans and compliance or non-compliance with the organization's OH&S management system generally.
- *Lagging performance indicators* are exclusively data on the prevalence of hazardous events (accidents and near misses), and of occupational ill health.

The two types of performance indicators derive from both proactive and reactive monitoring, although in practice most leading performance indicators are measured via proactive monitoring, and most lagging indicators are revealed by reactive monitoring.

Limitations of lagging performance indicators

As has been reiterated above, many organizations continue to focus on accident data as the primary measurement criterion. As a reduction in accident rates is the ultimate performance indicator, this emphasis, paradoxically, has a number of disadvantages (BSI, 2004):

- Most organizations have too few accidents to distinguish real trends from random effects.
- If more work is done by the same number of people in the same time, increased workload alone may account for an increase in accident rates.
- The length of absence from work attributed to injury may be influenced by factors other than the

seriousness of injury such as poor morale, monotonous work and poor management–employee relations.

- Accidents are often under-reported (and occasionally over-reported). Levels of reporting can change. They may improve as a result of increased workforce awareness and better reporting and recording systems.
- Accident rates are influenced by the proportion of personnel carrying out 'risky' work. If additional staff are recruited for low-risk work, the accident rate (e.g. per person employed) will go down, although at the 'sharp end' the risk is not reduced. On the other hand, an organization's accident rate might appear to improve when high-risk tasks are taken over by contractors.
- A substantial time delay can occur between safety management failures and the evidence of harmful effects or vice versa.

Key performance indicators

Key performance indicators (KPIs) are an abridged selection of leading and lagging performance indicators that should be used by senior management to review the implementation and effectiveness of plans and workplace risk controls.

Risk assessment

A risk assessment (accident prediction) for a particular location, process or activity involves the following steps (Boyle, 2002):

- Identifying what types of accident could happen, commonly referred to as 'hazard identification';
- For each type of hazard identified, estimating its associated 'risk', namely:
 - the likely frequency or probability of occurrence; and
 - the likely mean harm (in terms of injury severity and its associated financial cost);
- For each type of hazard identified, deciding whether the combination of frequency and harm (the risk) is tolerable to the organization;
- When the risk associated with the hazard is not considered tolerable, deciding what measures will be taken to reduce the frequency, or harm, or both.

This decision should involve cost–benefit analysis, drawing on information about the financial outlay associated with the measures proposed and the extent to which the measures will reduce risk.

It will be noted that this process is, in effect, making a business case for accident reduction. This emphasis on making a business case is not common but it renders the process rational and transparent and could be a significant motivational factor for accident prevention activity. However, employers may be obliged to spend more on prevention than is justified by a business case founded only on the local ‘costs’ of accidents: many of the costs are borne not by employers but by society at large. The business case should in any event be supported by evidence that the most cost-effective solutions have been chosen.

Risk assessment techniques

The two techniques that will be dealt with, as examples of available methodologies, are as follows:

- Hazard and Operability studies (HAZOPS);
- Failure Modes and Effects Analysis (FMEA).

The techniques have to be used as part of an overall procedure and, as this procedure is common to both techniques, it will be described before their detailed description.

Overall procedure

There is an overall procedure for risk assessments and the seven main steps in this procedure are listed below, followed by a more detailed description of each step:

- establish the objectives for the risk assessment;
- define the system boundaries;
- identify continuing hazards;
- identify failure hazards;
- identify consequences;
- assess risks;
- decide if the risks identified are tolerable (when further risk reductions would not be justified by the costs involved) or broadly acceptable (where no further reduction in risk is necessary).

The procedure is iterative, with findings in later stages often making it necessary to go back and revise an earlier stage.

Establishing objectives

The risk assessment procedure should start with a clear statement of what is to be achieved, namely the objectives of the risk assessment. In addition, risk assessment can generate large amounts of written information and the whole process can become bogged down in trivia unless there are clear objectives to keep things on course.

Defining system boundaries

Defining the system boundaries sets the scope of the risk assessment project and definitions are usually recorded in the form of sketch maps, block diagrams and flow charts, each annotated as appropriate. As with the objectives, everything has to be written down and agreed to avoid confusion later. The inputs to the system will also have to be considered, as will the outputs, and both will have to be agreed and recorded.

Identifying continuing hazards

Continuing hazards are those that are inherent in the process itself, the materials being used or the layout of the plant and equipment. These hazards are there as part of the design and will be present even if everything is operating as it should.

Continuing hazards can be ‘designed out’, which is why many risk assessments are carried out at the design stage. However, removing continuing hazards from existing processes can be difficult. (In practice, most continuing hazards relate to health effects, for example high noise levels.)

Identifying failure hazards

Failure hazards are those that arise only when the system fails (as a result of technical failures and/or human errors) and the techniques considered later in this section are primarily concerned with this type of hazard.

Identifying consequences

Although there is a potentially wide range of possible consequences of failures, the risk assessment

techniques are normally restricted to those having significant risks of injury. This is primarily done to limit the scope of the assessment as a major problem with all of these techniques is a tendency to get immersed in trivial risks.

Assessing risk

Various risk categories are used for qualitative or quantitative assessment of risk but, as was mentioned earlier, the preferred approach is to estimate the likely harm in terms of injury severity and their associated costs.

Deciding whether the risk is tolerable or broadly acceptable

This will usually be a subjective judgement arrived at by consensus among the risk assessment team members. However, when risks are quantified, a comparison may be made with standards based on risk levels, in this context defined as the probability of a fatality per year (HSE, 1988, 2001).

When it is decided that a risk is tolerable or broadly acceptable, no further action is usually taken, other than to ensure that control measures are sustained. When the risk is deemed as unacceptable, the team may go on to consider what risk controls would be appropriate.

Having looked at the overall procedure for the use of risk assessment techniques, it is now possible to consider in more detail each of the techniques mentioned above. They would normally be used in their full form only when the consequences of an accident were very severe, but the principles can be applied to any type of accident prediction.

Hazard and operability studies (HAZOPS)

HAZOPS are primarily used in the design stage to identify hazards that could occur if the process or operation did not go as planned, that is the hazards arising from failures or malfunctions in the system. However HAZOPS can be used for existing systems.

HAZOPS is a qualitative procedure that systematically examines a process by asking questions

about what could go wrong. It is generally carried out by a small team of people with knowledge of the system, directed by a group leader experienced in HAZOPS. Essentially HAZOPS is a brainstorming exercise and can be very time-consuming if the focus is not kept on significant risks.

The questions asked by the HAZOPS team are generated by two sets of key words: property words and guide words.

- *Property words*. These are words chosen to focus attention on how the process operates, for example for a steam boiler, the words would include temperature, pressure and (water) level.
- *Guide words*. These are words chosen to focus attention on possible deviations from the design intention.

Property words have to be chosen to suit a particular system, as do guide words. However, the guide words listed below have general relevance, although others may have to be added for particular risk assessment exercises.

<i>No or not</i>	The complete negation of the design intention
<i>More</i>	Quantitative increase, e.g. higher temperature
<i>Less</i>	Quantitative decrease, e.g. lower temperature
<i>As well as</i>	Qualitative increase, e.g. an impurity
<i>Part of</i>	Qualitative decrease, e.g. component in mixture missing
<i>Reverse</i>	The logical opposite of the intention, e.g. liquid flowing in opposite direction to that intended
<i>Other than</i>	Complete substitution, e.g. wrong material

Once the property words and guide words appropriate to the system have been established, the team works through each combination of guide and property words, brainstorming to decide whether a deviation from the design intention could arise. Then they consider possible causes of this particular failure and the consequences if the failure occurred. The results of the exercise are usually recorded in a preprinted table with the headings listed below.

HAZOP table headings

The headings are:

- property word;
- guide word;
- deviation;
- possible causes;
- consequences;
- action required.

The 'action required' column is used to record either risk control measures that will prevent the failure or requirements for further information.

Failure modes and effects analysis

FMEA is used to identify the hazards resulting from failures in hardware. It starts by listing the hardware items and analysing their possible failure modes, it is therefore a 'bottom up' approach. However, FMEA can also be extended to include numerical methods and can thus cover preliminary costing as well as hazard identification.

FMEA starts at the component level and seeks to answer the following questions about each component.

- How can this component fail, that is, what are the failure modes?
- What could the effects be if the failure occurred?
- How could the failure mode be detected?
- What would be the risk associated with the effects?

The results of a FMEA are recorded in a table that has columns for details of the component being analysed and answers to the four questions listed above.

Investigating accidents

Adequate accident investigation requires the following:

- A clear idea of what is to be achieved by the investigation. For example it is often stated that the primary aim of an accident investigation should be to prevent recurrence. While this might have been true in the days before risk assessment, it is no longer the case. If there is a primary purpose

for accident investigation then it is to review the risk assessment results.

- A high level of competence in interviewing and observation. People tend to believe that because they can hold a conversation they can carry out an interview. This is not the case.
- High levels of analytical skills, particularly those required to analyse often conflicting views of how and why an accident happened.
- A high level of creativity in generating possible remedial measures.
- A detailed knowledge of human factors and, in particular, individual differences and human reliability.

Investigators should consider whether the hazardous event was associated with one or more of the following, a list derived from BSI (2004). It will be noted that the process now outlined mainly involves an analysis of shortcomings that could occur within each element of the work activity accident prevention model shown in Fig. 27.3. Thus the investigation should challenge each element of the workplace preventive arrangements:

- Was there an inadequate collection of 'standards' and leading and lagging performance indicators to provide a secure foundation for risk assessment?
- Were workplace precautions and control systems selected on the basis of an unsuitable or insufficient risk assessment?
- Was there poor implementation of workplace precautions and control systems?
- Were there failures of proactive monitoring to detect poor implementation of controls?
- Were controls implemented but ineffective?
- Were there failures of reactive monitoring, for example a failure to detect near misses, which would have revealed ineffective controls?
- Were controls not reviewed or improved in the light of evidence of proactive and/or reactive monitoring?

In each case, the investigator should dig deeply to establish the root causes for any deficiencies revealed. For example, non-compliance with workplace precautions and control systems may result from production pressures, promoted by the tacit support of supervisors.

Details of structured methods for accident investigation, notably an Events and Causal Factors Analysis, are presented by Boyle (2002).

The findings of accident investigations should be communicated to all relevant parties. The findings should lead not just to improvements in compliance with relevant aspects of the process model shown in Fig. 27.3, but also to improvements in the generic OH&S management model shown in Fig. 27.2, and not least to an evaluation of the organization's safety culture and the way the organization is managed more generally.

Managing errors and violations

Errors and violations differ in that the latter involve intent, whereas the former do not. This fundamental difference makes it necessary to deal with errors and violations as separate topics. This section on managing human errors and violations begins with the classification of human errors and the strategies for error reduction. There is then a discussion of behavioural safety and safety culture, which are more appropriate ways of managing violations.

Classification of human errors

A typical error classification is the one devised by Miller and Swain (1987) and reproduced in Table 27.1.

A particularly influential classification of errors is that produced by Rasmussen (1980) and extended by Reason (1990). The human failure

Table 27.1 Classification of errors (after Miller and Swain, 1987).

<i>Error type</i>	<i>Description</i>
Commission	Adding or including something that should not be there
Omission	Missing something out
Selection	Incorrect choice from a range of options
Sequence	Incorrect serial position of actions or events
Time	Too late or too early with an action
Qualitative	Not performing an action properly

types identified by Reason are illustrated in Fig. 27.4 and explanatory notes are given below.

Slips are attentional failures, *lapses* are memory failures. Unintended actions that arise from slips or lapses usually occur at the routine execution stage of a task when the human is on 'automatic pilot'. Therefore, they are relatively easy to detect by observation. As they are relatively easy to detect, it is also relatively easy to provide suitable remedial or compensatory actions.

Mistakes embrace 'rule-based' mistakes: the misapplication of a good rule, or the application of a bad rule, and 'knowledge-based' mistakes where the human is confronted with a wholly novel situation. Intended actions that are mistaken usually occur at the planning stage of a task and the consequences of such mistakes may not materialize until some time after the mistake has been made. In the design of buildings, mistakes that create risks during the construction or maintenance of the building would fall into this category, as would design errors that make machinery difficult to operate without risk. Because these errors remain in the system until circumstances occur that trigger their effects, they are often referred to as 'latent' errors (see Reactive safety management). Many of the inappropriate decisions by management on safety matters fall into this category of latent errors.

Violations are described by Reason as deliberate – but not necessarily reprehensible – deviations from those practices deemed necessary (by designers, managers and regulatory agencies) to maintain safe operation. They can be divided into *routine* violations, when people habitually break the rules, for example the failure to wear personal protection, and *situational* violations, when people break the rules because the rules are difficult, if not impossible, to comply with. It should be remembered that there are cases when safety rules are so complex and inappropriate that systems can only operate if there are frequent and regular rule violations condoned by management. In the UK, 'working to rule', that is accurately following all safety rules, has been used by workforces as an alternative to strikes as a means of putting pressure on management during pay negotiations. When this is possible, it is legitimate to doubt the appropriateness of the rule base being used.

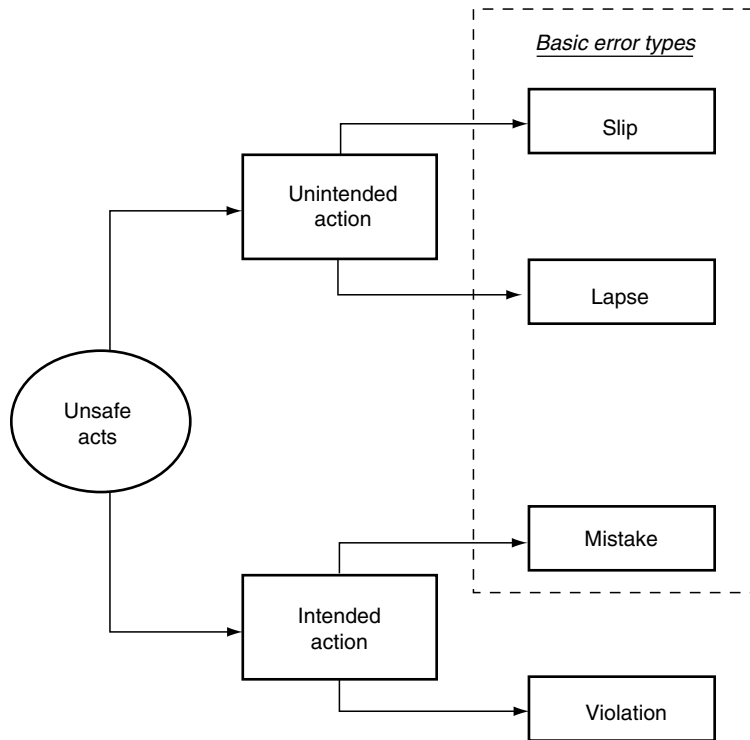


Figure 27.4 Human failure types (after Reason, 1990).

Error reduction strategies

Error reduction strategies are intended to reduce the likelihood of human errors and the most effective methods involve design, or redesign, of machines, equipment and tasks so that the possibility of error is 'designed out'. When errors cannot be designed out, it is essential to identify accurately the sorts of errors that can occur, as different types of error require a different approach, or combination of approaches, to prevention.

Various authors have attempted to establish the most effective methods of reducing the different types of error and Table 27.2 illustrates one such author's results (Mason, 1992).

Behavioural safety

Behavioural safety begins by identifying what it is that people are doing, or not doing, which could contribute to an accident (Boyle, 2002). Simple examples include the following:

Table 27.2 Error reduction strategies (after Mason, 1992).

<i>Error type</i>	<i>Error reduction methods</i>
Slips and lapses	Design improvement, training
Potential for mistakes	Training (team and individual), over-learning, refreshers; duplication of information, clear labelling, colour coding
Knowledge-based errors	Hazard awareness programmes, supervision, work plan checks, post-training testing
Violations	Motivation, underestimation of risk, balance between perceptions of risks and benefits, supervision, group norms, management commitment

- *doing* – riding on the forks of a lift truck, standing on swivel chairs, sitting with a poor posture and lifting heavy weights unassisted;
- *not doing* – not wearing personal protective equipment, not holding the hand rail when walking down stairs and not checking tyre pressures.

When the behaviours have been identified, any effective and reasonable means are used to change the behaviour. What will be effective in changing the behaviour, and what constitutes reasonable means, will vary according to the circumstances, but methods that have been used include the following:

- *Continuous supervision.* People tend to do what they are supposed to do when they know they are being supervised. This method is strengthened if it is combined with either, or both of the next two methods.
- *Sanctions for non-conformance.* Sanctions range from (non-judgemental) verbal reprimand to forms of disciplinary action, to dismissal. The last has the effect of ‘encouraging the others’ as it demonstrates the management’s intentions very clearly.
- *Rewards for conformance.* These can vary from verbal commendation to financial rewards, to early promotion.

If, by whatever means, the required behaviour can be enforced for a sufficiently long period of time, the behaviour becomes internalized, that is the person now considers that the behaviour is natural and normal and behaves in the desired way from choice. When this has happened, the behaviour is self-sustaining. An example was the imposition of seat belt wearing in the UK. When cars in the UK were first fitted with seat belts, few people wore them. However, when it became an offence not to wear a seat belt, people wore them to avoid prosecution. Having worn seat belts for this reason for a number of years, people now feel ‘uncomfortable’ without a seat belt and most people would continue to wear a seat belt even without the legal obligation.

Safety culture

A positive safety culture within an organization is now recognized as a crucial determinant of an

excellent safety performance (see Evolution of safety management principles and law).

Safety culture is a complex area and the aim of the present section is to provide a brief introduction to the following topics:

- what is meant by safety culture;
- how safety culture can be measured;
- how safety culture can be influenced.

Safety culture is normally defined (HSC, 1993; HSE, 1997) as follows:

The safety culture of an organization is the product of individual and group values, attitudes, competencies and patterns of behaviour that determine the commitment to, and the style and proficiency of, an organization’s health and safety programmes.

Organizations with a positive safety culture are characterized by communications founded on mutual trust, by shared perceptions of the importance of safety and by confidence in the efficacy of preventive measures.

For the present purposes it is only necessary to consider what people think, say and do in the context of the organization in which they work:

- | | |
|--------------|---|
| <i>Think</i> | This includes values, attitudes, beliefs, motivation, etc. |
| <i>Say</i> | This includes stated intent, verbal behaviour such as body language and written statements. |
| <i>Do</i> | This includes all other behaviour and physical responses. |

It is necessary to split ‘think’ and ‘say’ in this way because it is possible for people to think one thing and say another, but for the present purposes we will concentrate on attitudes (think) and behaviour (do).

Unfortunately, attitudes and behaviour are not always appropriately linked and the first step in dealing with safety culture is to identify that there is a mismatch between attitude and behaviour, as illustrated in Table 27.3.

There can be a variety of reasons why there is such a mismatch. For example, people may wear eye protection because it is an enforced rule rather than because they believe their eyes require protection, and people may not wear eye protection, despite thinking it is a good idea, because of peer

Table 27.3 Matching and mismatching attitudes and behaviour.

	<i>Positive behaviour</i>	<i>Negative behaviour</i>
Positive thought (attitude)	Wearing eye protection is a good idea; I wear my eye protection	Wearing eye protection is a good idea; I do not wear my eye protection
Negative thought (attitude)	Wearing eye protection is not a good idea; I wear my eye protection	Wearing eye protection is not a good idea; I do not wear my eye protection

group pressure. Although the wearing of eye protection is used as an illustration, the problem can apply to all safety activities, including those normally associated with managers. These include chairing safety meetings, planning for safety and organizing funding.

Note that behavioural safety and safety culture are different approaches to achieving the same aim, namely to have people with positive attitudes and positive behaviour:

- The behavioural safety approach involves changing people's behaviour in the hope that their attitude will also change.
- The safety culture approach involves changing people's attitudes in the hope that their behaviour will also change.

In practice, it is best to use a combination of both approaches.

One way to measure safety culture is by using an extensive attitude survey administered to all employees. However, there are less structured ways of measuring safety culture that can be used at any time:

- *Ask people about their attitudes.* Managers could identify a few important issues, draw up a short list of questions and ask a small sample of people to answer these questions.
- *Observe what people do and ask them why they do it.* In particular, managers could talk to people who are doing the correct things and find out what motivates them to do things correctly.
- *Monitor safety communications.* Managers could examine a sample of safety communications and identify the attitudes implied. For example, are rules made without consultation, are there impractical rules or are workplace precautions implemented without explanation?

Once the awareness of issues relevant to safety culture has been achieved, evidence of relevance to

safety culture issues will be identified in many aspects of how things are done.

Safety culture can be influenced in a large number of ways but it is always necessary to start with the nature of the problem identified – a 'shotgun' approach is unlikely to be successful. The sorts of measurement techniques we have just been discussing can be used to identify specific problems.

When attempting to influence safety culture, the following should be borne in mind:

- It is essential to address senior management issues first. Safety culture can be changed at lower levels in the organization but this change will not be permanent unless senior management continuously supports it.
- Many of the issues to be addressed may not be seen as the province of safety management. For example, a lack of trust on risk assessment issues may be part of a general problem of lack of trust on all issues. It is likely that liaison with functions other than safety will be required if these sorts of problems are to be solved.

What is to be changed is people's attitudes and beliefs and this is a notoriously sensitive and difficult task. If it is to be successful, the process must be taken slowly, with carefully thought out steps. Attempts at radical changes, or too rapid changes, are likely to result in resistance.

References

- Booth, RT. (2000). Challenges and opportunities facing the Institution of Occupational Safety and Health. *Journal of the Institution of Occupational Safety and Health*, 4(1), 7–21.
- Booth, RT. (2003). Occupational safety. In *Oxford Textbook of Medicine*. (eds D.A. Warrell *et al.*), 4th edn, pp. 961–5. Oxford University Press, Oxford.

- Boyle, AJ. (2002). *Health and Safety: Risk Management*, 2nd edn. IOSH Publications, Leicester.
- British Standards Institution (1996). *Guide to Occupational Health and Safety Management Systems*. BS8800. BSI, London.
- British Standards Institution (1999). *Occupational Health and Safety Management Systems – Specification*. OHSAS 18001. BSI, London.
- British Standards Institution (2004). *Guide to Occupational Health and Safety Management Systems*. BS8800. BSI, London.
- Committee on Safety and Health at Work (1972). *Safety and Health at Work. Report of the Committee 1970–1972*. Cmnd 5034. HMSO, London.
- Department of Employment (1974). *Accidents in Factories: the Pattern of Causation and the Scope of Prevention*. HMSO, London.
- Health and Safety Commission (1993). *Third Report: Organising for Safety*. ACSNI Study Group on Human Factors. HMSO, London.
- Health and Safety Executive (1988). *The Tolerability of Risk from Nuclear Power Stations*. HMSO, London.
- Health and Safety Executive (1997). *Successful Health & Safety Management*. Health and Safety Series booklet HSG65. HSE Books, Sudbury.
- Health and Safety Executive (2001). *Reducing Risks, Protecting People*. HSE Books, Sudbury.
- Heinrich, HW. (1969). *Industrial Accident Prevention*, 4th edn. McGraw Hill, New York.
- HM Chief Inspector of Factories and Workshops (1929). *Annual Report of the Chief Inspector of Factories and Workshops For the Year 1929*. Cmnd 3633. HMSO, London.
- Mason, S. (1992). Practical guidelines for improving safety through the reduction of human error. *The Safety and Health Practitioner*, 10, 5.
- Miller, D.P. and Swain, A.D. (1987). Human error and human reliability. In *Handbook of Human Factors* (ed. G. Salvendy). Wiley, New York.
- Rasmussen, J. (1980). What can be learned from human error reports?. In *Changes in Working Life*, (eds K.D. Duncan *et al.*). Wiley, London.
- Reason, JT. (1990). *Human Error*. Cambridge University Press, Cambridge.

Part 6

Control

Chapter 28

Work organization and work-related stress

Tom Cox, Amanda Griffiths and Stavroula Leka

Background

Work organization and work-related stress: the nature of acceptable evidence

The evidence

Failures of work organization: psychosocial and organizational hazards

Work organization, stress and health

Solving the problem

The legal context

Adapting the risk management paradigm

The Nottingham model

Risk assessment

Translation and risk reduction

Final comments

References

Background

Since the Second World War, developed countries, including Britain, have experienced an unprecedented and continuing period of change. This change has fundamentally reshaped the work that we do, the work organizations that employ us and the very nature of our working lives. There are two, possibly three, factors driving these changes: the rapid development of new information and communication technologies, the globalization of free market capitalism and the structural and economic transition of our societies (Cooper *et al.*, 2001; Cox, 2003). It is not surprising therefore that we have also experienced changes in the nature of the threats that challenge our safety and health at work (Cox, 2003). Increasingly, many of those threats arise from failures in the way we design and manage work and work organizations, from failures of 'work organization' and from the psychosocial and organizational hazards that arise from those failures. (Throughout this article, 'work organization' is used as shorthand for 'the design and management of work and work organizations'.) The concept of work-related stress is proving to be central to our understanding of those threats and the mechanisms involved. During the 1990s, it became obvious to all but the ill informed that the experience of work-related stress

had come to present one of *the* major challenges to the safety and health of working people. All of this is driving a critical reappraisal of what has been until now the traditional occupational health paradigm. It has opened the door to an increased contribution of applied psychology to occupational health and the emergence of occupational health psychology (Cox *et al.*, 2000a).

In the European Union, the challenge posed by failures of work organization and by work-related stress has been recognized by the governments of most Member States and by the Commission and its Agencies. It has prompted a concerted effort to develop a practical methodology for managing these issues within the framework of existing safety and health legislation. For many applied researchers and practitioners, managers, trades unionists and policy makers, the obvious way forward was the development of an evidence-based problem-solving process framed by our cumulative knowledge of risk management. This chapter discusses work organization in relation to occupational health taking the management of work-related stress as its framework. It describes the approach developed by the Institute of Work, Health and Organisations in Nottingham with the support of the Health and Safety Executive, several European bodies and a wide variety of private and public sector organizations.

Work organization and work-related stress: the nature of acceptable evidence

The starting point for the development of any problem-solving methodology is the evidence that a problem exists. Harvesting, collating and evaluating such evidence are, however, neither straightforward nor simple matters. The definition of what is and what is not *acceptable* evidence will determine, in large part, the outcome of such activities. This question of definition is, in turn, rooted in the purpose of the overall exercise. If that purpose is to produce a research methodology to promote knowledge through the publication process then the ideal or the perfect will take precedence over that which is adequate and sufficient. If the purpose is to apply a practical methodology capable of supporting action to protect and promote safety and health then the adequate and sufficient takes precedence over the ideal and perfect. (In all walks of life, there are arguments in favour of using methods that are 'good enough' over those that demand perfection, and such arguments often interlace with the notion of being 'fit for purpose'.)

The criteria for what is acceptable by way of evidence are naturally different in the two cases. In the former case, strict criteria based on ideal science (often experimental science) can be used to reduce uncertainty and avoid 'false-positives' in drawing conclusions. However, this approach is extremely conservative and slow. This is not always acceptable in the face of possible hazards of significance. Where such hazards may exist, a different approach to evidence is indicated. The imperative is to avoid false-negatives in drawing conclusions about those hazards, and, within the framework of the precautionary principle, to quickly develop and test solutions to the problems they pose. This difference in approach is obvious when reviews of the literature on work organization and work-related stress are compared.

The evidence

European data from a variety of national and transnational surveys of those in work, or who have recently worked, have identified *stress-*

related problems as among the most commonly reported sources of work-related ill health. For example, the European Foundation's 1996 survey of working conditions in the European Union (European Foundation for the Improvement of Living and Working Conditions, 1997) revealed that of the workers questioned, 57% believed that their work affected their health. The work-related health problems most frequently mentioned were musculoskeletal complaints (30%) and stress (28%). In England and Wales, similar data were drawn from the trailer to the 1990 Labour Force Survey, appended by the Health and Safety Executive, and from the subsequent follow-up survey in 1995 of self-reported work-related illness. Data from these surveys suggest that stress and stress-related illness were second only to musculoskeletal disorders as the major cause of work-related ill health (Hodgson *et al.*, 1993; Jones *et al.*, 1998). At that time, it was estimated that these stress-related problems resulted in about 6.5 million working days lost to industry and commerce each year. In terms of annual costs, estimated using the 1995–96 economic framework, the financial burden to society was £3.7–3.8 billion. Broadly similar data from the most recent survey (2001/2) clearly indicate a marked increase in the incidence of work-related stress. This unwelcome finding is supported by data from other occupational health surveillance systems (such as ODIN).

Survey data such as that referred to above are drawn from a variety of European (and other) countries across at least a 10-year window and gathered using a variety of instruments. These data are supplemented by at least two other sources: empirical data from more focused studies on particular work groups and occupations, again across a wide variety of countries and sectors using an equal variety of approaches and instruments, and practice data from both occupational and primary health care specialists. Although the purist may painstakingly dismiss each individual survey and study on the grounds that it is not methodologically perfect, there is a more pragmatic argument. After rejecting those with obviously fatal flaws, taking such studies and surveys together, the weight of evidence provided tells a reliable enough story to support the belief that

work-related stress is a real and major challenge to safety and health.

Besides making the case that stress is a major problem, the data from the European and the English and Welsh surveys make clear that it is largely work organization factors that trouble working people and are associated, at least in their minds, with the experience of work-related stress and ill health.

Failures of work organization: psychosocial and organizational hazards

Failures of work organization expose working people to psychosocial and organizational hazards. These hazards are summarized, together with examples, in Fig. 28.1. [This categorization is adapted from that first described by Cox (1993).]

A hazard is an event or situation that has the potential for causing harm. Work hazards can be broadly divided into the *physical*, which include, among others, the biological, biomechanical, chemical, mechanical and radiological, and the *psychosocial and organizational*. The International Labour Office (ILO, 1986) discussed psychosocial hazards in terms of the interactions among job content, work organization and management, and other environmental and organizational conditions, on the one hand, and the employees' competencies and needs on the other. Those interactions that prove to be hazardous influence employees' health through their *perceptions* and *experience* (ILO, 1986). Such a model

is conceptually close to many contemporary theories of work-related stress, such as the transactional model put forward by Cox (1978). A simpler definition of psychosocial and organizational hazards has been suggested by Cox and Griffiths (1996): 'those aspects of work design and the organization and management of work, and their social and environmental context, that have the potential for causing psychological, social or physical harm'.

The categorization offered here distinguishes between those failures of work organization that are related to the *content* of work and those that are associated with the *context* of work. It offers five clusters of problem under each heading. It is noted that issues of control and change can be written into all problems regardless of cluster. There is a reasonable consensus in the literature around such a scheme. For example, reference might be made to a somewhat similar but simpler categorization in terms of six factors suggested by Cartwright and Cooper (1997) and discussed by Cooper *et al.* (2001). Those authors distinguish between intrinsic job characteristics, a number of organizational factors and home-work interface and described them as being the main 'environmental' sources of job strain. Many other categorizations and listings exist (see, among others, Cooper and Marshall, 1976; Blomke and Reimer, 1980; Sharit and Salvendy, 1982; Levi, 1984; Baker, 1985; Lohar *et al.*, 1985; Karasek and Theorell, 1990; Sauter *et al.*, 1992; Warr, 1992).

As already noted, new forms of work can give rise to new hazards – not all of which are yet represented in the scientific literature. In the authors' experience, these are often context

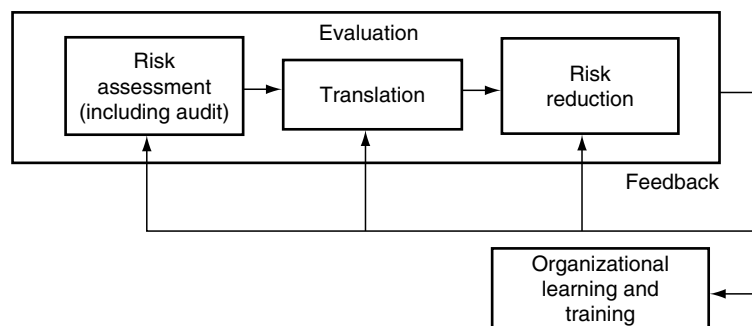


Figure 28.1 A framework model of risk management.

specific, although more generic factors relating to information and knowledge sharing such as poor feedback, inadequate appraisal and communication processes also appear to be of increasing importance (Griffiths, 1998). Among other things, the changing nature of work can make existing knowledge and skills obsolete and, in this scenario, opportunities for innovation and learning become increasingly important. At the same time, performance-related factors such as visibility (in which errors are highly visible), production responsibility (in which the cost of errors is great) and employee interdependence may also be developing as problems (Parker and Wall, 1998).

Although new problems may be emerging with changes in the nature of work and work organizations, old ones may also be finding new life. Examples are provided by issues of job insecurity, excessive working hours and managerial styles that are perceived as 'bullying' (Sparks *et al.*, 2001). However, a good example of an old problem revisited is qualitative underload.

Cooper *et al.* (2001) note that qualitative underload may be a source of job strain in terms of its detrimental effects on the report of anxiety, depression and job satisfaction (Kelly and Cooper, 1981). The distinction between the qualitative and quantitative aspects of workload is not a new one and can be traced back to, at least, the 1960s, using a factor analytical approach. French and his colleagues (1965) suggested a distinction between two different aspects of *overload*: quantitative overload (having too much to do) and qualitative overload (having tasks that are too difficult). Exposure to both types of overload appeared then, as now, to be related to the report of job tension.

Following on from the work of Frankenhauser and Gardell (1976), Cox (1978) applied this distinction to *underload* as well as *overload*: not having enough to do (qualitative underload) and having tasks that are too simple (qualitative underload). He later noted (Cox, 1985), as several others also did, that (1) in unskilled or semiskilled repetitive work, quantitative overload was often associated with qualitative underload, and that (2) in work that was irregular and unpredictable, such as firefighting, periods of quantitative and qualitative overload were interspersed with periods of

quantitative and qualitative underload. In both cases, those employed in such work often reported low levels of control although in different ways and possibly to different effect.

With the changing nature of work and work organizations, there is an increasing use of information and communication technology and, in some cases, this has reintroduced some of the problems of repetitive work and underload observed and researched some 40–60 years ago. At the same time, many organizations have exported repetitive assembly work to the developing countries and thus the risk associated with such work now falls largely on the working members of those countries.

Work organization, stress and health

The authors suggest that work-related stress mediates between employees' exposure to failures of work organization and subsequent effects on health. (There is an interesting question here as to whether work-related stress is simply a convenient and powerful hypothetical construct or a real experience rooted in the person's emotional architecture and processes. Put simply, in what sense does stress exist – in our data or in our experience? For the millions of working people who answered the survey questions about stress referred to earlier in this chapter, the answer is clear – stress is a real experience.)

There appear to be two types of hazard and their effects on health might be mediated by the experience of work-related stress: those physical hazards that evoke anxiety or fear (World Health Organization, 1995) and those hazards that are psychosocial or organizational in nature (Cox, 1993). In a sense, both result from failures of work organization, as defined here, although the link is often clearer in relation to psychosocial and organizational hazards.

Exposure to different types of hazard can contribute to different forms of harm. For example, exposure to organic solvents may have a psychological effect on the person through their direct effects on the brain, through the unpleasantness of their smell or through fear that such exposure

Table 28.1 Psychosocial and organizational hazards (adapted from Cox, 1993).

<i>Content of work</i>	
Job content	Lack of variety or short work cycles, fragmented or meaningless work, under-use of skills, high uncertainty, continuous exposure to people through work
Workload and work pace	Work overload or underload, machine pacing, high levels of time pressure, continually subject to deadlines
Work schedule	Shiftworking, night shifts, inflexible work schedules, unpredictable hours, long or unsociable hours
Control	Low participation in decision-making, lack of control over workload, pacing, shiftworking, etc.
Environment and equipment	Inadequate equipment availability, suitability or maintenance; poor environmental conditions such as lack of space, poor lighting, excessive noise
<i>Context to work</i>	
Organizational culture and function	Poor communication, low levels of support for problem solving and personal development, lack of definition of, or agreement on, organizational objectives
Interpersonal relationships at work	Social or physical isolation, poor relationships with superiors, interpersonal conflict, lack of social support
Role in organization	Lack of participation; role ambiguity, role conflict, and responsibility for people
Career development	Career stagnation and uncertainty, underpromotion or overpromotion, poor pay, job insecurity, low social value to work
Home-work interface	Conflicting demands of work and home, low support at home, dual career problems

might be harmful (Levi, 1981). Indeed, physical hazards can affect health through psycho-physiological as well as physicochemical pathways (Levi, 1984). Psychosocial and organizational hazards (Table 28.1) affect health largely, but not exclusively, through psycho-physiological pathways. For example, violence as a psychosocial and organizational hazard and *an act* may have a direct physical effect on its victim in addition to any psychological trauma or social distress that it causes. Both physical and psychosocial and organizational hazards have the potential for detrimentally affecting social and psychological health as well as physical health. One should not make the mistake of thinking about psychosocial and organizational hazards solely as risks to psychological health (Cox, 1993). Furthermore, significant interactions can occur both between hazards and in their effects on health.

Solving the problem

In a review of the scientific literature on work-related stress commissioned by the Health and

Safety Executive, Cox (1993) formally suggested that the risk management paradigm, used to good effect in dealing with physical hazards, should be applied to issues of work organization and problems of work-related stress. Over the last decade, the Institute of Work, Health and Organisations has conducted a programme of research and development designed to produce a *usable* and *useful* risk management methodology to these ends. The focus of the approach is 'upstream' and is largely preventive.

The major objective of the work conducted at Nottingham was the development and testing of a practical risk management methodology that could be used with different groups (etc.) in different organizations across sectors and countries. As a necessary result, the methodology is *process-based* and requires a manageable degree of tailoring for use in particular situations. It is the process that is transferable. The question of whether the outcomes resulting from the application of the process can be generalized across situations is an empirical one. This is an important point. It often translates into a debate as to whether standard 'off-the-shelf' instruments can be used instead of

the recommended tailored and process-based approach. The former are often instruments developed for purposes of research into stress and not for purposes of risk assessment. It is arguable whether they will prove adequate for the latter as they will, by their very nature, miss both situation-specific problems and include those that are irrelevant to any particular situation. Avoiding these sources of error is one of the purposes of the tailoring process.

The answer to this debate, and to several of the others referred to in this chapter, is framed by the existing safety and health legislation. This is briefly reviewed below.

The legal context

In 1989, the European Commission published the *Framework Directive on the Introduction of Measures to Encourage Improvements in the Safety and Health of Workers at Work* (European Commission, 1989). The duties imposed by this Directive had to be ‘transposed’ into each of the Member States of the European Union within their respective national legislative frameworks and within a specified time frame. The Directive required employers to avoid risks to the safety and health of their employees, to evaluate the risks that cannot be avoided, to combat those risks at source (Article 6:2), to keep themselves informed of the ‘latest advances in technology and scientific findings concerning workplace design’ (Article 10:1) and to ‘consult workers and/or their representatives and allow them to take part in discussions on all questions relating to safety and health at work’ (Article 11:1). Employers were also charged to develop a ‘coherent overall prevention policy that covers technology, organization of work, working conditions, [and] social relationships’ (Article 6:2). In addition, employers were required to ‘be in possession of an assessment of the risks to safety and health at work’ and to ‘decide on the protective measures to be taken’ (Article 9:1).

In Britain, many of these provisions had already been enacted through the *Health and Safety at Work etc. Act 1974* (Health and Safety Executive, 1990) but some of the 1989 requirements, such as

the duty to undertake assessments for *all* risks to health, had to be introduced in the *Management of Health and Safety at Work Regulations 1992* (Health and Safety Commission, 1992) and their revision (Health and Safety Commission, 1999). Employers were advised to consider work-related stress when undertaking their general risk assessments (Health and Safety Executive, 1995). In terms of the law, a risk assessment involves ‘a systematic examination of all aspects of the work undertaken to consider what could cause injury or harm, whether the hazards could be eliminated and, if not, what preventive or protective measures are, or should be, in place to control the risks’ (European Commission, 1996). In other words, employers in the European Union have a responsibility to take reasonable steps to protect their employees from those aspects of work or the working environment that are foreseeably detrimental to safety and health. What is being described in this safety and health legislation is the risk management approach with risk assessment as the initial step. It is easy to see from such descriptions that risk management is essentially a problem-solving methodology.

Adapting the risk management paradigm

The challenge has been to adapt the risk management approach to deal with work organization factors and work-related stress in line with existing European legislation. The use of risk management in health and safety has a substantive history, and there are many texts that present and discuss its general principles and variants (Stranks, 1996; Cox and Tait, 1998; Hurst, 1998) and that discuss its scientific and socio-political contexts (Bate, 1997). Most models incorporate five important elements or principles: (1) a declared focus on a defined work population, workplace, set of operations or particular type of equipment; (2) an assessment of risks; (3) the design and implementation of actions designed to remove or reduce those risks; (4) the evaluation of those actions; and (5) the active and careful management of the process. Perhaps to these should be added a sixth,

organizational learning and training. All of these fundamental elements and principles have been incorporated into the Nottingham process.

It has been argued that there cannot be an exact point-by-point translation of models designed for the management of physical risks to situations involving psychosocial and organizational hazards and the experience of work stress. This is neither a matter of real debate nor is it a problem as there is already a wide variety of effective risk management models in existence both across and within different areas of health and safety. The lack of any felt need to agree on one single model has not hampered progress in health and safety management – quite the reverse. Furthermore, the adaptation of the traditional risk management paradigm to deal with work-related stress does not have to aim at an exhaustive, precisely measured account of all possible stressors for all individuals and all health outcomes. The overriding objective is to produce a reasoned account of the most important work organization factors associated with the experience of stress and ill health (broadly defined) for a defined working group and one grounded in evidence. The account simply needs to be ‘good enough’ to enable employers and employees to move forward in solving the associated problems and comply with their legal duty of care (Griffiths, 1999).

The Nottingham model

At the heart of most risk management models are two distinct but intimately related cycles of activity: risk assessment and risk reduction. This is implicit in the European Commission’s *Guidance on Risk Assessment at Work* (European Commission, 1996): risk management involves ‘a systematic examination of all aspects of the work undertaken to consider what could cause injury or harm, whether the hazards could be eliminated, and if not what preventive or protective measures are, or should be, in place to control the risks’ (Section 3.1). These risk assessment and risk reduction cycles form the basic building blocks for the Nottingham model described here (see Fig. 28.1). They are linked by the processes involved in trans-

lating the output of the former into an input to the latter: translation.

The Nottingham model also includes consideration of ‘evaluation’ and ‘organizational learning and training’. Because all aspects of the risk management process should be evaluated – and not just the outcomes of the risk reduction stage – the ‘evaluation’ stage is treated as supra-ordinate to all the other stages. The risk reduction stage, in practice, tends to involve not only prevention but also actions more orientated towards individual health.

There are parallels between the risk management model and process (see below) developed at Nottingham, and described above, and organizational intervention processes developed by other researchers world-wide. When looking at the potential health effects of work organization, and particularly when attempting to go further and intervene, many applied psychologists have independently formulated somewhat problem-solving approaches and have identified issues that have proved to be common (e.g. Hugentobler *et al.*, 1992; Landsbergis and Vivona-Vaughn, 1995; Lindström, 1995; Schurman and Israel, 1995; Israel *et al.*, 1996, 1998; Kompier *et al.*, 1998; Kompier and Kristensen, 2000; Nytrø *et al.*, 2000; Goldenhar *et al.*, 2001). A key stage in all of these models and processes is the initial analysis (and assessment of risk). [The interested reader is referred to other publications for details of the overall procedure and of the other stages in that procedure – (Cox *et al.*, 2000b, 2002; Griffiths *et al.*, 2003)].

Risk assessment

The initial stage, risk assessment, is designed to identify for a defined employee group – with some certainty and in sufficient detail – significant sources of stress relating to its work and working conditions that can be shown to be associated with an impairment of the health of that group. Several issues follow from this definition of risk assessment, which have implications not only for the design and use of the assessment procedure but also for the overall risk management process. These are presented in Table 28.2. This process is

Table 28.2 Key principles of risk assessment.

Work with defined groups	Each risk assessment is carried out within a defined work group, workplace or function
Focus on working conditions not individuals	Risk assessments are executed in order to identify the aspects of work organization that give rise to the experience of stress and challenges to health and not on the individuals experiencing stress
Focus on 'big issues': significant sources of stress	The focus is on the problems that affect the majority of staff not on individual complaints
Provide evidence of effects of working conditions on health	The process is evidence driven
Use valid and reliable measures	All methods of data collection should be both reliable and valid; employees' expertise provides an important source of information
Maintain confidentiality of information	The confidentiality of information given by individuals must be guaranteed; individual information must be stored securely and not disclosed
Focus on risk reduction as the goal	The risk assessment is designed with risk reduction in mind. Risk assessment tools are designed to provide sufficient detail and context-specific information to allow for control measures to be taken; the emphasis is primarily on prevention and organizational level interventions
Involve employees	The use of participative methods and employee involvement is critical to success

The logic underpinning risk assessment can be operationalized through a six-step process (see Table 28.3).

represented diagrammatically in Fig. 28.2, which provides a schematic summary of the risk assessment strategy.

Much of the information used in the risk assessment is based on the expert judgements of working people aggregated to the group level. However, such judgements are sometimes labelled as *self-report* data and this is taken as grounds for dismissal. However, there is some evidence to suggest that for research into work-related stress it is more appro-

priate to measure work and working conditions as perceived, rather than objectively assessed by supervisors, colleagues, etc. (Spector, 1987; Jex and Spector, 1996; Bosma *et al.*, 1997). From a theoretical perspective, perceptions of work and working conditions may be, in fact, a better predictor of behaviour and health than more objective measures.

It is important to clearly understand that referring to perceived problems of work and working conditions does not place a value or ontological

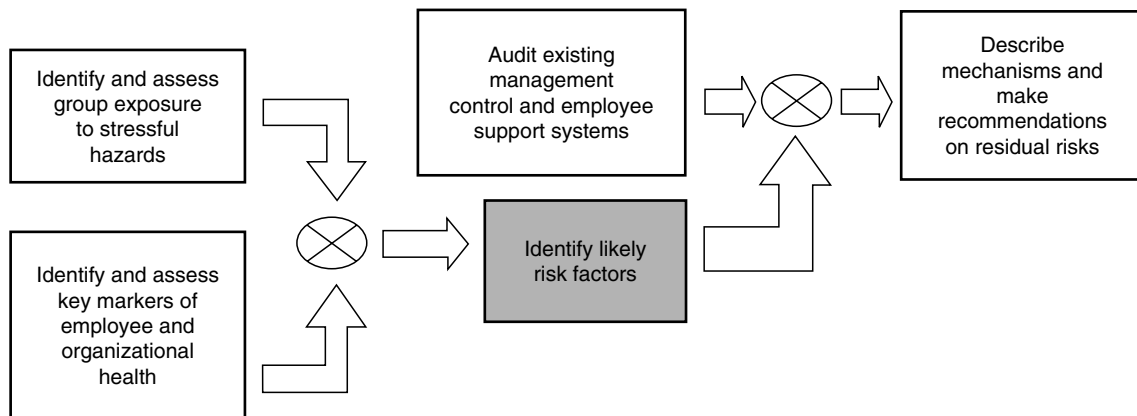


Figure 28.2 The risk assessment strategy.

Table 28.3 Six steps of risk assessment.

Hazard identification	Reliably identify the work organization factors that are in some way inadequate or unacceptable for specified groups of employees, and make an assessment of the degree of exposure. Since many of the problems that give rise to the experience of stress at work are chronic in nature, the proportion of employees reporting a particular aspect of work organization may be a 'good enough' group exposure statistic. There are various ways of measuring and presenting the strength of such consensus
Assessment of harm	Harvest and evaluate evidence that exposure to such work organization factors is associated with the experience of stress and/or impaired health in the group being assessed. This validation exercise should consider the possible detrimental effects of these factors in relation to a wide range of health-related outcomes, including symptoms of general malaise and specific disorders, and of organizational and health-related behaviours such as smoking and drinking, and sickness absence
Identification of likely risk factors	Logically or statistically explore the associations between exposure to the work organization factors identified as hazards and measures of harm to identify 'likely risk factors' at the group level, and to make some estimate of their size and significance*
Description of underlying mechanisms	Understand and describe the possible mechanisms by which exposure to the work organization factors – hazards – is associated with damage to the health of the assessment group or to the organization
Audit existing management control and employee support systems (AMSES)	Identify and assess all existing management systems both in relation to the control and management of the hazards and the experience of work-related stress, and in relation to the provision of support for employees experiencing problems
Draw conclusions about residual risk and priorities	Taking existing management control and employee support systems into account, make recommendations on the residual risk associated with the likely risk factors related to work-related stress and on priorities for action

*Transforming the exposure and health measures to provide binary data at the group level and then relating those measures using frequency based statistics such as odds ratios offers one statistical way forward that is consistent with the logic of a group-based risk assessment.

judgement on those problems: perceptions can be accurate as well as inaccurate and moderated by other factors (e.g. individual differences, see below). The reliability, validity and accuracy of perceptions and self-report data are *empirical* questions, and, as such, can themselves be the subjects of investigation. For example, social desirability effects (a common source of bias) can be tested for and screened out at several stages in the development of the risk assessment (Ferguson and Cox, 1993). Respondents' patterns of reporting can be examined for halo effects; any evidence of differential effects would reduce the likelihood of the assessment data being driven by halo or similar effects (e.g. negative affectivity). The use of different measurement techniques and different sources of data (triangulation) should reduce the likelihood of common method variance (Jick, 1979; Cox *et al.*, 2000c). It would be interesting to

apply this level of questioning to some of the more objective measures used in stress research. What, for example, is the reliability of a measure of blood pressure taken with an electronic sphygmometer? What are its accuracy and its validity as a predictor of heart disease?

The estimation of risk at the group level is effectively the estimation of risk for 'the average employee' and can be contrasted with the estimation of risk for any specified individual. When there is much difference between the two then individual differences obviously exist. There are several points to note here. First, the individual differences that exist can only operate through the person's interaction with their work environment. There is no other logical pathway by which their effects can be made manifest. Second, there is no evidence that the individual differences that exist in respect to the effects of stressors on health are any greater (or

less) than those that exist in relation to other health hazards. Therefore, the existence of individual differences does not negate the overall assessment exercise; rather, it adds an important extra dimension and opens up questions about moderators of the stressor–health relationship. It should be noted here that despite these arguments and the fact of any group-based risk assessments, employers still have a duty of care to the individual.

Translation and risk reduction

The process by which the information provided by the risk assessment is discussed, explored and used to develop interventions has been termed ‘translation’ (Cox *et al.*, 2002). In other models of risk management, such processes have often been ignored or underestimated in their importance.

Usually, the discussion and exploration of the likely risk factors identified in the risk assessment leads to the discovery of underlying organizational pathologies. In turn, their recognition often facilitates the design of economical action plans when those organizational pathologies are targeted rather than their manifestations or symptoms (likely risk factors). Translation often takes some time to accomplish satisfactorily. The development of an action plan involves deciding upon what is to be targeted, the methods used, their integration, those responsible, the proposed time schedule, the resources required and how these interventions will be evaluated. The emphasis in the Nottingham process is on prevention and organization-led interventions.

The interventions required by any action plan can often be integrated into on-going management activities and otherwise planned change. They need not be treated as ‘different’ compared with other management practices. Indeed prevention in relation to work-related stress is largely about good management practice. It is about achieving well-designed, organized and managed work in well-designed, organized and managed workplaces and organizations. Many interventions are simply examples of ensuring good management practice. Examples might be the introduction of regular team meetings or open forums, developing staff

newsletters, reviewing administrative procedures, introducing effective appraisal systems or adjusting rotas. Others, such as increasing staffing levels or installing new equipment, may incur extra costs, but may ‘pay off’ through improved attendance at work or reduced staff turnover, increased productivity and quality or improved creativity and innovation. Preventing the loss of one key member of staff, for example, may save the organization considerable disruption, recruitment and training costs. Detailed examples of interventions that have resulted from risk assessment work carried out in organizations by the authors and their colleagues are provided elsewhere (Cox *et al.*, 2000b; Griffiths *et al.*, 2003).

Although the evaluation of interventions is important, it is often overlooked or deliberately avoided. Not to evaluate is to miss an opportunity. Not only does evaluation tell the organization the extent to which actions have worked, but also why they have worked in that way. It also allows the reassessment of the ‘at-risk’ situation and the methods of assessment used. In all, it provides the basis for organizational learning and establishes a process for continuous improvement. Managing work-related stress is not a one-off activity but part of an on-going cycle of good management at work and the effective management of health and safety.

Final comments

Dealing with work organization issues is an increasingly important part of the challenge of safety and health at work. Many of the effects of such issues appear to be mediated by the experience of stress, and the evidence from survey data tells us that work-related stress is a major source of ill health among the working population. The complex aetiology of work-related stress provides us with an interesting challenge and its mechanisms and causes may never be completely understood in their finest detail. However, there is a moral, as well as a scientific and legal, imperative to act to reduce the harm associated with failures of work organization and work-related stress. The risk management paradigm provides a framework for

positive action – focused on prevention and on work organization. It has already proven successful in a wide range of organizational settings (Cox *et al.*, 2000b, 2002; Griffiths *et al.*, 2003).

References

- Baker, D.B. (1985). The study of stress at work. *Annual Review of Public Health*, 6, 367–81.
- Barling, J., and Griffiths, A. (2002). A history of occupational health psychology. In *Handbook of Occupational Health Psychology*, (eds J.C. Quick and L.Tetrick). American Psychological Association, Washington.
- Bate, R. (1997). *What Risk?* Butterworth-Heinemann, Oxford.
- Blomke, M. and Reimer, F. (1980). *Krankheit und Beruf*. Alfred Huthig Verlag, Heidelberg.
- Bosma, H., Marmot, M.G., Hemingway, H., Nicholson, A.C., Brunner, E. and Stansfeld, S.A. (1997). Low job control and risk of coronary heart disease in Whitehall II (prospective cohort) study. *British Medical Journal*, 314, 70–80.
- Cartwright, S. and Cooper, C.L. (1997). *Managing Workplace Stress*. Sage, Thousand Oaks, CA.
- Cooper, C.L., and Marshall, J. (1976). Occupational sources of stress: a review of the literature relating to coronary heart disease and mental ill health. *Journal of Occupational Psychology*, 49, 11–28.
- Cooper, C.L., Dewe, P.J. and O’Driscoll, M.P. (2001). *Organizational Stress*. Sage, Thousand Oaks, CA.
- Cox, T. (1978). *Stress*. Macmillan, London.
- Cox, T. (1985). Repetitive work: occupational stress and health. In *Job Stress and Blue Collar Work*, (eds C.L. Cooper and M.J. Smith). John Wiley & Sons, Chichester.
- Cox, T. (1993). *Stress Research and Stress Management: Putting Theory to Work*. HSE Books, Sudbury.
- Cox, T. (2003). Work stress: nature, history and challenges. *Science in Parliament*, 60, 10–11.
- Cox, T. and Griffiths, A.J. (1996). The assessment of psychosocial and organisational hazards at work. In *Handbook of Work and Health Psychology* (eds M.J. Schabracq, J.A.M. Winnubst and C.L. Cooper). Wiley & Sons, Chichester.
- Cox, S. and Tait, R. (1998). *Safety, Reliability and Risk Management*. Butterworth-Heinemann, Oxford.
- Cox, T., Baldursson, E. and Rial-Gonzalez, E. (2000a). Occupational health psychology. *Work & Stress*, 14, 101–4.
- Cox, T., Griffiths, A., Barlow, C., Randall, R., Thomson, T. and Rial-González, E. (2000b). *Organisational Interventions for Work Stress: A Risk Management Approach*. HSE Books, Sudbury.
- Cox, T., Griffiths, A., and Rial-González, E. (2000c). *Research on Work-related Stress*. Office for Official Publications of the European Communities, Luxembourg.
- Cox, T., Randall, R., and Griffiths, A. (2002). *Interventions to Control Stress at Work in Hospital Staff*. HSE Books, Sudbury.
- European Commission (1989). Council Framework Directive on the Introduction of Measures to Encourage Improvements in the Safety and Health of Workers at Work. 89/391/EEC. *Official Journal of the European Communities*, 32, No L183, 1–8.
- European Commission (1996). *Guidance on Risk Assessment at Work*. European Commission, Brussels.
- European Foundation for the Improvement of Living and Working Conditions (1997). *Working Conditions in the European Union*. Dublin.
- Ferguson E. and Cox T. (1993). Exploratory factor analysis: A users’ guide. *International Journal of Selection and Assessment*, 1, 84–94.
- Frankenhauser, M. and Gardell, B. (1976). Underload and overload in working life: outline of a multidisciplinary approach. *Journal of Human Stress*, 2, 15–23.
- Gardell, B. (1982). Work participation and autonomy: a multilevel approach to democracy at the workplace. *International Journal of Health Services*, 12, 31–41.
- Goldenhar, L.M., LaMontagne, A.D. Katz, T., Heaney, C. and Landsbergis, P. (2001). The intervention research process in occupational safety and health: An overview from the NORA Intervention Effectiveness Research Team. *Journal of Occupational and Environmental Medicine*, 43, 616–22.
- Griffiths, A. (1995). Organizational interventions: facing the limits of the natural science paradigm. *Scandinavian Journal of Work, Environment and Health*, 25, 589–96.
- Griffiths, A. (1998). The psychosocial work environment. In *The Changing Nature of Occupational Health* (eds R.C. McCaig and M.J. Harrington). HSE Books, Sudbury.
- Griffiths, A., Randall, R., Santos, A. and Cox, T. (2003). Senior nurses: interventions to reduce work stress. In *Occupational Stress in the Service Professions* (eds M. Dollard, A. Winefield and H. Winefield). Taylor & Francis, London.
- Health and Safety Commission (1992). *Management of Health and Safety at Work Regulations*. HMSO, London.
- Health and Safety Commission (1999). *Management of Health and Safety at Work Regulations: Approved Code of Practice and Guidance*. HMSO, London.
- Health and Safety Executive (1990). *A Guide to the Health and Safety at Work etc. Act 1974*. HSE Books, Sudbury.
- Health and Safety Executive (1995). *Stress at Work: A Guide for Employers*. HSE Books, Sudbury.
- Hernberg, S. (1994). Editorial: 20th Anniversary Issue. *Scandinavian Journal of Work, Environment and Health*, 20, 5–7.
- Hodgson, J.T., Jones, J.R., Elliott, R.C. and Osman, J. (1993). *Self-reported Work-related Illness*. HSE Books, Sudbury.
- Hugentobler, M.K., Israel, B.A. and Schurman, S.J. (1992). An action research approach to workplace health: integrating methods. *Health Education Quarterly*, 19, 55–76.

- Hurst, N.W. (1998). *Risk Assessment: The Human Dimension*. Royal Society of Chemistry, Cambridge.
- International Labour Office (1986). *Psychosocial Factors at Work: Recognition and Control*. Occupational Safety and Health Series no. 56. International Labour Office, Geneva.
- Israel, B.A., Baker, E.A., Goldenhar, L.M., Heaney, C.A. and Schurman, S.J. (1996). Occupational stress, safety and health: Conceptual framework and principles for effective preventions. *Journal of Occupational Health Psychology*, **1**, 261–86.
- Israel, B.A., Schulz, A.J., Parker, E.A. and Becker, A.B. (1998). Review of community-based research: assessing partnership approaches to improve public health. *Annual Review of Public Health*, **19**, 173–202.
- Jex, S.M. and Spector P.E. (1996). The impact of negative affectivity on stressor-strain relations: a replication and extension. *Work and Stress*, **10**, 36–45.
- Jick, T.D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, **24**, 602–11.
- Johnson, J.V. (1996). Conceptual and methodological developments in occupational stress research. An introduction to state-of-the-art reviews I. *Journal of Occupational Health Psychology*, **1**, 6–8.
- Jones, J.R., Hodgson, J.T., Clegg, T.A. and Elliot, R.C. (1998). *Self-reported Work-related Illness in 1995*. HSE Books, Sudbury.
- Karasek, R.A. and Theorell, T. (1990). *Healthy Work*. Basic Books, New York.
- Kelly, M. and Cooper, C.L. (1981). Stress among blue collar workers: a case study of the steel industry. *Employee Relations*, **3**, 6–9.
- Kompier, M.A.J., and Kristensen, T.S. (2000). Organizational work stress interventions in a theoretical, methodological and practical context. In *Stress in Occupations: Past, Present and Future* (ed. J. Dunham). Whurr Publishers, London.
- Kompier, M.A.J., Geurts, S.A.E., Grundemann, R.W.M., Vink, P. and Smulders, P.G.W. (1998). Cases in stress prevention: the success of a participative and stepwise approach. *Stress Medicine*, **14**, 155–68.
- Landsbergis, P.A. and Vivona-Vaughn, E. (1995). Evaluation of an occupational stress intervention in a public agency. *Journal of Organizational Behaviour*, **16**, 29–48.
- Leka, S., Griffiths, A. and Cox, T. (2003). *Work Organization and Stress*. World Health Organization, Geneva.
- Levi, L. (1981). *Preventing Work Stress*. Addison-Wesley, Reading, MA.
- Levi, L. (1984). *Stress in Industry: Causes, Effects and Prevention*. Occupational Safety and Health Series No. 51. International Labour Office, Geneva.
- Lindström, K. (1995). Finnish research in organizational development and job redesign. In *Job Stress Interventions* (eds L.R. Murphy, J.J. Hurrell Jr, S.L. Sauter and G. Puryear Keita). American Psychological Association, Washington, DC.
- Lohar, B.T., Noe, R.A., Moeller, N.L. and Fitzgerald, M.P. (1985). A meta-analysis of the relation of job characteristics to job satisfaction. *Journal of Applied Psychology*, **70**, 280–9.
- Nytrø, K., Saksvik, P.O., Mikkelsen, A., Bohle, P. and Quinlan, M. (2000). An appraisal of key factors in the implementation of occupational stress interventions. *Work and Stress*, **3**, 213–25.
- Parker, S., and Wall, T. (1998). *Job and Work Design: Organizing Work to Promote Well-being and Effectiveness*. Sage, London.
- Randall, R., Griffiths, A., and Cox, T. (2001). Using the uncontrolled work setting to shape the evaluation of work stress interventions. In *Occupational Health Psychology: Europe 2001* (eds C. Weikert, E. Torkelson and J. Pryce). I-WHO Publications, Nottingham.
- Sauter, S.L., Murphy, L.R., and Hurrell, J.J. (1992). Prevention of work-related psychological disorders: a national strategy. In *Work and Well-Being: An Agenda for the 1990s* (eds G.P. Keita and S.L. Sauter). American Psychological Association, Washington, DC.
- Schurman, S.J. and Israel, B.A. (1995). Redesigning work systems to reduce stress: a participatory action research approach to creating change. In *Job Stress Interventions* (eds L.R. Murphy, J.J. Hurrell Jr, S.L. Sauter and G. Puryear Keita). American Psychological Association, Washington, DC.
- Sharit, J., and Salvendy, G. (1982). Occupational stress: review and appraisal. *Human Factors*, **24**, 129–62.
- Sparks, K., Faragher, B. and Cooper, C. (2001). Well-being and occupational health in the 21st century workplace. *Journal of Occupational and Organizational Psychology*, **74**, 489–509.
- Spector, P.E. (1987). Interactive effects of perceived control and job stressors on affective reactions and health outcomes for clerical workers. *Work and Stress*, **1**, 155–62.
- Stranks, J. (1996). *The Law and Practice of Risk Assessment*. Pitman, London.
- Warr, P.B. (1992). Job features and excessive stress. In *Prevention of Mental Ill Health at Work* (eds R. Jenkins and N. Coney). HMSO, London.
- World Health Organization (1995). *Health Consequences of the Chernobyl Accident*. World Health Organization, Geneva.

Chapter 29

Control philosophy

Kerry Gardiner

Introduction
Control at the design stage
Control after the design stage
 Elimination
 Substitution
 Changing the process
 Ventilation
 Isolation or segregation

Maintenance and housekeeping
Education and training
 Management
 Workforces
Personal protective equipment
 Source, transmission and the individual
Summary
Reference

Introduction

The prevention or reduction of ill health at work relies on the elimination or control moderation of workplace contaminants (chemicals to psychosocial hazards). Effective measures can range from the simple to the esoteric, with this chapter aiming to provide a brief review of the hierarchical structure in which these reside and of which ventilation (Chapter 30) and personal protective equipment (Chapter 31) are part. More contaminant-specific means of control are discussed in other chapters (Chapters 15–29). Recently, the advent of a technique called ‘Control Banding’ which, as an approach that does not rely on measurement to specify control, is being promulgated and its use for SMEs encouraged.

It is common to split control methods into two groups: (1) software or administrative controls and (2) hardware or engineering controls. In the main, the hardware or engineering controls are preferable. This is not necessarily due to their effectiveness but more to their longevity. The software and hardware hierarchy is tabulated in Table 29.1.

Control at the design stage

It is much more effective, both in terms of outcome and cost, to instigate control measures when the

process or factory is still at the design stage. The following example, particularly for chemical hazards, includes a number of the issues that should be considered.

- 1 Careful a priori selection of the substances/process/people for use.
- 2 Enclosed system from raw material to product.
- 3 Prevent leakage of raw materials, intermediates, by-products, product and waste from equipment.
- 4 Vent unwanted contaminants to scrubbers, absorbers or incinerators.
- 5 Automate the process and control it remotely.
- 6 Consider the possibility of the amount of work in the workplace increasing over time, such as the installation of additional degreasing tanks, welding bays, sources of noise, etc. This should minimize the likelihood of the workplace being radically different from that at the time the controls were designed/commissioned.
- 7 ‘Interactions’ between two or more contaminants. For example, chlorinated organic solvents used in degreasing baths may decompose when near to sources of ultraviolet radiation (such as those from welding) and form phosgene (a strong upper respiratory tract irritant).
- 8 Provide a facility for the removal of residues from the system before opening it.
- 9 The minimization of maintenance requirements, for example:

Table 29.1 Engineering and administrative control techniques.

<i>Engineering</i>	<i>Administrative</i>
Appropriate design engineering	Appropriate administrative control by design elimination
Total or partial enclosure	Substitution
Local exhaust ventilation (LEV)	Isolation or segregation
Change the process	Maintenance and housekeeping
Shielding	Education and training
Personal protective equipment (PPE)	Personal hygiene

(a) continuous leak detection for fugitive emissions;

(b) careful design for infrequent but major tasks, such as the replacement of filter bags.

10 Chemical specification poses questions related to the following:

(a) is it necessary (are there alternative substances or forms?);

(b) its physical properties (e.g. can it be stored, its temperature and pressure during storage or use, its flammability?);

(c) its chemical properties (e.g. storage, impurities, likely intermediary products, degradation products, etc.);

(d) its toxicity and occupational exposure limit (OEL);

(e) special handling requirements;

(f) special hazards;

(g) can the plant contain it adequately?;

(h) can the excess or waste material be recovered?;

(i) emergency procedures;

(j) special hygiene or medical requirements.

Control after the design stage

Generally, however, the occupational hygienist has had little or no influence in the specification of control requirements at the design stage and is therefore left with a number of remedial options. Fortunately, in some forward-thinking companies the occupational hygienist is asked to contribute to the design process and 'sign off' their approval. There are many permutations of control but the majority lie within the broad headings described below.

Elimination

This is the most effective means of control because by definition, if the contaminant is no longer present then it can pose no risk. Unfortunately, most contaminants are being used or generated for, or as a result of, a specific purpose. However, it is a question one must ask and be sure that it has been addressed satisfactorily.

Substitution

Once a contaminant is believed to give rise to an unacceptable level of risk, and it is not possible to eliminate its use, one must look to substitute it for one generating a lower and more acceptable level of risk. If the risk is lower but still unacceptable then other means of control should be used. Care needs to be taken to ensure that a different problem is not created in terms of issues such as flammability or chemical interaction. Figure 29.1 is a flow diagram to assist in the process of substitution. Having identified an unacceptable risk, alternatives should be identified. If none is available then the risks should be managed by other methods. However, if alternatives are available then the consequences of their use, both in terms of the risk to health and their suitability in the process, should be considered. If alternatives appear to be compatible with the process and pose a lower risk to health than the original then they should be compared with each other. The most suitable alternative should be chosen and the change implemented. The effectiveness of this substitution should be evaluated and, if found to be acceptable, then at regular intervals its use should be reviewed with an aim to substituting again.

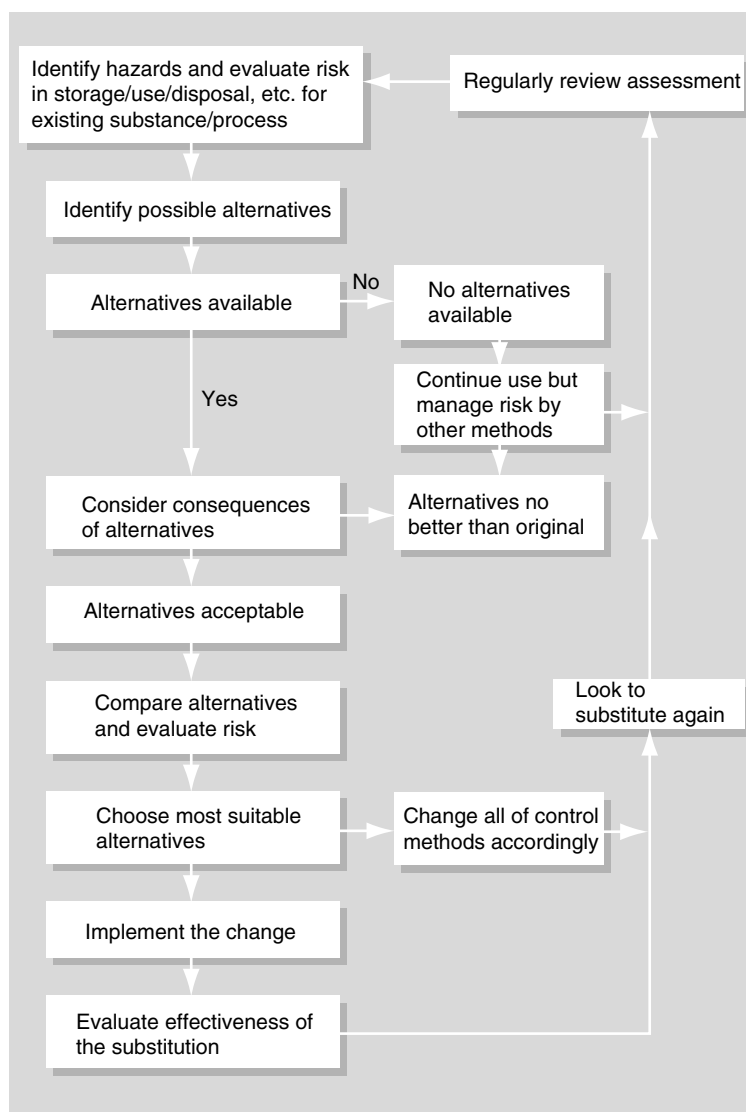


Figure 29.1 Flow diagram to assist in the process of substitution of a contaminant.

Classic examples of substitution include the use of: MDI instead of TDI (due to its lower volatility); zinc, barium or titanium dioxide instead of white lead in paint (due to their lower toxicity); phosphorus sesquisulphide instead of white phosphorus in matches (due to its lower toxicity); and synthetic mineral fibre instead of asbestos (perceived at the time to be of lower carcinogenicity).

There is also opportunity to substitute the state or form of the same substance for one which gives rise to less exposure. Clearly, a reduction in temperature or increase in pressure will change the

physical state of some contaminants with the likelihood that exposure decreases as follows: gas or vapour >> liquid > solid. In addition, for particulates it is possible to increase their aerodynamic diameter by pelletization or the formation of briquettes, thereby radically reducing the amount of inhalable aerosol.

Changing the process

Significant reductions in exposure can be achieved by quite minor modifications to the process or the

conditions under which it operates. The following are a few examples.

- Reduction of the temperature at which a process operates (especially volatile liquids) along with the use of drip trays.
- Reduction of the amount of agitation contaminants receive in a process – such as the liquid surface of a degreasing tank or the movement of bags along a conveyor belt when the rollers are spaced far apart.
- Reduction of the surface area from which liquid evaporation or splashes can occur by covering it with plastic balls or foam that float on the surface.
- Reduction of the pressure at which the process operates. If it is reduced to below that of atmospheric pressure it will ensure that any leaks are inward.
- Increase in the size of the individual solid contaminants from fine particulates to pellets or flakes or, if possible, to be kept in solution or as a slurry or paste.
- The use of automatic metering systems rather than manual contact with liquids or solids.
- The use of process-compatible bags in which the bag material itself is either of use or at worst is not detrimental to the process.
- The use of electrostatic spraying or dipping rather than spraying or manual application.
- The order in which work is undertaken.

Ventilation

Ventilation can be used to control a number of different contaminants (gases or vapours, dusts, heat, etc.) and is usually split into three types: (1) general; (2) dilution; and (3) local extract or exhaust ventilation (LEV). Clearly, within the remit of this chapter, LEV is of most importance due to its proximity to the source; however, its complexity in terms of specification and design warrants further explanation (see Chapter 30).

Isolation or segregation

This means of control attempts to remove individuals who are potentially exposed from the proximity of the source. It is simple and can be effective; however, ultimately it does nothing to remove the

hazard. There are a number of different means by which the potentially exposed individual can be segregated or isolated from the source of the contaminant and these will be considered in turn.

1 *Total enclosure* of the process (preferably under negative pressure, e.g. shot blasting).

2 *Partial enclosure* of the process with LEV (e.g. fume cupboard).

3 A physical *barrier* can be placed between the source and the receiver, thereby absorbing or reflecting the contaminant, as used for noise (Chapter 17) or ionizing radiation (Chapter 22). This can also be used to reduce the number of workers exposed.

4 The *distance* between the source and the receiver can be maximized (especially where the inverse square ($1/r^2$) law applies for point sources) such as with the thermal environment (Chapter 20), noise (Chapter 17) and ionizing radiation (Chapter 22).

5 The *time* of exposure can be controlled by either minimizing activities in the proximity of the source of exposure or by rotating the workforce, or preferably by ensuring that by examination of the cyclic nature of the process the tasks are carried out when there is no chance of exposure.

6 *Age* may be used as a means of selecting members of the workforce who are capable of carrying out heavy or hot or cold work because of their physical advantage. However, it would be preferable to adapt or control the workplace to such an extent that anyone could work there.

7 *Sex* may be used as a means of selecting members of a workforce who are not, or should not be, allowed to be exposed to certain substances. This is mainly to protect women of child-bearing age and potentially their fetuses. As in the previous example with age, sex may be used as a means of selection for physically demanding jobs – again profound doubts exist about the justification for this, both ethically and morally.

Maintenance and housekeeping

Proactive maintenance schedules minimize the likelihood of breakdown and spills, etc. with the provision of planned shut-down periods for major work. As far as possible, residual contaminant

ants should be removed before the system is opened up.

Despite having made every effort to prevent or minimize the release of contaminants into the workplace and the generation of excess and waste material, there will always be the unexpected leak or spill. It is therefore necessary to identify likely sites for fugitive emissions and to instigate well-planned and coordinated housekeeping. This relies on: (1) the correct procedure having been identified; (2) the appropriate remedial materials being readily available at the anticipated sites; (3) the personnel being suitably trained; and (4) the waste being correctly removed. However, as with a lot of control techniques it is often the simple technique that is effective, such as: the use of vacuum cleaners rather than brooms or compressed air; the removal of settled particulate from horizontal surfaces before secondary generation; and the immediate disposal of solvent-soaked rags rather than leaving them on bench tops, etc.

In some workplaces, the nature of the contaminants may make it necessary to have facilities to allow the workforce to maintain good personal hygiene. The ability of some contaminants to be rapidly absorbed via the skin, or to interact with it, makes it necessary for the workforce to have access to good washing facilities, along with the correct washing media. Care needs to be taken as some surfactants are aggressive to the skin, perhaps exacerbating the defatting action of solvents, etc. (see Chapter 4). It may also be necessary to supply work clothes and ensure that these are changed and washed at whatever frequency is deemed appropriate. The potential for a contaminant to be absorbed by ingestion also necessitates the need for designated areas for eating, drinking and smoking.

Education and training

The provision of information, instruction and training is required to supplement the more permanent means of control. It is necessary for management, supervisors and the individuals undertaking the work to be informed of the relevant information.

Management

Management should be aware of the safety and health hazards in their area (processes, operations and materials) and under what circumstances assistance is required to evaluate and/or control these. Of great importance is that managers and supervisors know everything the workforce have been told and the consequences of non-conformance, both for the individual and the company.

Workforces

In addition to their normal operating instructions, the workforce should be informed about the following:

- the specific means by which they can reduce their own exposure;
- the activities in the workplace during which hazardous chemicals are present;
- the means of identifying control defects resulting in the non-routine presence of hazardous chemicals in the workplace;
- potential hazards of non-routine tasks;
- potential health effects;
- how to use the control methods provided and the consequences of non-use (both the health effects and disciplinary action);
- explanation of the labelling system;
- explanation and location of material safety data sheets (MSDSs);
- reporting defects.

Personal protective equipment

This form of control is often inappropriate and ineffective, and is usually the least desirable for the operative. By the use of this technique, no attempt is made to reduce or eliminate the hazard. It should therefore only be considered or used as a last resort. However, there are four situations in which its use is defensible. First, when it is not technically feasible to control exposure by any other means, for example, for individuals who have to work inside the paint-spraying booths at car manufacturers. Second, for emergency procedures such as major spillages. Third, for maintenance work in which the usual controls have been

switched off to facilitate access; and fourth, when an assessment of the risks to health has shown that there is an immediate risk that needs to be controlled until such time as other means of control can be specified, installed and their effectiveness evaluated.

A great variety of equipment is included under the generic title of personal protective equipment, such as: respiratory protective equipment (RPE); hard hats; safety spectacles; face shields; safety boots; hearing protection (muffs or plugs); overalls (cloth, plastic or chain mail); gloves; and protective creams and lotions. A comprehensive dissertation is provided in Chapter 31.

Source, transmission and the individual

This philosophy is represented diagrammatically in Figure 29.2. In terms of effectiveness, it is also beneficial to consider the control of an individual's exposure as three distinct components: (1) control at the point of release or source; (2) prevent or control transmission of the contaminant to the individual; and (3) protection of the worker to minimize exposure and absorption. In occupational health terms, control is related to exposure to workplace contaminants, with the magnitude of absorption being dictated by the nature of the contaminant, the routes of entry into the body, the concentration and the duration of exposure.

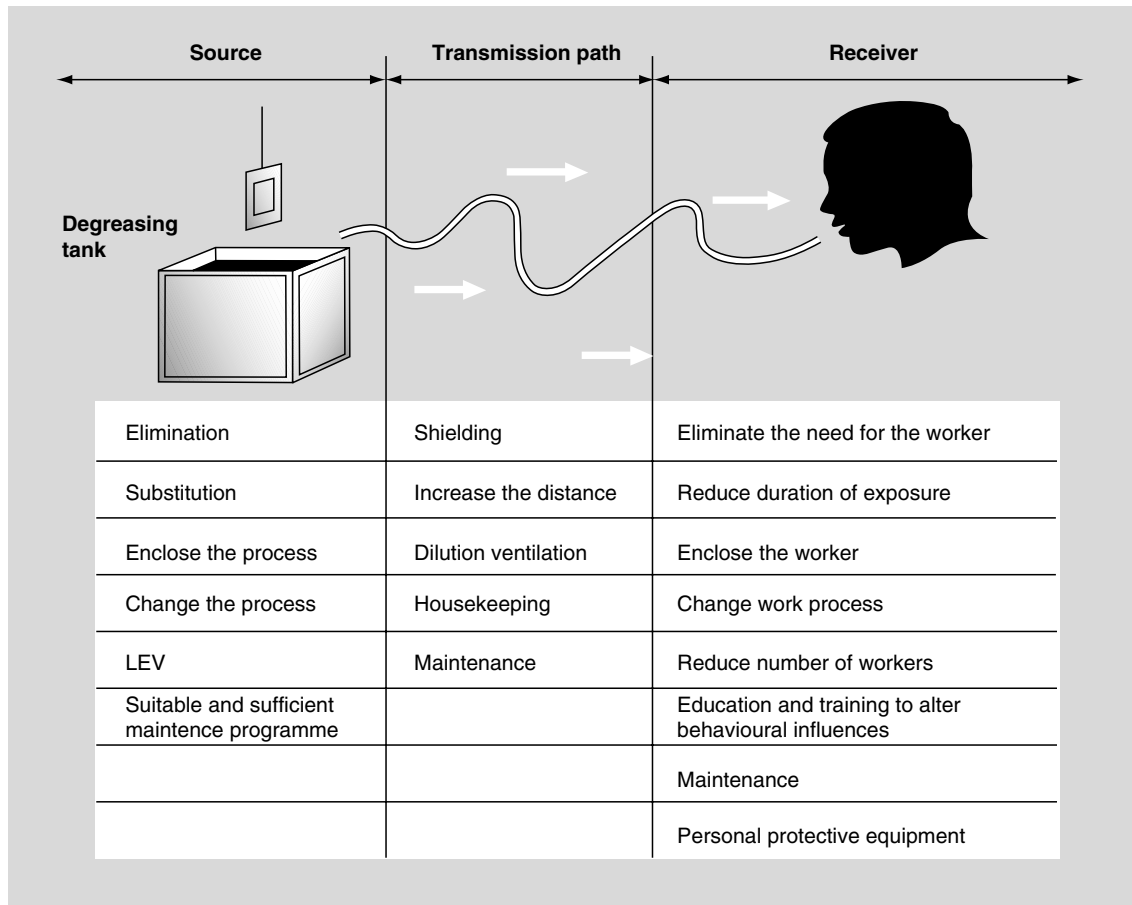


Figure 29.2 Movement of a contaminant from source to receiver with control techniques for each component (after Olishitski, 1988).

Clearly, to reduce this exposure it is necessary to influence one or more of these factors.

As mentioned earlier, significant progress has been made with the utilization of some basic parameters for the control of chemical risks. The HSE's 'COSHH Essentials', and the ILO's 'Control Toolkit' combine known facts about vapour pressure/dustiness, quantities, toxicity, nature of use, etc. to provide simplistic/prescriptive control-banding guidance. Much debate exists as to whether this has dumbed down the 'science' or whether the (probably) over-protective degree of control now available to many is a higher goal.

Other significant advances in control have been achieved by programmes aimed at achieving behavioural change (as in the BP Global Alliance ABC safety culture programme) or Bovis Lend Leases's Incident and Injury Free (IIF) culture promotion. On many occasions, where health and safety performance plateaus, such strategies of trying to assist the whole workforce change its perceptions and attitudes could liberate significant improvements. Clearly, when the problems are

predominantly health related (particularly psychosocial) then these techniques require some development but have huge potential.

Summary

A great many means of controlling exposure exist. It is most convenient to categorize these in terms of engineering and administrative controls, and in terms of the preferred location of control, i.e. source, transmission path and receiver. Table 29.1 provides a list of engineering and administrative controls and Figure 29.2 shows the movement of the contaminant from source to receiver, with a list of the suitable control techniques for each component.

Reference

Olishitski, J.B. (1988). Methods of control. In *Fundamentals of Industrial Hygiene*, 3rd edn (ed. B.A. Plog). National Safety Council, Chicago.

Chapter 30

Ventilation

Frank Gill

Introduction	Fletcher method
Units used	Garrison method
Air density	Capture and transport velocities
Reynolds number (Re)	Dilution ventilation
Pressure	Required volume flow rate for dilution ventilation
Volume and mass flow rate	Ducts and fittings
Zones of influence at terminals	Duct sizes
The measurement of airflow	Pressure losses
Pressure measurement	Suggested method of duct sizing and loss calculation
Smoke tracers	Balancing of multibranching systems
Air velocity instruments	Fans
Pitot-static tube	Fan power and efficiency
Rotating vane anemometers	Fan characteristic curves
Heated head anemometers	Fan types
Calibration	Propeller fan
Techniques for using air velocity instruments	Axial flow fan
Extraction ventilation	Centrifugal fan
Design features and volume flow rates	Matching of fan and system
Enclosures	Air cleaning and discharge to atmosphere
Hoods	Make up air
Slots	References
Prediction of performance	Further reading

Introduction

Ventilation is one of the most powerful tools that the occupational hygienist can use to control the working environment. This chapter will cover some of the design parameters that are required to make the best use of airflow and also provide instruction in the measurement of the performance of existing ventilation systems.

Units used

The most common units used in ventilation engineering are given in Table 30.1 in both imperial and Système International (SI) units together with their dimensions and conversions from one system to the other.

Air density

For most ventilation engineering problems, standard air density can be used. The exceptions are if high-temperature air is being dealt with or if the ventilation system is to be used in places where the barometric pressure is substantially different from normal, for example in deep mines or at high altitudes. Standard air density is taken as 1.2 kg m^{-3} (0.075 lb ft^{-3}), which corresponds to air at a barometric pressure of 1013.25 mb (760 mmHg) and at a temperature of 20°C (dry bulb temperature). Departures from these conditions can be corrected by using the expression:

$$\rho_0 = \rho_s \times b_0/b_s \times T_s/T_0 \quad (30.1)$$

Table 30.1 Common units used in ventilation engineering.

Unit	Dimension*	Imperial system	SI	Conversion factors
Length	L	Foot (ft) or inch (in)	Metre (m) or millimetre (mm)	$\text{ft} \times 0.305 = \text{m}$
Area	L^2	Square foot (ft ²)	Square metre (m ²)	$\text{ft}^2 \times 0.093 = \text{m}^2$
Air velocity	$L t^{-1}$	Feet per minute (ft min ⁻¹) feet per second (ft s ⁻¹)	Metre per second (m s ⁻¹)	$\text{ft min}^{-1} \times 0.0051 = \text{m s}^{-1}$ or $1 \text{ m s}^{-1} = 197 \text{ ft min}^{-1}$ $\text{ft s}^{-1} \times 0.305 = \text{m s}^{-1}$
Air volume flow rate	$L^3 t^{-1}$	Cubic feet per minute (ft ³ min ⁻¹) air changes per hour [†]	Cubic metre per second (m ³ s ⁻¹) air changes per hour [†]	$\text{ft}^3 \text{ min}^{-1} \times 0.000472 = \text{m}^3 \text{ s}^{-1}$ or $1 \text{ m}^3 \text{ s}^{-1} = 2119 \text{ ft}^3 \text{ min}^{-1}$
Pressure (force per unit area)	$MLt^{-2} L^{-2}$	Pounds force per square foot (lb _f ft ⁻²) inches of water column: 1 in H ₂ O = 5.2 lb _f ft ⁻²	Newton per square metre or pascal (N m ⁻²): millibar (mb) = 100 Pa	$\text{lb}_f \text{ ft}^{-2} \times 47.9 = \text{Nm}^{-2}$ (Pa) inch water $\times 249 = \text{N m}^{-2}$ (Pa) – Note: newton (N) = 1 kg m s ⁻²
Power (work done per unit time)	$ML^2 t^{-3}$	Horsepower (33000 ft lb _f min ⁻¹)	Watt (joule per second) (newton metre per second)	Horsepower $\times 746 = \text{watt}$ (W)
Air density	ML^{-3}	Pound per cubic foot (lb ft ⁻³)	Kilogram per cubic metre (kg m ⁻³)	$\text{lb ft}^{-3} \times 16.02 = \text{kg m}^{-3}$

*As used in dimensional analysis: L , length; t , time; and M , mass.

[†]It is sometimes convenient to express air volume flow rate as air changes per hour. To convert to a usable unit it is necessary to multiply the number of air changes by the volume of the room thus giving the result in cubic feet or cubic metres per hour.

where ρ_0 is air density at the non-standard conditions, ρ_s is standard air density, b_0 is barometric pressure at the conditions, b_s is barometric pressure at standard conditions, T_0 is absolute temperature at the conditions and T_s is absolute temperature at standard conditions.

Reynolds number (Re)

This number defines the nature of fluid flow from ‘streamlined’ at low velocities to ‘fully turbulent’ at high velocities. It is a dimensionless number that is the ratio of momentum forces to viscous forces of the fluid flowing. In air when Re is below 2000, streamlined flow exists and the energy required to move it is proportional to the velocity but, when Re is above 4000, fully developed turbulent flow exists and the energy is proportional to the velocity squared. This is important when calculating pressure losses through ductwork and fittings. In most ventilation systems Re is above 4000 and the square law holds. The exception is with fabric

filters when air velocities are extremely low and Re is below 2000.

Pressure

Air requires a pressure difference for it to flow and will always flow from the higher to the lower pressure. Pressure is a type of energy that appears in two forms: (1) static pressure (p_s) and (2) velocity pressure (p_v). The sum of these two pressures is known as total pressure (p_t).

Static pressure is the pressure exerted in all directions by a fluid that is stationary, but when in motion it is measured at right angles to the direction of flow to eliminate the effects of velocity. It can be positive or negative. On the suction side of a fan the static pressure is negative but positive on the discharge side.

Velocity pressure is the pressure equivalent of the kinetic energy of fluid in motion and is calculated from the following formula:

$$p_v = \rho v^2 / 2 \quad (30.2)$$

where ρ is the air density and v is the air velocity. Velocity pressure provides the force to move sailing craft and to damage buildings in a high wind. The formula shown above is widely used in the measurement of air velocity and in the calculation of pressure loss in ductwork and fittings. If a standard air density of 1.2 kg m^{-3} is used then the expression becomes:

$$p_v = 0.6v^2 \quad (30.3)$$

and if v is in metres per second, p_v will be in newtons per square metre. Velocity pressure is always positive.

Volume and mass flow rate

When a quantity of air is moving within a boundary of a duct or a tunnel, the volume flow rate (Q) is calculated from the formula:

$$Q = vA \quad (30.4)$$

where v in this case is the average air velocity over the cross-section of the duct, and A is the cross-sectional area of the duct at the place where the velocity is taken. If v is in metres per second and A is in square metres, then Q will be cubic metres per second.

The mass flow (M) is related to Q as follows:

$$M = \rho Q \quad (30.5)$$

where ρ is the density of air. If ρ is in kilograms per cubic metres then M will be in kilograms per second.

The volume flow rate is used when specifying the duties of fans or when calculating the dilution rates of pollutants. Mass flow rates are used to calculate quantities of heat required to temper the air in heating and air-conditioning systems.

Zones of influence at terminals

One important aspect of supply and extract ventilation influencing the success or failure of the design is the behaviour of airstreams close to the points of entry to and exit from the system. On the supply or discharge side, air leaves the system in a jet, which expands at an angle of approximately 20° in a conical shape. Thus, if a jet of

air is discharged from a parallel-sided duct at a velocity of 10 m s^{-1} , the air in the room will be influenced for a considerable distance from the exit. For example, at a distance of 30 diameters from the exit, along the centre line, the air velocity will be 1 m s^{-1} . The zone of influence on the suction side of the mouth of a system is spherical in shape. As the surface area of a sphere is proportional to the square of the radius, the air velocity decays inversely as the square of the distance. Thus within one duct diameter of the mouth the velocity has dropped to approximately one-tenth and is virtually non-directional. For this reason, extract points must be placed close to the source of emission, preferably within one diameter of it.

The measurement of airflow

When examining a ventilation system to assess its performance it may be necessary to know:

- 1 whether it is successfully capturing the pollutants at their point of release;
- 2 what the velocities of air are at various places in the system;
- 3 how much pressure or suction the fan is developing; and
- 4 how much pressure is being absorbed in different parts of the system, in particular by the filters or dust collectors.

Thus, the technique of measurement involves measuring air velocities, pressure differences and observing the path of the air by means of a tracer such as smoke.

Pressure measurement

In ventilation systems, pressure difference between various points is required thus a gauge pressure rather than an absolute value is measured. The simplest gauges are U-tubes filled with liquid (water or paraffin) usually with one limb inclined for more precision. They are known as 'manometers'. The two limbs are connected to the two places where a pressure difference is required. The height difference between the liquid in each column denotes the pressure. Most proprietary

gauges are calibrated in pascals but if a home-made gauge is used then the pressure can be calculated from the equation:

$$p = \rho_2 gh \quad (30.6)$$

where p is the pressure, g is the acceleration due to gravity, h the height of the column and ρ_2 is the density of the liquid in the gauge.

Manometers require to be set up exactly level and have problems with loss of fluid and bubbles; however if set up properly they do not require calibration and can be used to calibrate other instruments. More practical instruments for use in the field are preferred. Mechanical devices such as diaphragms linked to an analogue scale suitable for the range of pressures to be measured are useful for most applications. Electrical instruments using pressure transducers are suitable for non-flammable atmospheres. Both these types require regular calibration.

Smoke tracers

In order to identify airstreams outside the boundaries of ducts, a cloud of smoke is useful. Several methods of smoke generation are in use; commercial smoke tube kits are available, producing a cloud of white smoke. These smoke clouds are at the same temperature as the ambient air and would

follow the airflow into which they are introduced, thus the currents can be made visible. It is useful to 'puff' the smoke around the entries to hoods and fume cupboards to identify the flow patterns and speeds of slow-moving air currents. Leaks in joints and cracks in building work can also be tested using smoke clouds. Cigarette smoke is not suitable for this purpose as it is at a higher temperature than the ambient air thus giving a false idea of the flow patterns. In dusty atmospheres the movement of particles can be highlighted using a Tyndall beam apparatus or 'dust lamp'.

Air velocity instruments

Pitot-static tube

This is a device that, in conjunction with a pressure gauge, measures velocity pressure inside a ventilation system, irrespective of the static pressure at the point of measurement (Fig. 30.1). The velocity pressure can simply be converted to a velocity using the formula for velocity given above (Equation 30.2). The measuring head consists of two concentric tubes: one facing into the air stream parallel to the flow, the other with peripheral holes that will be at right angles to the air stream when the facing tube is correctly placed. The tubes are connected via flexible tubing to each side of the manometer or pressure gauge so that velocity

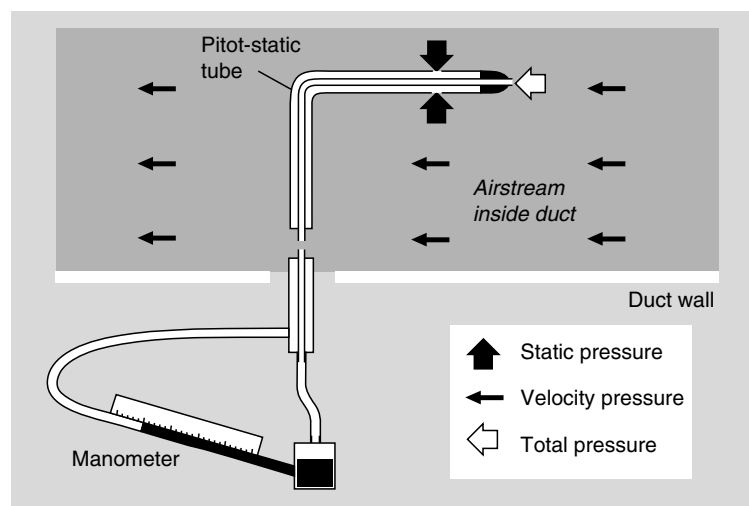


Figure 30.1 Principle of operation of the pitot-static tube.

pressure can be indicated. This device needs no calibration but the airflow measured must be in a section of duct free from obstructions, bends or unnecessary turbulence, i.e. in a straight length of ducting at least 10 duct diameters away from obstructions. As the velocity pressure is proportional to the square of the velocity, the pitot-static tube reading becomes more reliable as the velocity increases but below 3 m s^{-1} they should not be used.

Rotating vane anemometers

These are like small windmills, usually between 25 and 100 mm in diameter and enclosed in an annular shroud. The rotating vanes are either coupled mechanically or electrically to a meter. Those that are mechanically coupled must be used in conjunction with a stopwatch so that the meter can be read over a known period of time, but they require no power source and are ideal for use in flammable atmospheres. The electrically coupled ones read directly in air velocity units but the meters are battery or mains operated and they may not be used in flammable atmospheres unless intrinsically safe.

The size of most vane anemometers makes them unsuitable for use in narrow ducts or extract slots, although the 25-mm-diameter heads can be used in most situations. If the blade settings become altered or the bearings are damaged, due to rough handling or use above the designed velocity, they will give false readings and will require repair and recalibration. They should not be used in very dusty or corrosive atmospheres.

Heated head anemometers

This group of instruments relies on a stream of air to cool a sensitive head consisting of a hot wire, a heated thermocouple or a thermistor. Various electrical means are employed to convert the cooling power, suitably corrected for ambient temperature, to a reading on a meter calibrated in air velocity units. The field version of these instruments consists of a sensing head on the end of a probe with an electrical cable leading to a meter that is usually battery powered. The head is non-

directional unless a cowl is fitted, that is it will sense the airflow in whatever direction it is flowing but the cowl channels it into one direction. The probes and heads are small and can easily be inserted through a small hole in ducting. The air velocities they can measure are in the range of $0.1\text{--}30 \text{ m s}^{-1}$, often in one instrument, but the heads are fragile, susceptible to dust deposits and they require regular calibration. They cannot be used in flammable atmospheres unless they are specifically designed for the purpose.

Calibration

With the exception of the pitot-static tube, all the instruments mentioned above require regular calibration for reliable results to be obtained. This should be done in a wind tunnel in ideal flow conditions, where accurately known air velocities can be produced. Few organizations have such a device, but a small open-jet wind tunnel is available, which gives sufficiently precise air velocities for calibrating the instruments most often used in the field. An open-jet wind tunnel can be purchased at a fraction of the cost of a full one. The calibration can be shown on a chart in the form of a table or graph relating indicated velocity to true velocity. This chart should always accompany the instrument.

Techniques for using air velocity instruments

To obtain reliable results, not only must the instruments be in good condition and recently calibrated, but also the reading must be taken in the correct way, given the circumstances under which they are used.

The sensing head of the air velocity instrument must be placed in the air stream so that its axis is parallel to the streamlines, as most instruments are sensitive to 'yaw', that is the angle between the axis of the device and the direction of flow at the point of measurement. If the instrument is used at a place where the air streams are not parallel, it is very difficult to avoid yaw. Some such examples are given below.

- *At the discharge point from a ventilation system.* Here, the air streams expand in a conical shape, the

boundary angle of which is approximately 20° . If the discharge is a grille or diffuser then the sides may not be axial or parallel and the streamlines will be at an angle greater than 20° .

- *At the inlet to a ventilation system.* Air is being drawn in from all directions from a zone of air movement that is essentially spherical in shape, thus no air stream is parallel to the next.

- *Inside ductwork.* Bends, changes of section, dampers and other obstructions result in the air streams being distorted for some distance downstream.

In order to obtain reliable readings to establish the volume of air flowing in a system with some confidence, a measuring site must be chosen inside the ductwork where the air streams are parallel, as far as can be ascertained. Such a site should be in straight, parallel-sided ducting at least 10 duct diameters downstream of any obstruction or change of direction. Having chosen a suitable site, it will be necessary in most cases to drill a hole in the side of the duct of sufficient size to allow the insertion of the air velocity instrument or sensing head, and a plug should be available to block up the hole after taking the readings.

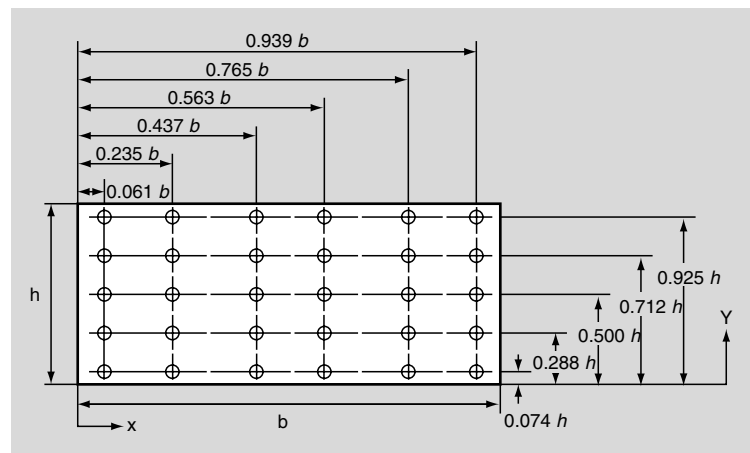
A further complication in the accurate measurement of airflow is the fact that air close to the boundary walls will move more slowly than that in the centre. Therefore, it is necessary to obtain an average velocity from over the whole cross-section

of the duct by measuring in more than one place across the duct. In order to ensure that no bias exists in the choice of places, the measuring area must be divided into imaginary sections of equal area and a representative velocity measured in each section. The British Standards Institution in BS 848 Part 1 (British Standards Institution, 1980) suggests ways of dividing the cross-section of a measuring station for both rectangular- and circular-shaped areas (Figs 30.2 and 30.3). Essentially, the technique involves dividing the area into equal annuli in circular ducts and according to the log-Tchebycheff rule in rectangular ducts. Velocity measurements are taken at each point and their arithmetic mean is calculated. If velocities are measured using the pitot-static tube then it must be remembered that the velocity pressure must be converted to a velocity before taking the arithmetic mean.

Extraction ventilation

One of the most useful methods of controlling the airborne concentration of toxic or nuisance pollutants is to extract them at source before they become dispersed into the workroom environment. Extract ventilation can remove gases, vapours, fumes, dust and larger particulates by means of enclosures, hoods or slots (Fig. 30.4) placed

Figure 30.2 Log-Tchebycheff rule for traverse points in a rectangular duct. To obtain the volume flow rate, the true mean air velocity must be multiplied by the area of the measuring cross-section. Based on BS 848 Part I (British Standards Institution, 1980).



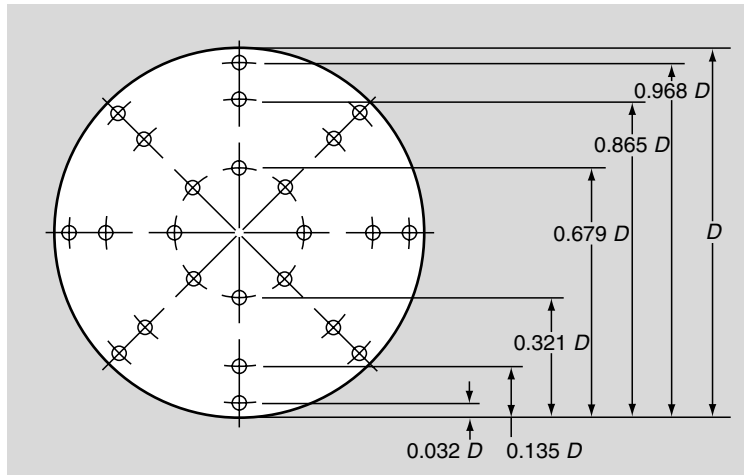


Figure 30.3 Measuring positions for placing pitot-static tubes in rectangular and circular ducting. Based on BS 848 Part I (British Standards Institution, 1980): extracts from BS 848 Part I (1980) are reproduced with the permission of BSI. Complete copies can be obtained by post from BSI Customer Services, 389 Chiswick High Road, London W4 4AL; tel. 0181 996 7000).

around or as close to the point of release of the pollutant as possible. The volumes of air drawn through these devices can be calculated and depend upon the capture velocity of the substance, the distance of the mouth of the extract from the point of release and the dimensions of the device.

The extract points need to be coupled via ducting, and in many cases the air should be conveyed to an air cleaner connected to a point of discharge that may or may not be outside the building. Fans or air movers are used to provide the motive power unless there is sufficient natural ventilation to give an adequate airflow. The design features of these extract devices, together with some simple formulae to calculate the air volume flow rates required to capture the pollutant successfully, are given below. First, however, it is necessary to give some definition of terms used.

- *Capture velocity.* The air velocity required at the source of emission sufficient to cause the pollutant to move towards the mouth of the extract and thus be successfully captured. Table 30.2 gives recommended capture velocities.
- *Face velocity.* The air velocity at the opening of a hood or enclosure.
- *Slot velocity.* The air velocity in slots.
- *Transport velocity.* The minimum air velocity required in all parts of the system, including ductwork and extract devices, to keep collected particles airborne and to prevent them from being

deposited on the sides or floor of the system until the collector is reached. Recommended transport velocities are given in Table 30.3.

- *Plenum.* The enclosed space behind the face of a hood or slot.

Design features and volume flow rates

Enclosures

The commonest forms of these devices are booths and fume cupboards, which contain the process from which the pollutant is released. The mouth of the device provides access for the operator but also provides an escape route for the pollutant. Therefore, at the face of the enclosure, it is necessary to maintain an air velocity that is sufficient to prevent the pollutants from escaping. In the case of chemical fume cupboards, the opening area can be varied by raising or lowering a sliding front.

With all these devices it is advisable to enclose as much of the process as possible and to minimize the area of access compatible with the task being enclosed. In this way the volume of air required can be minimized. If such an enclosure reduces visibility for the operator then toughened glass or Perspex can be used for those parts interfering with vision. It may be possible to cover the opening with a flexible curtain to further reduce the face area. Such a curtain could be made of plastic or canvas sheeting hung in narrow overlapping strips to

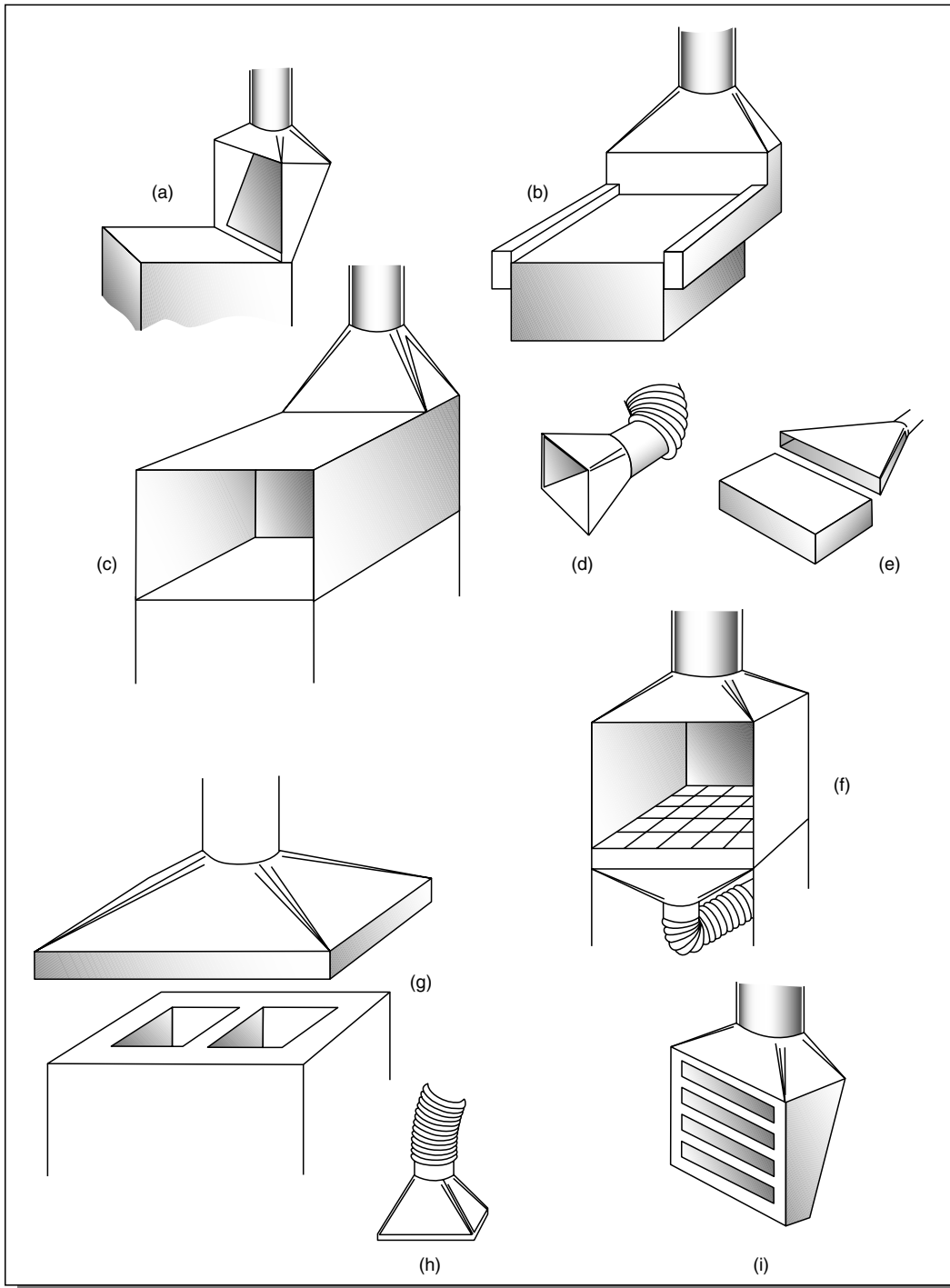


Figure 30.4 Extraction ventilation devices: (a) open face hood; (b) slots on tank; (c) booth or enclosure; (d) hood on flexible duct; (e) slot; (f) supplied and extracted enclosure (push-pull); (g) canopy hood; (h) canopy hood on flexible duct; (i) hood with face slots.

Table 30.2 Recommended capture velocities.

<i>Source conditions</i>	<i>Typical situations</i>	<i>Capture velocity ($m s^{-1}$)</i>
Released into still air with no velocity	Degreasing tanks, paint dipping, still-air drying	0.25–0.5
Released at a low velocity or into a slow-moving air stream	Container filling, spray booths, screening and sieving, plating, pickling, low-speed conveyor transfer points, debugging	0.5–1.0
Released at a moderate velocity or into turbulent air	Paint spraying, normal conveyor transfer points, crushing, barrel filling	1.0–2.5
Released at a high velocity or into a very turbulent air stream	Grinding, fettling, tumbling, abrasive blasting	2.5–10.0

Table 30.3 Recommended transport velocities.

<i>Pollutant</i>	<i>Transport velocity ($m s^{-1}$)</i>
Fumes, such as zinc and aluminium	7–10
Fine dust, such as lint, cotton fly, flour, fine powders	10–12.5
Dusts and powders with low moisture contents, such as cotton dust, jute lint, fine wood shavings, fine rubber dust, plastic dust	12.5–17.5
Normal industrial dust, such as sawdust, grinding dust, food powders, rock dusts, asbestos fibres, silica flour, pottery clay dust, brick and cement dust	17.5–20
Heavy and moist dust, such as lead chippings, moist cement, quick-lime dust, paint spray particles	over 22.5

allow the passage of tools or materials with the minimum of disturbance. The front of the enclosure could also be 'washed' by a gentle vertical air curtain.

In order to improve pollutant control it may be possible to supply the enclosure with air in addition to extracting from it. If this is done it will be necessary to extract at least 15% more air than is supplied to prevent the excess air from escaping through the access opening.

If the enclosure covers moving machinery that is producing the airborne pollutant, such as a grinding wheel or circular saw blade, it is important to take into account the trajectory of any particles released so that the enclosure will capture them. With such an extract, the transport velocity must be maintained in all parts of the system up to the dust collector, including the enclosure itself.

The required volume flow rate is calculated by multiplying the area of the opening by the face velocity necessary to prevent the pollutant from escaping. This velocity depends upon the toxicity

of the substance and the momentum it has as a result of the way it has been released. For example, if the work being enclosed is being sprayed with paint then particles of paint are likely to rebound from the workpiece with considerable velocity, whereas if the workpiece was being dipped in a tank of paint, only the vapours could escape at a low speed. Thus the face air velocity needs to be much higher in the former than in the latter. As a general rule, face velocities should not fall below $0.5 m s^{-1}$ and for the more toxic pollutants or high-momentum particles, face velocities in excess of $1.5 m s^{-1}$ may be required. Thus an opening of $1\text{-}m^2$ area with a face velocity of $1 m s^{-1}$ will require a volume flow rate of $1 m^3 s^{-1}$.

In the case of fume cupboards, the older types had a variable face velocity depending upon the position of the sliding front, such that the wider the opening the lower the velocity and vice versa, but the modern ones have a bypass arrangement or a variable performance fan so that the face velocity remains reasonably constant whatever the position of the front.

Hoods

Often it is not possible to enclose work for reasons of access, for example, workpieces may have to be lowered from above or passed sideways from one process to the next. Therefore, hood ventilation, although less effective than enclosures, may have to be used. Hoods with width-length aspect ratios of less than 0.2 are called *slots* and are dealt with separately.

Hoods can be placed over the work like a canopy or at the side or rear of the workplace, depending upon the accessibility required. It is inadvisable to place canopy hoods over the top of the work if the operators have to lean over into the path of the air as it rises into the hood. In such circumstances, a side or rear hood is more likely to keep the pollutants away from the breathing zone of the worker. When deciding upon the position of a hood it is necessary to take advantage of the natural currents of the substance to be captured. Substances that are hotter than the ambient air temperature or gases that are lighter than air will naturally tend to rise. Therefore, as far as possible, the hood should be sited above the point of release to facilitate capture. Similarly, heavier, cooler substances should be captured from the side. If it is necessary to pull against the natural current of air then a much higher capture velocity must be provided. However, it should be remembered that a concentrated stream of pollutants will obey buoyancy or gravity forces, but if well mixed with air, the mixture will have a density so close to that of air that such forces can be ignored. An example of this concerns vapours from certain cellulose paints, which, in a concentrated state, will be heavier than air but, even at saturation conditions, a mixture of the vapour and air will not be sufficiently different in density from air to have any great effect.

With large hoods, the air velocity distribution across the face can be uneven, being higher close to the duct entrance and lower at the extremities. To overcome this, either the face of the hood can be divided up into a series of slots or the plenum can be divided by guide vanes or air splitters carefully spaced to ensure even distribution.

In order to minimize the air volume flow rate required, the hood should be placed as close as

possible to the point of release of the pollutant and it is worth noting that the required quantity varies with the square of the distance.

Slots

For reasons of accessibility even hoods may be too bulky to allow work to flow satisfactorily and it may be necessary for operators to lean over the work on two sides while the other two sides are required to be free for the movement of materials or to improve visibility. In these circumstances, hoods with aspect ratios of below 0.2, called slots, can be used, the narrow sides being no more than 50 or 75 mm in width.

Slots are commonly used on degreasing tanks, cleaning baths and electroplating tanks to remove the vapours released from their surfaces. If a wide surface is to be ventilated, it is important to have two slots, one on each side, so that the extraction distance is minimized, as experience has shown that it is difficult to pull the air into a slot from a distance of more than 750 mm. Slots are coupled to the extract ductwork via a manifold or a transformation piece, whose position in relation to the slot will influence the distribution of air along the slot. If it is placed at one end and the slot is too long, very little air will be drawn into the opposite end and then air splitters would be required to improve the flow distribution. Slots that are longer than 2 m require to have more than one connection to the ductwork.

Improvement to the capture can be made by boosting the extract by a supply slot on the opposite side. However, care must be taken not to provide too high an air velocity over the surface of a tank in case the liquid in the tank is unnecessarily evaporated.

Prediction of performance

Fletcher method

The required volume flow rate can be calculated from the formula below given by Fletcher (1977), which can be applied to both hoods and slots. The formula relates centreline air velocity to the dis-

tance from the hood and is dependent upon the aspect ratio of the hood. In order to capture the pollutants successfully, the source should be on the centreline.

$$\frac{VA}{Q} = \frac{V}{V_0} = \frac{1}{0.93 + 8.58\alpha^2} \quad (30.6)$$

where

$$\alpha = XA^{-\frac{1}{2}} \left(\frac{W}{L} \right)^{-\beta}$$

and

$$\beta = 0.2(xA^{-\frac{1}{2}})^{-\frac{1}{3}}$$

where Q is the required volume flow rate, x is the distance of the source from the hood along the centreline, A is the area of the hood, L is the length of the hood, W is the width of the hood, V is the centreline velocity at distance x from the hood and would normally be the capture velocity and V_0 is the mean velocity at the face of the hood. To assist in the numerical solution of the formula, a nomogram (Fig. 30.5) is provided, which relates: V/V_0 , $x/A^{\frac{1}{2}}$ and W/L .

Garrison method

The method described by Garrison (1983) can be used for circular hoods, which predicts the centreline velocity at two distances from the end of a

circular duct of diameter d . These distances are: $0.5d$ and $1.0d$, i.e. one-half of a diameter and one diameter from the mouth of the circular duct. Using the same symbols for velocities and distance as above:

$$V = FV_0 \quad (30.7)$$

when $x = 0$, $F = 1$, i.e. $V/V_0 = 1$. Table 30.4 gives values of F for $x = 0.5d$ and $x = 1.0d$ for various profiles of duct end. Note that x is measured from the end of the duct not the edge of the profile.

Capture and transport velocities

Capture velocities required to move the pollutants vary with the type of substance and the speed at which it is being released, and can range from 0.25 to 10 m s⁻¹. For example, the emission of vapour

Table 30.4 Values of F for various duct ends.

Duct end profile (see Fig. 30.6)	F	
	$x = 0.5d$	$x = 1.0d$
Plain	0.26	0.08
Flanged	0.30	0.10
Flared	0.40	0.18
Rounded	0.69	0.33

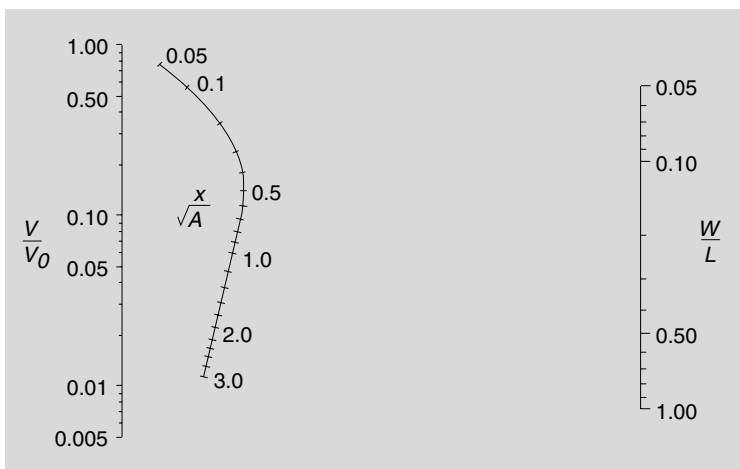


Figure 30.5 Nomogram to assist in the calculation of required volume flow rate. Note: the presence of a flange on the hood can reduce the quantities by 20–25%. Crown copyright is reproduced with the permission of the Controller of HMSO.

from a degreasing tank is of a low kinetic energy and would be successfully captured by an air speed of 0.25 m s^{-1} , whereas a particle emitted from a grinding wheel or a sand-blasting process would have a high kinetic energy and might require a capture velocity of 10 m s^{-1} . Ranges of capture velocities are given in Table 30.2 but it is possible to establish such a velocity for oneself by trial and error using a nozzle attached to a vacuum cleaner.

The transport velocity is somewhat harder to establish and varies with the size and density of the particle (Table 30.3). As with capture velocities, several authorities publish values for different substances based on experience. As a general rule, an air velocity in the duct of over 20 m s^{-1} will keep most particles airborne but bends and obstructions inside the duct can cause local changes of velocity which may result in particles being deposited. Care must be exercised in the design of the ductwork system so that local dust deposits are not allowed to build up and accumulate.

Dilution ventilation

This method of ventilation is less positive than extraction and should only be used when extraction ventilation is not practicable; the pollutants have a low toxicity and the release volume does not fluctuate.

Before calculating the required volume flow rate to dilute the pollutants to an acceptable level, it is necessary to establish their rates of release into the space to be ventilated. This may not be easy as the processes involved may be complex. But if the process is in operation it may be possible to sample the concentrations in known flow rates of air, thus establishing the rate of release of the pollutants. It may be necessary to resort to continuously recording instruments and in this way it is possible to establish whether the release of the substances is constant or fluctuating. In some cases, the estimation of the release rate can be related to the consumption of a particular raw material, for example if a degreasing tank requires to be replenished with a known volume of degreaser in a given time then

it is reasonably safe to assume that the amount has evaporated into the air of the workroom and the rate of release can be calculated.

Required volume flow rate for dilution ventilation

Consider a point source of pollution emitting an airborne substance at a steady rate of r into diluting air flowing at a rate of Q . Assuming thorough mixing, the resulting concentration of pollutant C can be estimated as follows:

$$C = r/Q, \text{ or rearranged as: } Q = r/C \quad (30.8)$$

In order to express Q in the units of cubic metres per second ($\text{m}^3 \text{ s}^{-1}$), when C is in milligrams per cubic metre (mg m^{-3}), r will have to be in milligrams per second (mg s^{-1}).

The value of C will be obtained from the published standards related to the law of the country, e.g. OEL, TLV, PEL, MAK, etc. It is always advisable to set a more stringent standard than that set by law, for example select C as one-quarter of the published standard in order to allow a margin of safety.

The value of r is the most difficult to establish as it has to be based on the amount of pollutant released into the atmosphere, which would depend on the consumption of the source material and its release rate. For example, if dilution ventilation was required to reduce the airborne concentration of painted items being placed on racks to dry then the amount of paint used during the working period together with the volatile proportion of that paint would have to be taken into account.

The mode of introducing the diluting air is important. It can be done in one of three ways:

- 1 allowing the unpolluted diluting air to pass over the worker before reaching the source;
- 2 ensuring thorough mixing of the diluting air and the pollutant before it reaches the worker;
- 3 introducing the diluting air at a temperature slightly lower than the rest of the room and at a very low velocity, such that the clean air displaces the pollutant and moves it upward towards an extract point.

The choice depends upon circumstances and the layout of the workplace.

Ducts and fittings

Air needed to provide satisfactory ventilation, whether it is for supply or extract, is normally carried in ductwork to and from its source. It is unusual to have a direct route for this purpose, therefore bends and changes of section are required to fit the ventilation system to the needs of the building and the associated plant. The shape and size of the ducts and fittings are determined by the ventilation engineer. Each section of duct or fitting absorbs energy from the air to overcome the friction resulting from the passage of air through it. That energy requires a source of motive power for its generation, either a fan or some other air mover. In order to determine the size and duty of the fan and its associated prime mover, the energy loss for each section of the system is calculated and totalled. The following paragraphs describe these calculations in more detail.

Ducting is usually made of galvanized sheet steel, either in circular or rectangular section, although a variety of other materials, including stainless steel, brick, concrete, PVC, canvas, fibreglass and other plastics, are sometimes used.

Duct sizes

There are several factors that should be considered when deciding upon the cross-sectional area of the duct and usually a compromise is made between them. For a given volume flow rate, the larger the duct the lower the air velocity inside and the less energy absorbed, but the larger the capital cost of the material. A circular cross-section is more economical in material than a rectangular one, but in some buildings the space available is more suited to the rectangular shape. If a particular transport velocity has to be maintained, once the required volume flow rate is determined the duct cross-section is fixed as the one that provides that velocity, using Equation 30.4.

Pressure losses

The energy losses due to friction are expressed as a pressure loss. These losses can be determined by calculation or by the use of charts and nomograms. As many published texts reproduce these aids there is no need to duplicate them here, but reference should be made to the Chartered Institute of Building Services Engineers (CIBSE) *Guide Book C* (1986), and the British Hydro-mechanics Association and the British Occupational Hygiene Society (BOHS) publications for further details.

As far as straight lengths of ducting are concerned, pressure losses are quoted per unit length of duct, for example newton metres (N m^{-2}) per metre run of duct or inches of water per 100 ft of duct. This pressure loss is proportional to the air density and the square of the air velocity and inversely proportional to the fifth power of the duct diameter. As the duct sides are parallel, there is no change in air velocity from one end to the other, therefore the pressure losses calculated are both static and total pressures. The losses in most fittings (bends, junctions, etc.) are calculated by multiplying the velocity pressure (p_v) at a point in the fitting by a factor determined empirically for the geometric shape of that fitting; the resulting pressure loss is in total pressure. It is important to work in total pressures for ventilation calculations, as fittings such as expansion pieces have changes in velocity pressure from one end to the other, resulting in gains in static pressure while still sustaining a loss in total pressure; working in total pressure throughout avoids any confusion.

Once the pressure loss for each section of the ductwork system has been calculated, including any dust collection device, all the total pressures should be added together to determine the overall total pressure loss of the system. Thus the duty of the fan required is computed, i.e. the calculated volume flow rate at the total pressure loss. The velocity pressure of any discharge velocity must be added to the total, as it represents energy that the fan has to provide. If this value is forgotten or ignored it could result in the fan being undersized for the duty it is expected to perform.

A complication arises if more than one off-take is connected to the system, for example when several extract or supply fittings are connected to one duct leading to the fan. In such cases the fan total pressure required is only that required to move the air between the two extremities of the system, for example between the fan discharge point and the extract branch requiring the highest pressure, usually the one furthest from the fan, known as the ‘index branch’. Normally, there will be sufficient pressure to cater for intermediate branches.

Suggested method of duct sizing and loss calculation

It is first necessary to draw a sketch of the layout of the system in which the volume flow rates for each terminal should be marked. In order to identify each section of the system, a number should be assigned either to every section or to every junction where the air changes speed or direction. Labelling the junctions rather than the sections facilitates network analysis, particularly if analogue or digital computers are used for calculations; programs are available for this work. In order to simplify the calculations, they should be approached by drawing up a table with the headings as shown in Table 30.5. For each section the appropriate column only need be filled in and the pressure losses calculated. The cumulative pressure loss column enables a running total pressure loss to be maintained from the furthest outlet to the fan so that at any point in the system it is possible to see what the pressure difference is between that point and atmospheric pressure.

With the solution of the ductwork problem, it is necessary to make three decisions that will influence the design.

1 Of what materials should the ducting be made?

Unless corrosive gases or vapours are to be carried, galvanized sheet steel is the normal material for reasons of strength and ease of manufacture of the fittings. Corrosive gases require stainless steel or plastic ducting. Nevertheless, whatever materials are chosen the calculations are carried out as if it was steel and a correction factor is applied depending upon the surface roughness of any other material.

2 What velocity should be chosen for the air in the duct? If there are particles to be carried, the velocity chosen should be the transport velocity for those particles (see Table 30.3), but as transport velocities are high, noise levels from the passage of air in the duct will also be high. If a quiet area is to be ventilated, the air speeds should be below 5 m s^{-1} , but in many industrial settings other noises will prevail and higher velocities can be used.

3 Should the ducting be circular or rectangular in cross-section? If space is limited then a rectangular section may make better use of what space there is, but if high air velocities are to be carried then circular ducting is better for rigidity and has a lower pressure loss for a given cross-sectional area. An example of a typical duct sizing calculation is given by Gill (1995).

Balancing of multibranching systems

If a system has been designed with damper control of intermediate branches then an initial balance will have to be established, and if an existing system is thrown out of balance by injudicious alteration of dampers, rebalancing is required. This task must be undertaken systematically because it should be remembered that the movement of any one damper will alter the airflow rate to

Table 30.5 Headings for table to be used in designing an extract system.

Section	Length (m)	Volume			Air velocity (m s ⁻¹)	Velocity pressure (Pa)	Pressure		
		flow rate (m ³ s ⁻¹)	Duct dimension (mm)	Duct area (m ²)			loss per metre (Pa m ⁻¹)	Section pressure loss (Pa)	Cumulative pressure loss (Pa)

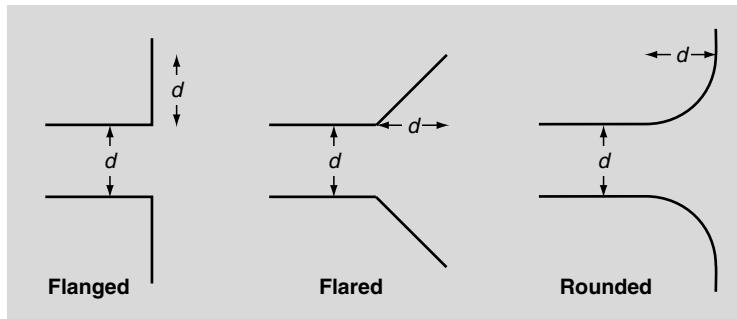


Figure 30.6 Shapes for ends of round ducts.

every other branch. The correct procedure to follow is laid down in the *CIBSE Commissioning Code: Series A, Air Distribution* (CIBSE, 1971).

Fans

A fan consists of a rotating impeller on a shaft to which the power source is connected and a casing that guides the air through the impeller. A pressure difference is created between the inlet and outlet by the rotation of the impeller.

The Fan Manufacturers Association have agreed on the following definitions for fan pressure.

- *Fan total pressure.* The total pressure at the fan outlet minus the total pressure at the fan inlet.
- *Fan velocity pressure.* The velocity pressure based upon the mean air velocity at the fan outlet.
- *Fan static pressure.* The fan total pressure minus the fan velocity pressure.

Fan power and efficiency

Power is the work done on air to move it against a particular pressure. In dealing with fans there are two components of power: the theoretical power required to move air (*air power*) and the actual power required to be provided at the shaft to achieve the air power (*shaft power*). The reason for these two is that fans are not 100% efficient and some power is lost in the turbulence between the blades, the friction of the air as it passes through the casing and the losses in the bearings that support the rotating parts. Air power (P_a) can be calculated from the expression:

$$P_a = Q \times p \quad (30.9)$$

where Q is the volume flow rate and p is either the fan's static or total pressure. If Q is in cubic metres per second and p is in newtons per square metre then the power is in newton metres per second, which is *watts*. The resulting units are known as 'air power (static)' if static pressure is used and 'air power (total)' if total pressure is used.

Fan efficiency is obtained from the measurement of the input power to the shaft of the fan using the expression:

$$\text{Fan efficiency} = \text{air power/shaft power} \times 100\% \quad (30.10)$$

As above, if air power (static) is used then the efficiency is known as 'fan static efficiency' and if air power (total) is used then the efficiency is known as 'fan total efficiency'. The efficiency of a fan varies with its duty and is given by the fan manufacturer, if requested, for a particular performance.

Fan characteristic curves

If a fan is run at a constant speed and its volume flow rate is altered by varying the resistance against which it has to operate, curves showing the variation of pressure, power and efficiency can be plotted against a base of volume flow. These curves, known as 'fan characteristic curves', give the performance of the fan over the whole range of resistances for which it is designed. Manufacturers' catalogues quote these curves either as graphs or tables.

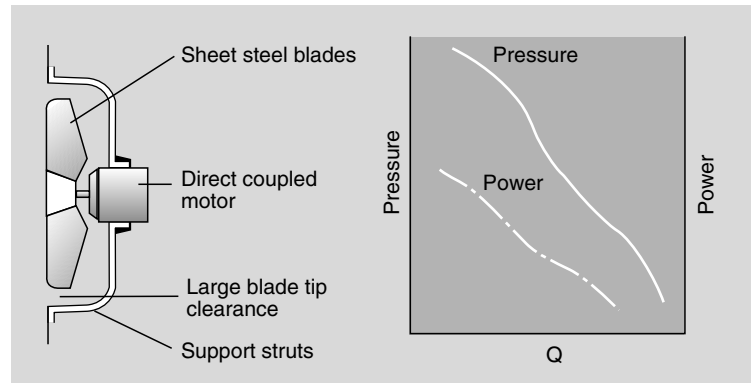


Figure 30.7 Propeller fan and graph showing its characteristic curves.

Typical characteristic curves are shown in Figs 30.7–30.9. Their shape depends upon the geometric design of the fan.

Fan types

Some of the more common types of fan are described below.

Propeller fan

This type of fan (Fig. 30.7), which is a type of axial flow fan, is useful for general ventilation work when there is little resistance to airflow. It is not suitable for use on ductwork or air filtration but is most commonly used for unit heaters, air cooling on car radiators and refrigeration evaporators, and is often seen mounted on an aperture in an outside

wall to provide general supply and extract ventilation for rooms, workshops and warehouses. Propeller fans are generally low in efficiency, but as most are low in power this is no great problem. Aerofoil blades fitted to some of the larger ones improve their efficiency from 55% for the sheet steel blades up to 70% for the aerofoil.

Axial flow fan

This type of fan (Fig. 30.8) has a casing that is cylindrical, the shaft of the impeller being at the centre of the casing and running parallel to the sides. The impeller blades are usually of aerofoil section and rotate with their tips close to the casing. Although similar in principle to the propeller fan, the axial fan produces much higher pressures, up to 1100 N m^{-2} (Pa) per stage, although some single-stage fans can produce even higher pres-

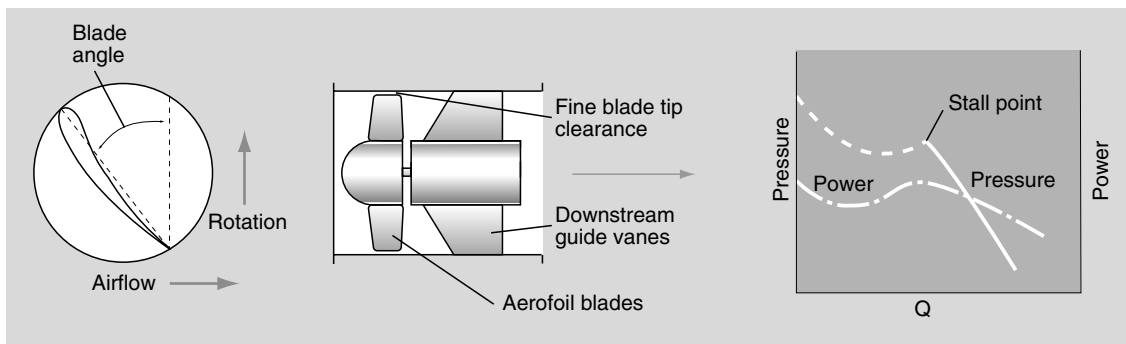


Figure 30.8 Axial flow fan and graph showing its characteristic curves.

tures. If a higher pressure is required, a second stage is fitted or a centrifugal fan is used.

The performance of the fan can be changed by altering the angle of the blades, as shown in Fig. 30.8. With some axials it is possible to unclamp the blades to reset them at another angle, whereas others can be altered while in motion.

The rotation of the impeller imparts a swirl to the air that is undesirable from a performance point of view and so some form of flow straightener is required to help recover some of the rotational energy. If the fan is to be used close to a finned heat exchanger this is not a problem as the fins act as a flow straightener, but if the fan is to be used on ducting it is advisable to use one fitted with guide vanes. These guide vanes take the form of static blades attached to the casing at an angle sufficient to give a counter-swirl to the air, thus it leaves the fan moving parallel to the casing. If a two-stage fan is necessary, the second stage can be arranged to rotate in the opposite direction to the first to counteract the swirl, thus avoiding the need for static guide vanes. Some axial flow fans available have this facility.

The power efficiency of axial flow fans can be as high as 78%, they are compact and they can fit neatly into a ductwork system, often at the same diameter as the ducting. If the drive motor is directly coupled to the impeller in the air stream, the fan takes up no more space than a short length of ducting.

However, an axial flow fan has several disadvantages, one of which is the high noise levels that are associated with the higher speed, higher pressure fans. Silencers are often necessary to keep noise to an acceptable standard. Another disadvantage is the problem of 'stall' that occurs if the fan is expected to work against a system that has a higher resistance than one for which the fan was designed. Stall condition in a fan is similar to hydraulic cavitation on a water pump. The symptoms of stall in a fan are: fluctuating fan pressure and power, change in note and variable flow rate. It is unwise to run a fan in stall condition as the bearings can become damaged. Reducing the blade angle of the fan can move it out of stall condition but the overall performance will be reduced as a result.

It is also unwise to use an axial flow fan with the motor in an air stream where the air is at a high temperature or containing dust or corrosive chemicals. Fans with alternative layouts that put the motor out of the air stream, for example bifurcated, are available and should be used in these situations.

Centrifugal fan

With this type of fan (Fig. 30.9), air enters the eye of an impeller via a suction sleeve. A rotating impeller, which is not unlike a paddle-wheel, throws air to its periphery, where it is collected by a casing in the shape of a volute. The blades of

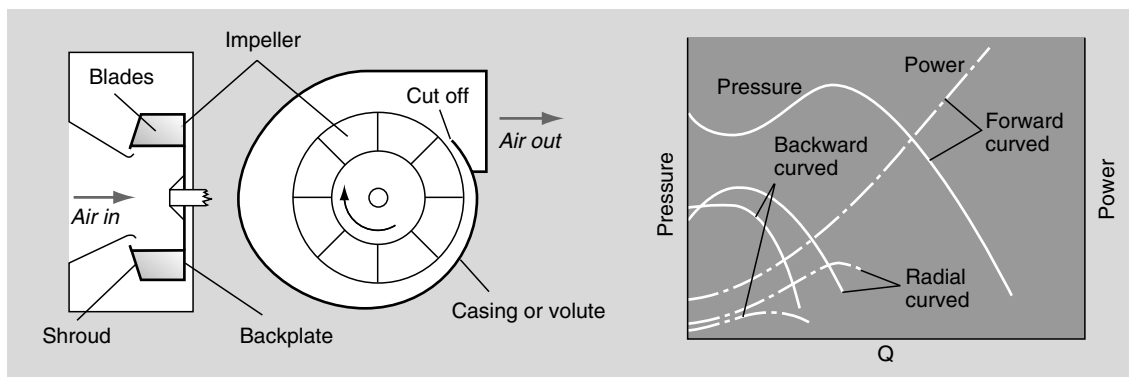


Figure 30.9 Centrifugal fan and graph showing its characteristic curves.

the impeller can be either forward-inclined, backward-inclined or radial in relation to the direction of rotation. Each configuration results in a fan with its own idiosyncrasies. All centrifugals can produce high pressures.

The forward-bladed fan (Fig. 30.10a) has an impeller with many closely spaced but narrow blades that produce a distinctive pressure characteristic, in that at a certain range of pressures three different flow rates are possible, dependent upon the resistance of the system against which it is working. The power characteristic is known as 'overloading' at high flow rates because the curve continues to rise as the volume flow is increased. This type of fan is the most compact of the centrifugals for a given volume flow rate and is used where space is limited. By reason of its pressure and power characteristics, the forward-bladed fan should never be run without being coupled to a ventilation system, nor should two be coupled together in parallel unless under the guidance of the manufacturer or an experienced fan engineer. Forward-bladed fans are less power efficient than the backward-bladed models.

Backward-bladed fans (Fig. 30.10b) have fewer, deeper blades than the forward-bladed models but are less compact for a given duty. The power characteristic is 'non-overloading' and this family of fans is the most power efficient. If aerofoil section blades are used, efficiencies exceeding 85% are possible, particularly on the larger fans. This

makes them popular for the high-powered, continuously running operations such as those used in power stations and deep mines.

Radial-bladed fans (Fig. 30.10c) have similar characteristics to the backward-bladed ones but are less power efficient. The most common of this type of fan is the paddle-bladed model in which the impeller is made of a number of crude, flat plates riveted or bolted to a central hub. It is used for applications where highly corrosive or abrasive air is handled such that the blades may wear or rot away. Replacement blades can be easily fitted to such an impeller that is inherently self-cleaning.

Matching of fan and system

The most efficient part of the pressure characteristic of a fan is just to the right of the peak pressure on the top of the final downward slope (Fig. 30.11). A fan should be chosen so that the duty required lies on that part of the curve. Having calculated the pressure loss of a ventilation system for a given volume flow rate, a point can be plotted and a curve drawn on linear graph paper, which represents the system. The curve will be of the form:

$$p \propto Q^2 \quad (30.11)$$

that is a parabola and will pass through the origin. If the fan total pressure characteristic is plotted on the same scale, the point of intersection of the two

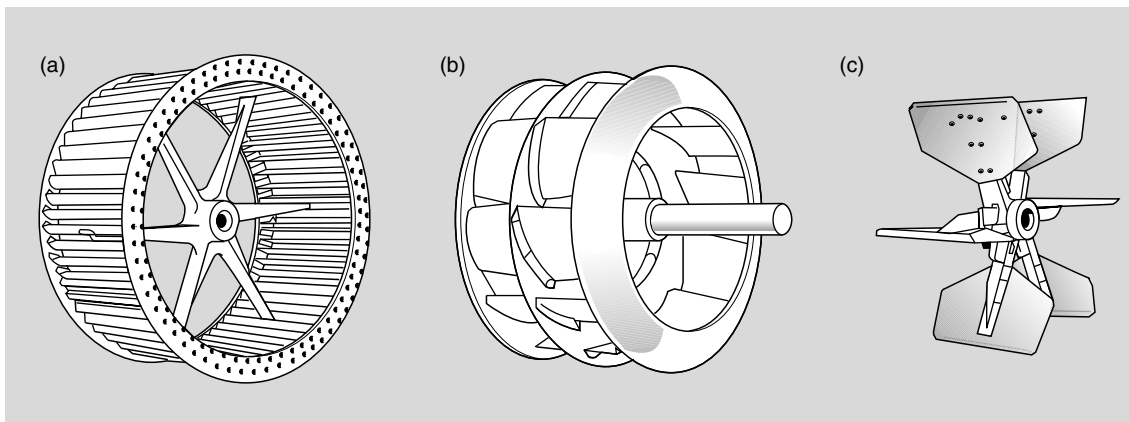


Figure 30.10 Centrifugal fan impellers: (a) forward-bladed; (b) backward-bladed; (c) radial-bladed.

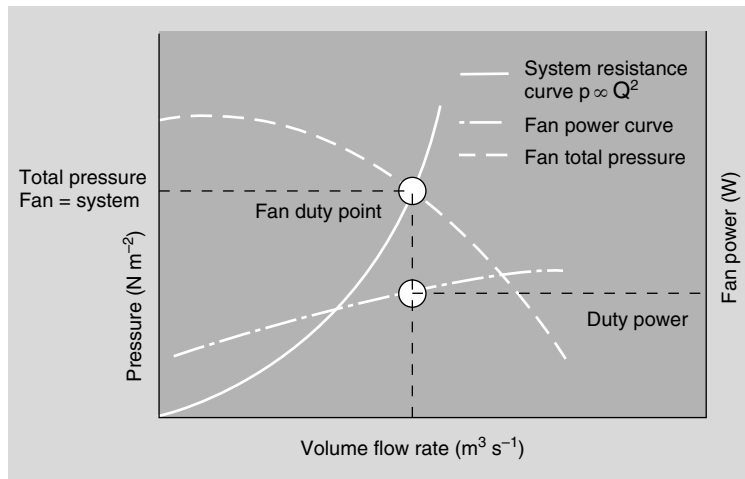


Figure 30.11 Matching of fan and system.

curves will give the duty point that would occur if that fan was installed on that system.

Air cleaning and discharge to atmosphere

It is a statutory infringement in many countries to discharge air into the atmosphere without first rendering it free from pollutants as far as is reasonably practicable. Much of the air that is extracted from workplaces requires some form of cleaning before it is discharged. There are many excellent texts on the engineering control of atmospheric pollution, which include the techniques available for removing substances from ventilation air. Therefore, no details will be given here.

If dry particulates are to be removed from the dry air then dry dust collectors of various types are available. The larger particles can be removed by dry centrifugal methods such as cyclones, whereas bag filters will remove the smaller dusts. Particles that can be easily electrically charged can be collected by electrostatic collectors, provided that there is no fire or explosion hazard in the air or the collected material. Dust extracted from humid air or wet or sticky particles must be collected by some wet method such as venturi scrubbers, wet centrifugals or wet orifice collectors. Unfortunately, these devices collect dust as a sludge that needs disposal.

Air-containing gaseous chemicals should be cleaned by absorption or chemical scrubbing according to the nature of the substance. Whether the air is cleaned or not, it must be discharged into the atmosphere in such a way that it does not re-enter the building or any other building in the vicinity before it has been diluted to negligible concentrations. The best techniques involve discharging the air as high into the atmosphere as possible and at a high velocity. Devices such as cowls and weather caps hinder the upward throw of discharged air and are not recommended. Care should be taken to allow for local wind effects, including turbulence caused by adjacent buildings and any risk of 'blow back' through the ductwork should be prevented.

Make up air

Extract ventilation systems can remove very large quantities of air from buildings. This air will be replaced from outside which, if not specifically supplied, will find its way in through openings in doors and windows, cracks in the walls and roof and anywhere else that is open to the atmosphere. At certain times of the year, particularly in winter in extreme northern and southern latitudes, this will result in cold draughts and unsatisfactory working conditions with regard to temperature.

Ventilation systems can cause cold draughts resulting in non-use. Whenever an extract ventilation system is installed, a make up system should be provided where the volume flow of air supplied is equal to that extracted. The supply should be tempered to the correct conditions to suit the comfort of the building occupants.

The provision of adequate heating or cooling is expensive in both capital and fuel costs. The following simple formula will enable a power calculation to be made to estimate the costs of heating or cooling:

$$\text{Heat required} = mc_p(t_i - t_o) \quad (30.12)$$

where m is the mass flow rate of the air, c_p is the specific heat of air (1.02 kJ kg^{-1}) and t_i and t_o are the temperatures inside and outside the building. From this formula it can be seen that to heat $1 \text{ m}^3 \text{ s}^{-1}$ of air from 0°C to 21°C requires 25.7 kW of energy:

$$\text{Heat required} = 1 \times 1.2 \times 1.02 \times 21 = 25.7 \text{ kW} \quad (30.13)$$

References

- British Standards Institution (1980). *Fans for General Purposes. Part 1: Methods of Testing Performance*. British Standards Institution BS 848 Part 1.
- Chartered Institute of Building Services Engineers (1971). *CIBSE Commissioning Code: Series A, Air Distribution*. CIBSE, London.
- Fletcher, B. (1977). Centreline velocity characteristics of rectangular unflanged hoods and slots under suction. *Annals of Occupational Hygiene*, **20**, 141–6.
- Garrison, R.P. (1983). Velocity calculation for local exhaust inlets – empirical design equations. *American Industrial Hygiene Association Journal*, **44**, 937–40.
- Gill, F.S. (1995) Ventilation. In *Occupational Hygiene*, 2nd edn (eds J.M. Harrington and K. Gardiner), pp. 395–7. Blackwell Scientific Publications, Oxford.
- American Conference of Governmental Industrial Hygienists (2002). *Industrial Ventilation. A Manual of Recommended Practice*, 23rd edn. ACGIH, Cincinnati.
- Ashton, I. and Gill, F.S. (2000). *Monitoring for Health Hazards at Work*, 3rd edn. Blackwell Scientific Publications, Oxford.
- British Occupational Hygiene Society (1987). *Controlling Airborne Contaminants in the Workplace*. BOHS Technical Guide No. 7. Science Reviews Ltd, Leeds.
- Burgess, W.A., Ellenbecker, M.J. and Treitman, R.D. (1989). *Ventilation for the Control of the Work Environment*. John Wiley, Chichester.
- Chartered Institute of Building Services Engineers (1986). *Guide Book C*. CIBSE, London.
- Fletcher, B. (1978). Effect of flanges on the velocity in front of exhaust ventilation hoods. *Annals of Occupational Hygiene*, **21**, 265–9.
- Fletcher, B. and Johnson, A.E. (1982). Velocity profiles around hoods and slots and the effects of an adjacent plane. *Annals of Occupational Hygiene*, **25**, 365–72.
- Gill, F. (1999). Prevention and control of exposures. In *Occupational Health Risk Assessment and Management* (eds S.S. Sadhra and K. G. Rampal), pp. 197–218. Blackwell Science, Oxford.
- Harrington, J.M., Gill, F.S., Aw, T.C. and Gardiner, K. (1998). *Occupational Health, Pocket Consultant*, 4th edn. Blackwell Scientific Publications, Oxford.
- Miller, D.S. (1978). *Internal Flow Systems*. British Hydro-mechanics Research Association, Cranfield.
- Woods of Colchester Ltd (1978). *Woods Practical Guide to Fan Engineering*. Woods, Colchester.

Chapter 31

Personal protective equipment

Robin M. Howie

Setting up an effective personal protective equipment programme

Assess risks and identify where control is required

Implement all reasonably practicable controls

Identify who needs residual protection

Inform wearers of consequences of exposure

Select personal protective equipment that is adequate to control residual exposure

Involve wearers in personal protective equipment selection process

Match personal protective equipment to each individual wearer

Carry out fit tests of respiratory protective equipment

Ensure that personal protective equipment does not create risk(s)

Ensure personal protective equipment is mutually compatible

Train wearers in the correct use of their personal protective equipment

Minimize wear periods

Supervise wearers to ensure correct use of personal protective equipment

Maintain personal protective equipment in an efficient and hygienic condition

Inspect personal protective equipment to ensure it is correctly maintained

Provide suitable storage facilities for personal protective equipment

Record usage, maintenance and inspection data

Monitor programme to ensure continuing effectiveness

Provide personal protective equipment free of charge

Personal protective equipment standards

The reality of personal protective equipment performance in the workplace

Respiratory protective equipment

Personal hearing defenders

Protective clothing

General conclusions and recommendations

References

The objectives of this chapter are to outline the role of personal protective equipment (PPE) in a risk control programme and to provide advice on setting up and implementing a PPE programme to ensure that PPE wearers can achieve the required levels of protection without being exposed to either unacceptable stress or discomfort.

PPE, such as protective clothing, respiratory protective equipment (RPE) or personal hearing defenders (PHD), is very widely used. This is primarily because PPE is perceived to provide effective and relatively inexpensive protection, whereas alternative techniques, such as substitution, segregation or other means of control (see Chapter 29) either cannot be applied or are perceived to be expensive. In addition, many employers believe that they can delegate their legal responsibility to protect their employees' health to the employees themselves by providing PPE.

It is a fundamental principle of occupational hygiene practice that all possible action should be taken to prevent or reduce risks at source rather than relying on PPE to provide the only or main protection for the wearer. This is because PPE reduces exposure only for the individual wearer and because competent hygienists have recognized for many years that PPE performance in the workplace is difficult to guarantee and is invariably lower than predicted by either standard laboratory tests or simulated workplace tests. This principle is enshrined in Article G of the European 'Framework' Directive, which requires that the employer gives 'collective protective measures priority over individual protective measures' (Commission of the European Communities, 1989a), and is reiterated in Article 3 of the PPE 'Use' Directive (Commission of the European Communities, 1989b). These duties specified in the Commission of the

European Communities (CEC) Directives should be incorporated into the national legislation of all Member States of the European Union.

Notwithstanding the above, there will be work situations where alternative means of adequately controlling risks are not technically possible and the use of PPE is unavoidable, for example during incidents such as chemical spillages, between recognizing a risk situation and implementing prevention or control, during routine maintenance operations, to supplement inadequate controls or to provide additional security in case primary control measures should fail.

When PPE has to be used it must provide adequate protection without itself leading to any risk or increased risk and without imposing unacceptable discomfort or encumbrance on the wearer. To achieve this, PPE must be selected taking into account the nature and extent of the hazards and risks, the wearers' individual characteristics, their jobs and working environments and any other items of PPE that may have to be worn simultaneously.

To ensure that PPE provides adequate ongoing protection, it is not sufficient to simply provide nominally adequate equipment; it is essential to set up a comprehensive programme to ensure that the PPE can continue to be correctly used and, if reusable, is maintained in an efficient and hygienic condition.

Setting up an effective personal protective equipment programme

The steps required to run and establish an effective programme covering PPE such as RPE, PHD or protective clothing etc. are summarized below:

- assess risks and identify where prevention or control is required;
- implement all reasonably practicable controls;
- identify who needs residual protection;
- inform wearers of consequences of exposure;
- select PPE adequate to control residual exposure;
- involve wearers in PPE selection process;
- match PPE to each individual wearer;
- carry out fit tests of RPE;
- ensure PPE does not create risk(s);

- ensure PPE is mutually compatible;
- train wearers in the correct use of their PPE;
- minimize wear periods;
- supervise wearers to ensure correct use of PPE;
- maintain PPE in an efficient and hygienic condition;
- inspect PPE to ensure it is correctly maintained;
- provide suitable storage facilities for PPE;
- record usage, maintenance and inspection data;
- monitor programme to ensure continuing effectiveness;
- provide PPE free of charge.

The above subheadings are described in greater detail below.

Assess risks and identify where control is required

The essential first step to achieve effective management of safety and health in any workplace is an assessment to identify any likely occupational hazards and to quantify any risks (see Chapters 15–26). The assessment should identify all unacceptable risks and the individuals at risk and provide the information needed for designing collective and administrative means of prevention or control and for selecting adequate and suitable PPE (see Chapter 29).

Implement all reasonably practicable controls

If hazardous substances or processes must be used, all reasonably practicable means of reduction of risk at source must be considered before adopting PPE, for example to enclose any process involving hazardous substances (see Chapter 29) or to apply local exhaust ventilation to such processes, etc. (see Chapter 30).

Identify who needs residual protection

From the assessment of likely risks and of the effectiveness of any collective or administrative procedures applied to reduce these risks, all persons still potentially at risk should be identified and the level of residual protection still required should be quantified.

Inform wearers of consequences of exposure

To ensure that all workers fully utilize all control measures, they should be made fully aware of the risks to their health and safety in the workplace and the potential consequences if these risks are not adequately controlled. If the correct use of control measures involves inconvenience or reduction in productivity, particularly for those on piecework, control measures may not be correctly used unless those at risk perceive some benefit to themselves. Similarly, as many types of PPE are inherently uncomfortable, some wearers may refuse to wear such equipment unless they are convinced that the imposed discomfort can be justified in terms of reduced risk to themselves. To ensure that wearers are aware of the benefit of wearing the PPE provided, it is important to ensure that all persons exposed to risk in the workplace have a perception of the risk or risks to which they are exposed.

Select personal protective equipment that is adequate to control residual exposure

PPE should be selected that reduces any risks to acceptable levels. Given the general downward trend in OEL and the increasing number of substances assigned a Maximum Exposure Limit (MEL) rather than an Occupational Exposure Standard (OES), it is considered prudent to limit personal exposures to levels lower than the OEL unless such reduction is not required to protect health, e.g. in the case of a substance assigned an OES, it is considered prudent to reduce personal exposure to not more than 25% of the OES and in the case of a substance assigned an MEL or a Control Limit, to not more than 10% of the MEL or Control Limit. PPE should be selected on the basis of demonstrated workplace performance. If a manufacturer cannot supply workplace data or written assurance as to the level of performance that can realistically be achieved in the workplace, his PPE should not be used.

Involve wearers in personal protective equipment selection process

Many types of PPE impose some level of discomfort on the wearer and as the level of discomfort may reflect the degree to which the PPE and the wearer's body have to mutually deform to achieve adequate fit, wearers should be fully involved in the PPE selection process to ensure that the equipment selected does not impose unacceptable discomfort. On a simple psychological basis, involving wearers in the selection procedure and in all aspects of the PPE programme gives them a stake in ensuring the programme's overall effectiveness. The importance of such involvement is recognized in Article 8 of the 'Use' Directive (Commission of the European Communities, 1989b), which requires the 'consultation and participation of workers and/or their representatives'.

Match personal protective equipment to each individual wearer

Even in a small workforce there may be large differences in the physical characteristics of PPE wearers; consequently, PPE that fits one person may not fit another. For example, a respirator that correctly fits a large male might not correctly fit a petite female. Although no one would consider it sensible to buy safety shoes in only one size, many users appear to consider that one size of respirator, earmuff or gloves can correctly fit all wearers.

Given the range of differences in body sizes in a typical workforce, the employer should ensure that the equipment provided adequately fits all of those required to wear them. It might thus be necessary to provide a range of equipment of different shapes and sizes to ensure that all potential wearers can be adequately protected. When the potential wearers include persons with non-European characteristics, the supplier should be requested to provide test data demonstrating that the PPE provides such wearers with adequate protection.

For PPE that imposes significant stress on the wearer, for example heavy chemical protective suits that can cause heat strain or breathing appar-

atus (BA) that may weigh up to 18 kg or which requires the wearer to breathe pure oxygen, it may be necessary to apply tests to ensure that only medically fit and suitable persons will be required to wear the equipment.

Carry out objective fit tests of respiratory protective equipment

Current UK guidance requires that each RPE wearer carries out fit tests to ensure that any RPE provided adequately fits his other face. Guidance on how to carry out such tests is given in HSE (2003). However, it must be stressed that such tests identify only gross misfits between the RPE and the wearer (BSI, 2001). The results from such tests should not be taken as any evidence that the RPE fits the wearer or as any indication of likely protection in workplace (see Howie, 2000).

Ensure that personal protective equipment does not create risk(s)

Some types of PPE can create or exacerbate risks. For example, eye protectors or full-facepiece respirators can affect the wearer's field of vision, PHD can reduce the ability to hear communication or warning signals or the approach of vehicles and chemical protective clothing can reduce the body's ability to lose metabolic heat, etc. If the required PPE encloses a significant proportion of the area of the body, the possibility of the wearer being at risk due to thermal strain should be addressed and action taken if harmful levels of thermal strain are likely (see Chapter 20).

If wearers perceive that the consequences of such potential risks are more severe than the risks against which the PPE is provided to protect, they may refuse to wear the PPE. For example, coalminers traditionally refused to wear PHD because they needed to hear the 'strata creak'. The risk from roof falls was thus perceived to be more serious than the risk from noise.

Care should therefore be taken to ensure that any risks likely to be created by the equipment are fully investigated and reduced to acceptable levels.

Where some residual risk remains, action should be taken to minimize the risk created by the PPE at source. For example, if PHD are used when moving vehicles may be present, the vehicles should be fitted with flashing warning lights so that realization of the approach of the vehicles does not rely on aural warning.

For PPE intended to prevent ocular or percutaneous problems, the manufacturer should be able to provide guidance on maximum wear periods for foreseeable conditions of use (CEC, 1989c). The user should therefore require the potential supplier of such PPE to provide the relevant information. Equipment should only be purchased from manufacturers able to supply such information. If the supplier cannot provide this information, the supplier should be reported to the local trading standards officer, who has the duty to enforce the requirements of the 'Product' Directive.

Ensure personal protective equipment is mutually compatible

Where two or more types of PPE must be worn simultaneously, the different types of PPE can interact to reduce the protection provided by one or both items. For example, for safety helmets worn with full-facepiece respirators, the facepiece can force the safety helmet to tip backwards, and the front head-harness buckle of some facepieces can reduce the space between the forehead and the helmet, so that an impact on the front of the helmet can force the buckle into the forehead. Other examples of potentially adverse interactions are respirators worn together with eye protectors when the respirator can prevent the correct fitting of the eye protectors, or eye protectors worn with earmuffs when the legs of the eye protectors can prevent a good seal between the head and the muff. In such situations it is essential that each item of PPE provides the required level of protection without affecting the effectiveness of any other PPE. Care must therefore be taken when selecting PPE for such situations to ensure that all PPE that may have to be worn together are mutually compatible.

Although it is important to select PPE that does not individually cause unacceptable discomfort, it should be appreciated that wearing two or more types of PPE together can result in previously 'comfortable' equipment being considered 'uncomfortable'. Care must therefore be taken to ensure that the overall ensemble is acceptable to the wearers. The user should therefore seek guidance from the supplier, for example ask what type of safety helmet can be worn together with his respirator. Although it may be difficult for individual users to ensure mutual compatibility when the various items of PPE are manufactured by different companies, it may be simpler to buy equipment made by the same manufacturer, as a manufacturer who produces several types of PPE is required by Annex 11 of the PPE 'Product' Directive to be able to supply information regarding their mutual compatibility (CEC, 1989c).

If suppliers are unable to provide the required information, they should be reported to the relevant enforcing authority, i.e. in the UK, the local trading standards office.

Train wearers in the correct use of their personal protective equipment

Wearers and their supervisors should be thoroughly trained in how to fit the PPE correctly, how to assess whether the equipment is correctly fitted, how to inspect the PPE to ensure that it has been correctly manufactured and, for reusable equipment, whether it has been adequately cleaned and maintained.

The potential consequences of PPE failure should be reflected in the thoroughness of the training, i.e. the greater the risk, the more thorough the training. Training should aim to convince wearers that the equipment provided will protect; few people are likely to wear uncomfortable PPE unless they are satisfied that it will protect. Generating such conviction is generally most easily achieved by practical training sessions that allow the wearer to assess how well the PPE can perform when fitted and worn correctly. Where possible, the training should involve practical sessions using suitable training tools, for example for RPE using

quantitative fit tests to identify those wearers unable to fit the given RPE. However, as noted above, such tests do not identify either that the RPE can fit an individual wearer or the level of protection likely to be achieved in the workplace.

Training should cover the correct use of PPE ensembles; for RPE and protective clothing ensembles, for example, to fit the RPE first, not to wear the RPE facepiece or head harness over the hood of protective clothing and to remove RPE last after completing all required decontamination procedures. Wearers of RPE should be aware that correct fit may depend on there being no facial hair lying between the facepiece and their face and that stubble can be even more damaging than a beard. They should therefore be aware that being 'clean-shaven' means having shaved immediately before the start of each shift during which RPE might have to be worn and shaving again during the shift if they have heavy beard growth.

Supervisors should be trained in how to ensure that wearers are likely to be able to fit their PPE correctly, for example to ensure that RPE wearers are clean-shaven or that wearers of PHD do not have hairstyles that might prevent earmuffs sealing adequately to the sides of the head. Maintenance and inspection staff should be trained in the relevant procedures, with particular emphasis being given to ensuring that they are able to carry out such tasks without placing themselves at risk due to contamination on the PPE, for example when cleaning PPE that may be contaminated with asbestos or isocyanates. For highly complex PPE that may be used in situations of acute risk, for example breathing apparatus, manufacturers often offer training courses for wearers and for maintenance and inspection staff.

Minimize wear periods

Many types of PPE are inherently uncomfortable. Since the acceptability of a given level of discomfort can decrease with increasing wear time, it can be important to reduce wear times as far as possible. If it is possible to identify those processes that generate most of the risk, it may be possible to achieve adequate reduction of exposure by wear-

ing the PPE only during such processes. That is, the imposed discomfort could be reduced, and the likelihood of the equipment being correctly worn increased, if wearers were able to remove their PPE when these processes are not being carried out. Care would have to be taken to ensure that contamination of the wearer or PPE in such circumstances does not itself constitute a risk. The risk assessment noted above should therefore include identifying whether it is possible to reduce wear durations by identifying periods when the PPE can be safely removed.

Supervise wearers to ensure correct use of personal protective equipment

Supervisors should ensure that PPE is correctly worn at all times when wearers may potentially be at risk, that wearers are correctly prepared for wearing their PPE and that the correct personal decontamination procedures are followed when necessary. The overall effectiveness of any PPE programme can be critically dependent on the actions of shop-floor supervisors who are close enough to the wearers to actively enforce correct usage of PPE and any other control methods adopted. The responsibilities and authority of supervisors in the PPE programme should be specified in the supervisors' job descriptions so that they are fully aware of their role and that production does not take precedence over health and safety.

Maintain personal protective equipment in an efficient and hygienic condition

Reusable PPE will need to be cleaned, serviced and maintained to ensure both the ongoing efficiency of the equipment and that the wearer is not exposed to contamination caused by poorly cleaned equipment. Many wearers will be unwilling to fit obviously dirty or faulty equipment. In setting up a PPE programme, the persons responsible should ask themselves 'would I be prepared to wear the equipment provided?'. So called 'low-maintenance' PPE must be cleaned, serviced and maintained as rigorously as any other reusable

PPE: the term 'low maintenance' is therefore misleading.

The legal duty to maintain equipment is placed squarely on the employer. The employer cannot legally delegate such a responsibility unless the person to whom the responsibility is delegated is 'competent'. Even where competent persons are available, the employer still has the responsibility to ensure that the competent persons carry out their duties in a competent manner.

Inspect personal protective equipment to ensure it is correctly maintained

Given that the duty is placed on the employers to ensure that PPE is correctly serviced and maintained, regular inspections and testing of serviced PPE enable them to ensure that the equipment is complete and in good condition.

Provide suitable storage facilities for personal protective equipment

Where reusable PPE is used, suitable storage facilities should be provided for the clean PPE as it is clearly inefficient to service and maintain the equipment if it is then going to be left lying around in the wearers' lockers or in dirty locations where it may be stolen, become contaminated or damaged or may be used by unauthorized or untrained persons. Such provision of suitable storage facilities is recommended in some UK guidance, for example guidance on the Noise at Work regulations (Health and Safety Commission, 1988).

Record usage, maintenance and inspection data

Although there is no legal duty to record usage and maintenance data, there is a legal duty to record inspection data, for example see Regulation 9 (4) of the COSHH Regulations (HSC, 2002). Usage records provide the employer with a means of checking if all those who should wear PPE actually do so, to ensure that those who should wear PPE actually do so and to ensure that those who do not are identified so that corrective action can be

taken. The maintenance and inspection records provide employers with 'proof' that their legal duties to maintain and inspect relevant items of PPE have been met.

Monitor programme to ensure continuing effectiveness

As with any programme, it is generally inadequate to put a programme into operation and assume that it will continue to function adequately without any further action. It is therefore prudent to routinely check the operation of the programme, to retrain and 'reindoctrinate' all personnel involved at suitable intervals and to take any remedial action that may be required.

Provide personal protective equipment free of charge

If the PPE is to be used solely for protection against occupational risks, the equipment must be provided and maintained free of any charge to the wearer. Reusable equipment should be provided as a personal issue item unless very thorough cleaning and decontamination procedures will be reliably followed before use by any other wearer.

When PPE may be used for both occupational and non-occupational purposes, for example crash helmets supplied to motorcycle couriers, if agreed between the employer and the user that the PPE may be used for travel to and from the place of work, the employer may levy a charge proportional to the occupational versus non-occupational usage. However, it would be prudent to provide the crash helmet free of charge.

Personal protective equipment standards

Almost all PPE, other than that intended to provide only minimal protection, is now designed to meet internationally agreed standards and is manufactured using internationally agreed quality control procedures. Standards thus have a significant effect on the type and nominal performance of the PPE available in the marketplace.

Over a period of years, many industrialized countries have developed their own standards for PPE or have adopted the standards of a larger neighbour or previous colonial power. In the European Community it was recognized that the multiplicity of standards among the Member States constituted a potential barrier to trade and it was agreed that the individual national standards organizations would develop harmonized standards that would apply in all Member States of the Community. Over the past 20 years about 160 European Standards (CEN) for PPE have been developed, covering PPE ranging from safety helmets to safety footwear and fall-arrest harnesses.

The PPE 'Product' Directive (Commission of the European Communities, 1989c) requires that all PPE crossing internal borders or imported into the European Community must carry the 'CE' mark, demonstrating compliance with the relevant European or national standard. Where no relevant PPE standards are available, the manufacturer or importer must prepare a technical file defining the performance of the PPE before a CE mark can be issued. CE marked equipment must either undergo annual retesting or manufacture must be conducted in accordance with agreed quality systems to ensure that equipment continues to meet the relevant performance criteria.

The adoption of CEN standards has generated a major problem for some users of RPE in that for half- and full-facepiece unpowered dust and gas respirators there are no standards for complete devices, there being separate standards for facepieces and for filters. It is assumed that any CE marked filter fitted with a standard connector should be able to be fitted to any relevant CE marked facepiece to yield a correctly functioning respirator. However, some filter and facepiece combinations may never have been tested to demonstrate the correctness of the above assumption. Users should therefore only use filter and facepiece combinations for which the supplier can supply written evidence that the different components do combine to give a correctly functioning respirator.

Although the CEN standards specify minimum performance criteria, the standard tests are intended only to provide manufacturers with a means of assessing the initial and ongoing per-

formance of their products vis-à-vis the requirements of a standard. As the standard tests might either be a poor simulation of actual usage in the workplace (e.g. leakage tests for RPE or attenuation tests for PHD last only about 30 min although equipment in the workplace may have to be worn for many hours) or might not test a critical parameter of performance (e.g. protective clothing is not currently tested to assess the potential for generating thermal stress), the results of such standard tests may thus not be a suitable basis on which to select PPE for use in the workplace.

To ensure that the PPE selected is likely to perform adequately in the workplace, only equipment for which relevant workplace data are available should be purchased. As noted above for the provision of information, any supplier unable to provide the required information should be reported to the relevant enforcing authority. In the UK, the relevant authority is the Health and Safety Executive (HSE) as such data should have been generated by the manufacturer or importer to ensure compliance with the requirements of Section 6 of the Health and Safety at Work Act (HSC, 1974).

The reality of personal protective equipment performance in the workplace

For RPE, PHD and protective clothing there is extensive evidence of problems regarding the levels of protection that can realistically be achieved in the workplace. In addition, for these types of PPE there are further concerns regarding how the equipment can be safely used. These problems are addressed below.

Respiratory protective equipment

In the past, it was generally assumed that the major difficulty with RPE was getting the wearer to wear the equipment. In his 1908 report, the Chief Inspector of Factories reported that his inspectors had instanced 'the old difficulty of getting workers to wear the respirators provided' (His Majesty's Stationery Office, 1908). Many users of today still

encounter the 'old' problem of 1908. It was generally assumed that if the RPE would only be worn, the anticipated protection would be obtained.

In the UK, RPE is now selected on the basis of assigned protection factors (APF) derived from measured performance achieved in real workplaces where the APF is defined as the 'level of respiratory protection that can realistically be achieved in the workplace by 95% of adequately trained and supervised wearers using a properly functioning and correctly fitted respiratory protective device' (British Standards Institution, 2001).

In selecting RPE, the minimum required APF is calculated as:

Required APF = contaminant concentration outside the PPE/contaminant concentration in wearer's breathing zone (31.1)

For example, when selecting a respirator for use in a workplace in which the airborne concentration of contaminant is 8 mg m^{-3} and the OEL for the contaminant is 0.5 mg m^{-3} : required APF = $8/0.5 = 16$.

For such a workplace, an item of RPE providing an APF of at least 16 could nominally provide adequate protection if the device is matched to each individual wearer, his job and his workplace, is properly fitted and worn, and is properly maintained if reusable.

As it is considered prudent to limit personal exposures to not more than 25% of an OES or not more than 10% of an MEL, the required APF in the above situation will be 64 in the case of a substance assigned as OES and 160 in the case of a substance assigned an MEL. Where RPE with the necessary APF is not available, it will be necessary to improve the other control techniques applied.

The APF assigned to the different classes of Respiratory Protective Devices are given in BS 4275 and HSG 53, (HSE, 1998a; BSI, 2001). Note that only the most up-to-date editions of these references should be used as the current assigned protection factors are substantially lower than the protection factors used in earlier editions.

There is evidence that the protection provided by conventional RPE may be reduced if wearers sweat heavily. RPE should therefore be selected with care if wearers also need to wear protective

clothing that encloses a large proportion of the body. In such situations it would be prudent to use only powered or air-fed equipment.

For reusable RPE, it will be necessary to change particulate and gas filters on a regular basis. Particulate filters need to be changed because dust collection can increase airflow resistance, so increasing breathing resistance in unpowered devices or reducing airflow rates in powered devices. For particulate filters that rely on electrostatic properties of the media, the collection of dusts, oils or some organic solvents can cause filter efficiency to fall substantially. Clear written guidance should therefore be sought from the supplier to ensure that the contaminants for which the filter will be used are unlikely to affect filter performance or to find out when filters should be changed. The efficiency of particulate filters is described using the nomenclature P1, P2 P3 where 'P' indicates a particulate filter and the number indicates filter efficiency, '1' being the lowest and '3' the highest.

Gas and vapour filters need to be changed regularly because the capacity of such filters is finite and exposure to moisture or other contaminants that may be more strongly retained by the filter medium can cause previously retained contaminants to be released into the wearer's breathing gas.

Gas and combined gas and particulate filters are covered by EN 141 and some special filters are covered by EN 371 and EN 372 (Comité Européen de Normalisation, 1990, 1992a, b). EN 141 has adopted a nomenclature such as 'A2' where the first letter indicates the substances against which the filter should be used and the number indicates the capacity, '1' being the lowest capacity and '3' the highest. Thus an A2 filter is a medium capacity filter for use against organic contaminants with boiling points $> 65^{\circ}\text{C}$ as specified by the manufacturer. The breathing resistance imposed by a gas filter tends to increase with capacity, as does cost. If the filter also provides protection against particulates, the filter nomenclature becomes, for example, 'A2P3', where the 'P3' indicates a medium capacity filter for use against organic contaminants with boiling points $> 65^{\circ}\text{C}$ as specified by the manufacturer which is also fitted with a high-efficiency particulate filter.

It is not possible to reliably calculate the likely breakthrough time for gas filters in the workplace, even if contaminant concentrations are known, given the complexity of filter chemistry and the possible interactive effects of moisture and other contaminants. Some users who have access to suitable analytical facilities consider it worthwhile to test used filters to determine their residual capacity. This ensures that filter usage is optimized as regards both protective performance and cost.

In all situations where gas filters are to be used, written guidance should be sought from suppliers who should be provided with as much information as possible as regards likely contaminants, contaminant concentrations, any other contaminants likely to be encountered and information on any unusual factors such as unusually high work rates or high ambient temperatures or humidities.

Filters that have been exposed previously to very high contaminant concentrations without breakthrough can release the contaminant on subsequent use. Filters should be changed immediately any odour or other subjective effect is detected by the wearer. Such warning effects should be used to supplement a planned filter replacement regimen and should not be relied upon as the indicator of when to change filters as exposure to some materials can cause olfactory fatigue, thereby reducing the sensitivity of subjective detection and for some substances the odour or irritation thresholds can be higher than the relevant OEL.

As for users of particulate filters, users of gas filters should seek clear, written guidance from the supplier to ensure both that the correct gas filters are selected and that the filters are changed when necessary.

Personal hearing defenders

Proposed UK guidance for selecting PHD suggests that the level of attenuation that can be assumed to be achieved in the workplace should be derived from the results obtained in standard laboratory tests by subtracting two standard deviations from the mean attenuation at each frequency (HSC, 2004). The guidance suggests two standard deviations may be subtracted to take account of poor

fitting. However, even if the laboratory attenuations were achievable in the workplace, the 'assumed attenuation' as given by subtracting one standard deviation would result in one wearer in six getting less attenuation than the assumed attenuation. If two standard deviations are subtracted, only 2.5% of wearers would obtain less attenuation.

However, many studies have shown that PHD attenuations in the workplace are lower than the laboratory attenuations [see review by Berger *et al.* (1996)]. For example, Hempstock and Hill (1990) reported that by comparison with equivalent laboratory type data, the field attenuations for earmuffs could be 2.5–7.5 dB lower and for ear inserts could be up to 18 dB lower and that the workplace mean attenuations were lower, and the standard deviations were larger, than in the laboratory. Therefore, had attenuations been based on the mean minus two rather than minus one standard deviation, the relative reduction in the field would have been even larger. Laboratory test data, particularly if based on the mean-minus-one standard deviation, simply do not provide a secure basis from which to derive likely workplace performance. In practice it would be prudent to assume a maximum attenuation of 5 dB from insert defenders and 10 dB from earmuffs [see review by Howie (2001)].

A further difficulty with ear inserts is that such devices should be removed and completely refitted at least every 60–90 min because the insert is ejected from the ear canal by jaw movement as a result of talking, chewing or yawning, as highlighted by Abel and Rokas (1986) and Cluff (1989). Although auto-ejection of ear inserts has been known about for many years, no UK supplier of inserts currently specifies that they should be regularly refitted. The need to regularly refit ear inserts may preclude their use in some workplaces unless good hand-washing facilities are available along with supervision to ensure that both the correct refitting and hand-washing procedures are followed. For example, in the food industry it might not be acceptable for wearers to remove and refit their inserts given the potential for contaminating food. When toxic or abrasive substances are involved, it would be unacceptable to risk contam-

ination from the wearers' fingers being transferred into the ear canal.

The available information indicates that the performance of earmuffs can fall rapidly over the first 4 or 6 weeks of usage (Rawlinson and Wheeler, 1987). If the attenuations provided by the earmuffs only marginally achieve the required in-ear noise exposure levels, the muffs should be replaced at least every 4 weeks unless the supplier can provide information demonstrating that a longer replacement frequency is adequate for the given environment.

In addition, when reliance is placed on personal hearing protectors, wearers should undergo annual high-quality audiometric testing if exposed to ambient noise levels below about 95 dB(A) and progressively more frequent testing as ambient noise levels increase as the 'ultimate' test of hearing protector effectiveness as the legal duty is to protect hearing, not to reduce personal noise exposures.

As for all types of PPE, the supplier should be required to supply sufficient information to enable the user to ensure that the equipment selected will provide effective protection.

Protective clothing

The two major factors that should be addressed when selecting protective clothing are the overall level of protection afforded and the likelihood of the clothing generating or exacerbating heat stress that could restrict safe periods of wear or preclude the clothing being worn with some types of PPE.

The overall performance of protective clothing is highly dependent on the nature of the challenge against which protection is required. If the challenge is simple physical contact or splash, simple laboratory tests might provide adequate information regarding the likely levels of protection that could be achieved in the workplace. However, if the clothing is required to protect against airborne contaminants such as aerosols or gases, it is necessary to take account of all routes by which contaminants could breach the protective layer.

For gases or aerosols, the protective layer could be breached by penetration or permeation through

the fabrics, seams or fasteners or could leak past seals at body openings such as at the neck, wrists, waist or ankles. The mechanism by which contaminated air can be drawn into protective garments is that movement of the wearer's body creates and destroys voids between the body and the garment; air is drawn inwards as the voids are created and expelled outwards as the voids are destroyed. A pre-war Home Office document indicated that such 'suction effect produced by movement' could cause the inward leakage of mustard gas to contaminate the inside of protective clothing (HMSO, 1938). When selecting protective clothing for use against gases or vapours, the suppliers should be required to supply information regarding the overall protection afforded by their products in real workplaces when worn as specified in their user instructions.

Some types of protective clothing may reduce or prevent the body losing heat generated by metabolic processes. This effect of protective garments was recognized in the 1938 Home Office document referred to above, which indicated that air-raid wardens who had to wear heavy anti-gas suits with a full-facepiece respirator and hood might be restricted to three periods of 30 to 60 min hard work per 24 h. As the situation to which the above document refers was that the wearers could be involved in saving life and that lives could be lost if the protected wearers were not available, the limitation was clear recognition of a serious health threat caused by the PPE itself. For a person working moderately hard, for example producing about 100 W of useful external work, the body has to lose about 400 W of heat if a body efficiency of about 20% is assumed. If clothing prevents the loss of this heat, heat will be stored in the body leading to an increase in core temperature. For small increases the effect is only discomfort. However, if the storage rate is excessive, the effects can range from heat cramps to death.

BS 7193 in conjunction with BS EN 27243 (BSI, 1994, 2000) give guidance on permissible working conditions for persons wearing personal protective equipment. Such guidance is based on limiting heat storage so that any increase in deep body temperature does not exceed 1°C. BS EN 27243 uses the wet bulb globe temperature (WBGT) index for

classifying the thermal environment and gives limiting values for different work loads. BS 7193 is valid only for workers carrying out moderate tasks and will underestimate heat strain in those working at higher activity levels. If the limiting values are exceeded, BS 7193 directs the reader to BS EN 12515 (BSI, 1997). Note that the scientific validity of BS EN 12515 is dubious and that this standard should therefore not be used for other than non-enclosing PPE in normal room temperature conditions.

Protective garments often have to be fitted well before potential exposure to contaminants and may need to be worn after exposure while carrying out any decontamination procedures. That is, total wear times may be substantially greater than the time nominally required to carry out the primary task.

As noted above there is evidence that the protection provided by conventional RPE may be reduced if wearers sweat heavily. RPE should therefore be selected with care if wearers also need to wear protective clothing that causes significant heat strain.

It is therefore essential that full consideration be given to the possible thermal consequences of the work environment, the wearer's energy expenditure and the PPE ensemble worn.

General conclusions and recommendations

In any workplace, the major effort should go into reducing risks as far as technically possible by means other than by the use of PPE (see Chapter 29). Unless PPE is being used in an emergency situation or the risks are minimal, PPE should only be used as one component of a comprehensive programme to prevent and reduce risks to safety and health.

Where PPE must be used, the levels of protection assumed should be based on information derived from tests in representative workplaces or from information provided by the supplier. Only if the above information is not available, should equipment be selected on the basis of standard laboratory test data.

Unambiguous written guidance should be sought from the suppliers of any PPE regarding: (1) the protective performance of their product in real workplaces over likely wear durations; (2) any potential problems due to incompatibility with other types of PPE; (3) any risks that may be generated by their product; and (4) relevant servicing and maintenance information. If any suppliers are unable or unwilling to supply the required information, their products should not be bought and their inability to supply such information should be reported to the relevant enforcing authority.

PPE should be selected in conjunction with the wearers to ensure that any equipment selected provides adequate protection without generating either unacceptable risks or discomfort for the wearer. The overall performance of PPE is best summed up by a quotation from the 1988 draft *Approved Code of Practice for Carcinogenic Substances* (HSC, 1988): 'but PPE, particularly RPE, depends for its effectiveness on the wearers willingness to wear it'.

In practice it would be prudent to assume a maximum attenuation of 5 dB from insert defenders and 10 dB from ear muffs and to ensure that PHD wearers undergo regular high quality audiometry. The thermal consequences of personal protective equipment should be assessed using BS 7193 and BS EN 27243.

References

- Abel, S.H. and Rokas, D. (1986). The effect of wearing time on hearing protector attenuation. *Journal of Otolaryngology*, 15, 293–7.
- Berger E.H., Franks, J.R. and Lindgren, F. (1996). International review of field studies of hearing protector attenuation. In *Scientific Basis of Noise-Induced Hearing Loss* (eds A. Axlesson, H. Borchgrevink, R.P. Hamernik, P. Hellstrom, D. Henderson and R.J. Salvi), pp. 361–77. Thieme Medical Publishers, New York.
- British Standards Institution (1994). *Hot Environments – Estimation of the Heat Stress on Working Man, Based on the WBGT-index (Wet Bulb Globe Temperature)*. BS EN 27243. BSI, London.
- British Standards Institution (1995). *Acoustics. Hearing Protectors*. BS EN 24869–1. BSI, London.
- British Standards Institution (1997). *Hot Environments – Analytical Determination and Interpretation of Thermal Stress using Calculation of Required Sweat Rate*. BS EN 12515. BSI, London.
- British Standards Institution (2000). *Ergonomics of the Thermal Environment: Guide to the Assessment of Heat Strain of Workers Wearing Personal Protective Equipment*. BS 7193. BSI, London.
- British Standards Institution (2001). *Guide to Implementing an Effective Respiratory Protective Device Programme*. BS 4275. BSI, London.
- Cluff, G.L. (1989). Insert-type hearing protector stability as a function of controlled jaw movement. *American Industrial Hygiene Association Journal*, 50, 147–51.
- Comité Européen de Normalisation (1990). *Respiratory Protective Devices – Gas Filters and Combined Filters – Requirements, Testing, Marking*. EN 141:1990. CEN, Brussels.
- Comité Européen de Normalisation (1992a). *Respiratory Protective Devices – AX Gas Filters and Combined Filters Against Low Boiling Organic Compounds – Requirements, Testing, Marking*. EN 371:1992. CEN, Brussels.
- Comité Européen de Normalisation (1992b). *Respiratory Protective Devices – SX Gas Filters and Combined Filters Against Specific Named Compounds – Requirements, Testing, Marking*. EN 372:1992. CEN, Brussels.
- Commission of the European Communities (1989a). *Council Directive of 21 June, 1989 on the Introduction of Measures to Encourage Improvements in the Safety and Health of Workers at Work*. 89/391/EEC. CEC, Brussels.
- Commission of the European Communities (1989b). *Council Directive of 30 November, 1989 on the Minimum Health and Safety Requirements for the Use by Workers of Personal Protective Equipment at the Workplace* (third individual directive within the meaning of Article 16(l) of Directive 89/393/EEC). CEC, Brussels.
- Commission of the European Communities (1989c). *Council Directive of 21 December 1989, on the Approximation of the Laws of the Member States Relating to Personal Protective Equipment*. 89/686/EEC. CEC, Brussels.
- Health and Safety Commission (1974). *Health and Safety at Work etc. Act, 1974*. HMSO, London.
- Health and Safety Commission (1988). *Draft Approved Code of Practice for the Control of Carcinogenic Substances*. HMSO, London.
- Health and Safety Commission (2002). *Control of Substances Hazardous to Health Regulations*. HSE Books, Sudbury.
- Health and Safety Commission (2004). *Proposals for New Control of Noise at Work Regulations implementing the Physical Agents (Noise) Directive (2003/10/EC)*. Consultative Document 196. HSE Books, Sudbury.
- Health and Safety Executive (1998a). *Respiratory Protective Equipment, a Practical Guide for Users*. Health and Safety Series Booklet HSG 53. HSE Books, Sudbury.
- Health and Safety Executive (1998b). *Reducing Noise at Work*. L108. HSE Books, Sudbury.

- Health and Safety Executive (2003). *Fit Testing of Respiratory Protective Equipment Facepieces*. Information Document 282/28. HSE Books, Sudburg.
- Hempstock, T.I. and Hill, E. (1990). The attenuation of some hearing protectors as used in the workplace. *Annals of Occupational Hygiene*, **34**, 453–70.
- His Majesty's Stationery Office (1908). *HM Chief Inspector of Factories Annual Report*. HMSO, London.
- His Majesty's Stationery Office (1938). *Personal Protection Against Gas*. HMSO, London.
- Howie, R.M. (2000). Are your lips sealed? *Health and Safety at Work*, February, 30–1.
- Howie RM (2001) Hear say? *Health and Safety at Work*, **August**, 30–1.
- Rawlinson, R.D. and Wheeler, P.D. (1987). The effects of industrial use on the acoustical performance of some ear muffs. *Annals of Occupational Hygiene*, **31**, 291–8.

Chapter 32

Occupational health and hygiene management

Lawrence Waterman and Karen Baxter

Introduction	Measuring system performance
Organization context	The health and hygiene team
Business risk management	Stakeholder involvement
Corporate social responsibility	Directors/trustees
International trends	Workforce
Legal compliance	Community
Management in practice	Health-care system
Hazards and risks	Regulators
Health hazards	Formal management systems
Qualitative assessment of risks	Continual improvement
Categorizing risk	Auditing/verification
Quantitative assessment of risk	Linking health and hygiene management to safety, environment and other issues
Control programmes	Similarities
Staff training/consultation	Differences
Behaviour	Health and hygiene in the boardroom
Monitoring and maintenance	Managing business risk
Record-keeping	Corporate social responsibility
Occupational health support	Setting organizational objectives and targets
Pre-employment health status checks	Leadership
Health surveillance	The future
Case management	Notes
Rehabilitation	
Well-being	

Introduction

Occupational hygiene, as a discipline and a practice, is only of social worth if it helps to shape safer, healthier workplaces. To achieve this requires a sound technical knowledge of a wide variety of health risks and control techniques – but it also requires an ability to influence, change and lead organizations. For all but the very few who work in social isolation, exposures to risks to health at work occur within organizations, many of which are large and complex and most of which, whatever their size, have a large number of interfaces with others. This chapter is about how to successfully reduce occupational ill health through the correct application of occupational

hygiene skills within a broader organizational framework.

The theories of organizations have spawned many schools – from Taylorism and the measurement of work to systems theory (generic statements about all organizations) and the analysis of institutions (focusing on the unique characteristics of any enterprise). Nevertheless, there are several aspects of all forms of modern management that represent something akin to a mantra for managers and directors of organizations. If an organization is to improve its performance, and that may mean becoming more effective (that is achieving desired results, for example increasing the rate of production) or becoming more efficient (that is ensuring that the resource demand is sustainable,

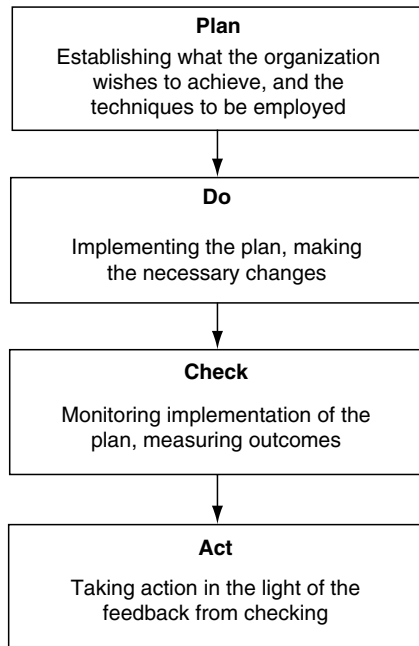


Figure 32.1 Core steps common to all management systems.

e.g. achieving the same level of production but with less materials by reducing wastage), it will need to implement a management programme based upon four key steps (Fig. 32.1).

The effective management of health and hygiene in the workplace requires this same basic approach. Nevertheless, there are aspects specific to the subject – monitoring includes the application of occupational hygiene techniques that are not utilized in other fields – and the general management strategy needs to be moulded to address the particular hazards and risks associated with and caused by the work under consideration. This lends a certain character, almost a personality, to the management of occupational health and hygiene, but the similarities with general management remain greater than the differences.

In any management programme it is essential to be clear about the objectives from the beginning. In the WHO Declaration on Occupational Health for All, adopted October 1994,¹ there were three strands to the argument for the development of a comprehensive approach to health in the workplace:

1 The acknowledgement of the fundamental right of each worker to the highest attainable standard of health.

2 The focal point for practical occupational health activities is the workplace, with employers having the responsibility for planning and designing a safe and healthy workplace, work environment and work organization, as well as for maintaining and constantly improving health and safety at work.

3 The importance of universal access to occupational health services to support this effort.

With supplementary aims of assisting people with disabilities or ill health conditions to stay in or re-enter the workforce, and a recognition of the collective, collaborative effort that is required involving the workforce and its representatives – the objective of a healthy and present workforce is the *raison d'être* of the management of health and hygiene at work.

With this overarching set of objectives in mind, this chapter looks at:

1 the organizational environment in which the management of occupational health and hygiene finds itself and contributes to;

2 the elements of an integrated programme and how they need to be managed and coordinated;

3 the role of stakeholders in management success and failure, formal systems and why they have been developed, linking health and hygiene to other related issues such as environmental management and quality programmes; and finally

4 the role of leadership in organizations.

The chapter ends by considering how these themes are likely to develop over the next few years.

Organization context

Business risk management

In the light of certain scandals and criticisms of business practice in the late 1980s and 1990s, the London Stock Exchange commissioned a series of reports (Cadbury, Greenbury, Hampel) and adopted a set of rules for risk management based on the Turnbull Report.² The now mandatory Combined Code requires UK stock market-listed com-

panies to identify, record and control their significant risks in a suitable manner. To achieve this, a systematic framework is required to direct managers to regularly review risks to the organization and develop the arrangements for prevention, mitigation and recovery. Statements have to be made in the company annual reports confirming the effectiveness of these systems. The risks that have to be dealt with in this way are those which are significant or, in financial terminology, 'material' risks. Thus they may encompass health hazards that could lead to significant lost working time through absence, substantial compensation claims, make it difficult for the organization to recruit and retain staff or damage reputation and thus corporate value.

Corporate social responsibility

Much of the management of organizations, in practice, is essentially the management of business risks. The London Stock Exchange requirements are an illustration of the rise to prominence in recent years of *corporate governance* and *corporate social responsibility* as key issues. The willingness of an increasing number of individuals in advanced industrialized countries to specify that their investments, such as their pension contributions, should be placed in ethical investment trusts or similar vehicles has created an influential lobby for social responsibility. Clearly, there is a coming together of various interest groups. At one time it was thought that long-term investors (pension and insurance funds are the two major categories of institutional investors in the UK) were only concerned with investment security and growth. In contemporary investment markets such investors are looking for indicators of good management, and the effective management of health and safety risks are increasingly being used as a sign of good management in general.

Further, with the concerns over stability and sustainability of organizations in the wake of scandals that have rocked global stock markets in recent times, the ethical funds' approach and that of general investment funds is looking increasingly similar.³ It should also be noted that consumer pressure in this area may not be restricted to environmental concerns, as recent market research has

demonstrated a public interest in health and safety.⁴

International trends

The trends towards more scrutiny of organizations and their internal controls, over risks that could harm them or even threaten their survival, have led to a number of initiatives both inter-related and independent. Transnational bodies such as the European Union and the Organization for Economic Co-operation and Development (OECD) have published consultation and guidance papers on corporate social responsibility. The Global Reporting Initiative has encouraged greater transparency in the methods organizations adopt to communicate their objectives and performance. The reshaping of the global environmental agenda from the Rio Summit to the 2002 conference on sustainability in Johannesburg is another sign of how these different strands are being woven together. They all help to create the context in which private and public bodies alike operate and seek to manage their risks including those that may affect the health of their employees.

Legal compliance

There is one further defining aspect to the context for occupational health and hygiene – the legal framework. Although it varies greatly from country to country, from sets of highly specific regulations to goal setting laws that define processes rather than minimum standards, the law is always a great shaper of the health and hygiene programme that any organization develops. For many larger organizations, the minimum standards established by statute are subsumed within a higher set of minimum corporate standards, which in turn are developed in annual programmes to improve performance and seek to achieve best practice.

However, even where a management programme has largely been drawn up by reference to occupational hygiene good practice, it is usually necessary to evaluate it against the particular and specific legal duties that health and safety law places on the enterprise. For example, when new strategies are being drafted, a compliance register

that relates the legal obligations of the organization's activities forms an important and helpful planning tool. It must also be recognized that for many businesses, particularly smaller ones, achieving and maintaining legal compliance is the main and perhaps sole driver for occupational health and safety programmes. For multinational companies seeking to establish a single corporate standard, as a minimum this has to meet the highest legal standard set for any aspect of risk management across the countries where company sites are located.

Management in practice

To understand what needs to be in place to effectively manage occupational health and hygiene, it is necessary to discuss the activities that have to be carried out, the team that carries out those activities and the organizational arrangements. The starting point in modern health and safety management is the risk assessment.

Hazards and risks

The terms hazard and risk are used in everyday speech as if they are interchangeable, but technically a *hazard* is something with the potential to cause harm; a *risk* arises only when people are exposed to the hazard. Thus high levels of noise emissions are a health hazard, but it only turns into a risk when one or more people are exposed to that noise. This means that characterizing a hazard in some senses is based on evaluations that rest on theoretical occupational hygiene principles – considering the toxicity of substances and the physiological effects of physical agents, whereas risks require a probabilistic analysis, considering the likelihood and severity of potential harm arising from the type of exposures that occur in the workplace. In occupational health management, therefore, the driver for everything we do is the identification of hazards – the potentialities for causing harm – and the evaluation of the risks that can arise.

Making sensible judgements about risk, how people may be harmed and how serious that

harm may be helps us determine what precautions it is appropriate to take. That is why the oft-quoted standard definition of occupational hygiene is *the recognition (hazards), evaluation (risks) and control (precautions)* of threats to health caused by work. Compared with the risk of accidents, this process is more important in protecting health, because many of the hazards are not obvious, and the risks are much harder to pin down. Often it is only the risk assessment that raises managers' and workers' understanding of the risks in their work and the necessity to take precautions.

Health hazards

Health hazards fall into one of the major categories:

- chemical;
- biological;
- physical;
- ergonomic;
- psychosocial.

The requirement is to identify the hazards associated with the work, to establish what dusts, fumes, noise, etc. could arise from the work when it is being carried out normally, and the same again during maintenance or other activities which may be considered 'non-standard'. It is usually helpful to start with the classification of work activities. In addition, it is necessary to consider the workplace itself, particularly with a view to identifying contamination problems.

For a complex workplace, it is common to divide it into departments or sections in accordance with the line management structure. For each task and activity, it is helpful to systematically think through the hazards intrinsic to the work or which could arise. Records and reports of musculoskeletal problems, high levels of work-related sickness absence and industrial diseases such as dermatitis will be indicators to be taken into account. When carrying out such a review, it is worth remembering that there are three sources of workplace health hazards:

- *health hazards found within the workplace*, for example chemicals (contaminated sites), asbestos in buildings or the presence of rats in sewers and the possibility of *Leptospira* in their urine;

- *health hazards brought into the workplace*, for example chemical feedstock and paints;
- *health hazards created in the workplace as a result of the activities carried out there*, for example manual handling of materials, dust generated by machining timber, vibration from drilling equipment, stress associated with organizational structure, workload and culture.

Qualitative assessment of risks

It is unlikely that many hazards identified will not already be subject to controls that serve to reduce the risks arising. Such precautions may be embedded in standard practice for a particular trade or skill, or have been determined by previous risk management efforts. Therefore, any new, formal risk assessment needs to be of the residual risk, that is the risk to health that remains even after the existing controls are implemented. The purpose of health risk assessments is to determine what needs to be done to protect the health of workers and others who may be adversely affected by their activities. This requires the work activities to be correctly characterized, and to evaluate what more should be done to prevent or control exposures – and because work changes and new control techniques become available, this is a continual process.

It is crucial that the identification of opportunities to improve such protection should take into account the arrangements already in place. The residual risk is judged by looking at the harm that could arise, and the likelihood of that harm arising, despite all the precautions that are already being taken. When assessing the risks, the following will influence the judgements made.

- *Legal*. Many health risks are subject to legal controls and standards, compliance with which represents acceptable levels of control – in this sense, the statutory authorities have undertaken risk assessments on every employers' behalf and determined the minimum control regime. For example, in the UK many Occupational Exposure Limits for airborne hazardous substances have been set on the basis that they are effectively 'No Observed Adverse Effect Levels', and if exposures

are maintained below such a level no further control action is legally required.

- *Guidance*. Even where there are no formal regulatory risk standards, there are often sources of guidance that should be checked. First, statutory authorities (such as the Health and Safety Commission and Health and Safety Executive in the UK, OSHA and NIOSH in the USA) publish much guidance – some specific to particular hazards (such as asthmagens), some related to particular workplaces (such as motor vehicle repair workshops). Second, there are trade and other associations that agree standards that, on the basis of 'reasonable practicability', have substantial backing in law. For example, the advice from the Association of the British Pharmaceutical Industry – ABPI – issuing guidance on setting in-house exposure standards for pharmaceutical substances,⁵ and the Chartered Institution of Building Services Engineers – CIBSE – guide to ventilation systems and rates.⁶

- *Experience*. Despite the often long timescale associated with ill health compared with accidental injuries, many of the risks that are being assessed will be familiar to an experienced assessor. For example, poor housekeeping and work practices will be identified by an experienced occupational hygienist and attention drawn to their relationship with elevated skin contamination.

- *Concerns*. Risk assessment expresses current views on the acceptability of risk. The worries and concerns of the workforce and the managers, or the absence of such concerns, will have an impact on the judgement as to the significance of the risk and be reflected in the priorities assigned to the control measures. Of course there are dangers here, as health risks may be less well understood, and misperceptions of the level of risk engendered may be encouraged. For example, in one workplace there may be a dismissal of the significance of risks associated with noise exposure, whereas another may have seen staff in the past requesting 'radiation protection' aprons for display screen work.

One of the techniques utilized by organizations, typically with large groups of workers engaged in similar tasks and/or working in similar environments, is to establish by job analysis and

observation the membership of similarly exposed groups (SEGs). This facilitates making risk assessments efficiently, and when quantitative risk assessments based on exposure monitoring are required, such monitoring can be limited to a representative SEG sample (if this generates too high a variance of measured exposures, the SEG membership should be readjusted).

Categorizing risk

Risk assessment decisions to be made are ones of categorization: ranging from an intolerable risk which requires work to cease or not be started to trivial risks which need not trouble the assessor.⁷ The judgement is based on the following five criteria that relate to the potential severity of harm, the likelihood of harm and the people likely to be harmed.

1 All other factors being equal, something that represents a risk of death or serious, permanent harm to health is regarded as a much higher risk than something that, even if it were to happen, will only result in a temporary health effect from which the person(s) concerned fully recover with no long-term consequences. For example, entry into confined spaces – the availability of a breathable, non-toxic atmosphere is safety critical.

2 All other factors being equal, something that represents a risk to a person who is vulnerable to harm is regarded as a much higher risk than that to someone who is likely to be able to withstand the exposure. For example, there is a special duty of care to prevent eye injury to a worker with only one eye.

3 All other factors being equal, something that represents a risk to a number of people is regarded as a much higher risk than that which will only harm an individual – and the higher the number potentially affected, the higher the risk. Examples include the use of solvents in polishes or paints, which not only contaminate the local work area, but can also affect staff throughout a building if distributed by the ventilation system.

4 All other factors being equal, something that represents a risk to members of the public, particularly vulnerable people such as children or older people, is regarded as a much higher risk than that

which only presents a risk to workers. This is partly because the range of controls available cannot usually extend to managing the way in which non-employees behave, whereas this can be a key control for workers such as the wearing of PPE. It should be remembered that occupational exposure limits are set with healthy workers in mind, and not those whose compromised health status keeps them out of the workforce.

5 All other factors being equal, something that represents a risk that has a great likelihood of actually happening is regarded as a much higher risk than that which could conceivably occur but is regarded as very unlikely. For example, one would expect greater controls over the handling of a potent respiratory sensitizer, such as isocyanates, compared with a weak skin irritant such as some synthetic fibres and wool.

Evaluating these criteria, it should be possible to categorize the health risks as trivial, tolerable, moderate, substantial or intolerable and, accordingly, to not only identify where additional controls are possible but also assign priorities for implementing such controls (Table 32.1).

There are two *caveats*:

1 The risk assessments should make sense to the stakeholders including workers and their managers.

2 It may prove difficult or impossible to complete the risk assessment without measuring the exposures – specifically where there are agreed exposure standards (statutory or in-house).

Quantitative assessment of risk

There are a number of reasons for monitoring levels of airborne dusts, gases and vapours, noise

Table 32.1 Simple risk estimator [from BS 8800, BSI (1996)].

	<i>Slightly harmful</i>	<i>Harmful</i>	<i>Extremely harmful</i>
Highly unlikely	Trivial risk	Tolerable risk	Moderate risk
Unlikely	Tolerable risk	Moderate risk	Substantial risk
Likely	Moderate risk	Substantial risk	Intolerable risk

emissions, vibration exposures and other health risks in the workplace.

1 For the purposes of risk assessment, such measurements may assist in establishing whether or not exposures are above a threshold for action, such as UK (and EU) statutory action limits for noise exposure. Such data will assist in characterizing the risks in accordance with the schema discussed above.

2 It may be difficult to establish which workers, perhaps identified through their job descriptions and titles, are exposed to significant risks without monitoring. Monitoring within such groups will also confirm or challenge the validity of the assignment of individuals to SEGs.

3 Exposure monitoring may also be conducted to evaluate the effectiveness of control measures and their maintenance, to commission new equipment or work areas, to evaluate the impact of work practice changes, to determine the specification for PPE and to meet statutory requirements.

4 Monitoring can establish a baseline, for example when commissioning a new work area, which can then be used as a comparator for future routine monitoring to confirm the maintenance of exposure control.

Thus, exposure monitoring is not only a standard technique available through the application of occupational hygiene methodologies – it also represents a crucial element as part of and alongside risk assessment in the management of health risks. It should be noted that almost all monitoring is an imposition on the workers concerned, and it is resource intensive – it is important, therefore, to only conduct monitoring when the data generated will be of practical benefit in making control decisions.

Control programmes

The purpose of risk assessment is to identify and characterize significant risks, in order to exercise control over them such that health is protected. The control measures that are identified to further reduce risk, perhaps completely preventing exposure to some particular risks, are the crucial outputs from the assessment process. The controls available range from those that eliminate the risk, to

those that reduce it and contain it by design without reliance upon individuals' behaviour (as this is more reliable), to personal protection for individuals. The types of controls available have been placed in a preferred hierarchy by the EU and other legislators, targeting the protection of groups ahead of individuals. The strategies available include:

- *Elimination of the risk*, such as replacing manual by mechanical handling, purchasing ready-diluted acids and alkalis, thus eliminating the dilution process.
- *Substitution of materials, processes or equipment* to reduce the risk, such as replacing organic solvent-based paints and adhesives with water-based alternatives, choosing different surface coatings to isocyanates, replacing a noisy/dusty/vibration technique, such as scabbling, to form a key on poured concrete with alternatives (use of retardants to maintain the concrete in its uncured state until the next phase of concreting can take place, chemical keying, use of needle guns), using chemical feedstocks in pellet rather than powder form.
- *Technological solutions* – engineering controls, such as the installation of local exhaust ventilation to capture dust from cutting equipment or welding fumes, the use of a water spray to capture dust from a disc cutter, moulding building blocks with handholds for easier handling, the use of noise/vibration-dampening mounts on equipment, the use of rubber rather than metal crates to catch machined components and constricting work within fume cabinets.
- *Containment* – restricting health risk exposures to the fewest staff such as the use of hearing protection zones that may only be entered by authorized persons (this presupposes that precautions will be taken by the authorized persons for their own protection).
- *Work planning* – sharing exposures amongst the exposed population so that individuals are not exposed above what is determined as an acceptable threshold (usually set, with a safety margin, below the level at which adverse health effects are predicted). This type of control requires cooperation between supervisors, managers and the operatives, for example in handling vibrating tools.

- *Management techniques* – in addition to ‘sharing’ exposures, there are techniques that are applicable to minimizing psychosocial risk exposures such as bullying, aggression and violence, excess workload and other pressures.
- *Personal protective equipment* – as a temporary measure while other controls are being designed and implemented, or as a back-up to controls where the risk is very high, or if it is not possible to reduce the health risk to an acceptable level by the techniques described – the selection and supply of PPE may be appropriate. However, PPE at best only protects the user and then only to the extent that it is correctly selected and properly used.

Staff training/consultation

The risk assessments themselves should be completed by staff who are competent – that is they have the necessary skill, knowledge and experience to do so. In many cases this requires teams of people to work together, with experts supplementing the intimate knowledge of the work processes held by those engaged directly in them. The workers, their supervisors and managers also need to have a good understanding of the hazards that have been identified, the significance of the risks and the way in which the controls are to be employed. From induction training to updates, toolbox talks and formal training courses there is a requirement to educate all those engaged in work that engenders significant health risks.

With good will, such training can be transformed into an opportunity to discuss, consult and develop approaches to health risk management that are ‘owned’ by the workforce as well as senior management or occupational health advisers. Within the EU there are formal requirements for consultation, but in addition there is an overriding practical consideration in that it provides an opportunity to ensure that major issues are not missed and that the control strategy is acceptable to the workforce as a whole. As far as possible the workforce, either directly or through their elected representatives, should be considered a key resource in the management of health risks.

Behaviour

The core of occupational hygiene practice is the recognition, evaluation and control of health risks in the workplace. The controls are not exclusively technical, however, and require workers to carry out their tasks in accordance with views about best practice. The application of safe working methods, the correct use of control measures such as exhaust ventilation and the wearing of PPE are all dependent on the behaviour of the workers and their managers. For these reasons, in recent years there has been a growing use of behavioural science in establishing a culture of safe and healthy working within the workplace. The main method employed is the introduction of feedback from the observation of work practices – typically by fellow workers but in some cases by supervisors – with encouragement for most of the feedback to be positive, praising and rewarding staff for correct work practices. This reinforcement of safe working is carried out so that it becomes habit-forming, behaviours adopted and implemented almost without thinking, so that constant stimulation is no longer required. By these means the aim is to reduce the observable occasions of unsafe and unhealthy working. The consequential reduction in exposure has a beneficial effect on reportable harm.

Monitoring and maintenance

The controls introduced, typically a combination of all of the options described above, require routine maintenance and checks so that they remain effective. This generates a planned maintenance programme for ventilation systems, managerial arrangements for the issuance of PPE, and staff training and similar mechanisms to sustain the efficacy of the controls. When risk assessments are conducted and controls specified, the specification should also address this long-term issue. Related to the maintenance of controls is the routine monitoring of exposures, ventilation rates, etc. so that the data that originally stimulated the controls and the baseline checks conducted at commissioning may be compared with current levels. Evidence that exposures are rising may then be used to stimulate further control efforts without waiting

for health surveillance to indicate that health is being affected.

Record-keeping

It is essential that proper records are kept. Statutory requirements apply to risk assessments for all but trivial risks, the testing and maintenance of control measures such as local exhaust ventilation and personal exposure data. In addition, the health management system benefits from periodic review and reappraisal, for which baseline records are a vital source of information. There is a management dictum that developed when quality systems were being introduced: 'if it isn't recorded it hasn't happened' – and the need to prove that health is being protected, to demonstrate this aspect of corporate governance of risks, requires a degree of written record-keeping (Fig. 32.2).

Occupational health support

The occupational health expertise required for the management of an effective programme includes the support required for the risk assessment and control activities. There are a number of potentially medically orientated activities that assist and supplement the core effort of reducing significant exposures to health risks. These include the following.

Pre-employment health status checks

The need to ensure that the work and the worker are fitted for each other is satisfied by a variety of means, from simple pre-employment questionnaires to detailed, highly specific medical examinations for occupations such as diving. Although care should be taken to minimize discrimination, there

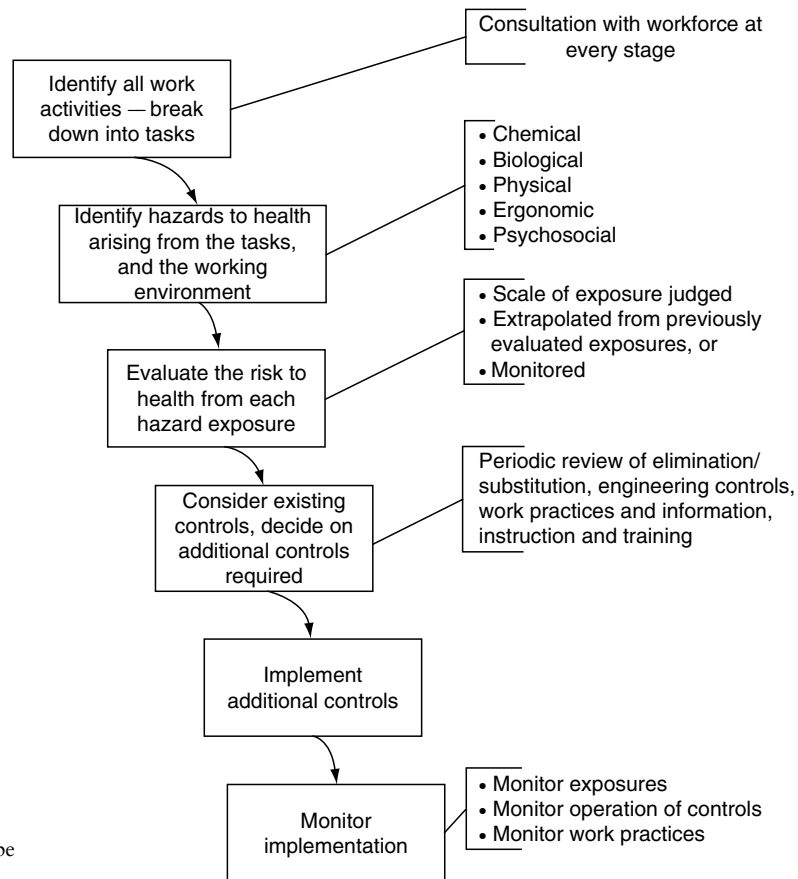


Figure 32.2 Key actions that should be recorded.

are certain minimum health standards for many jobs, which, when met, reduce the risk of each employed person being put at personal risk in their subsequent work. These checks also apply to promotion or other movements within an employment.

Health surveillance

Risk assessments can identify certain work that engenders significant health risks. Although controls are put in place to protect against such risks and prevent adverse ill-health effects, if there are methods for detecting signs and symptoms of such effects they can be used to check the effectiveness of the controls. Health surveillance is widely used to monitor precautions against hand–arm vibration, dermatitic agents, asthmagens and other agents.

Case management

Pre-employment screening and health surveillance are typically applied to categories and groups of workers, but once a worker has presented with an ill health condition that may be associated with work, the individual case has to be managed. Case management is focused on getting people back to work as soon as is possible and consistent with their own recovery and well-being. Often the occupational health physician or occupational health nurse has to lead, as the individual's GP would not usually be in a good position to identify the steps required before a partial and managed return to work is possible – and the beneficial effects of returning to the workplace earlier would be denied to workers recovering from injury or ill health.

Rehabilitation

A subgroup of cases to be managed represents workers who need active rehabilitation in order to return to the workplace. This can apply to mental health and physical disabilities, and a combination of targeted treatments combined with workplace adaptations may be required. Such adaptations may include modified workload, shorter working hours during recovery, physical changes to the workplace or other ways of facilitating a return to work prior to or in the absence of a full recovery.

Well-being

The emphasis thus far has been on the prevention of harm, of the ill health that can arise from exposures to harmful agents within the workplace. The complementary occupational health effort is in the enhancement of health status through programmes that encourage, for example, healthy lifestyle choices such as good diet and non-smoking. The occupational health support that has been discussed in terms of preventing and responding to ill health can also play a positive role in working towards a healthier workforce. This applies to the interest in developing work–life balance programmes designed to adjust working patterns. Regardless of age, race or gender, the aim is for everyone to find a rhythm to help them combine work with their other responsibilities or aspirations. Increasingly, employers are developing a wide range of work–life balance options, covering flexible working arrangements and flexible benefit packages such as:

- flexi-time;
- staggered hours;
- time off in lieu;
- compressed working hours;
- shift swapping;
- self-rostering;
- annualized hours;
- job-sharing;
- term-time working;
- working from home;
- breaks from work;
- cafeteria benefits;
- assisted access to a local gym.

Measuring system performance

For its 10-year occupational health strategy in June 2000, the UK Government adopted just three measurable performance indicators: reductions in the incidence of work-related ill health for employees and for members of the public affected by the work of others, and a reduction in the number of days lost due to work-related ill health rehabilitation. Negative measures that have to be used to compare organizational performance with these targets clearly include the monitoring of the numbers of

people whose health is adversely affected, and the number of days lost to normal working that this causes. In addition, it is also possible to identify positive performance indicators, such as the numbers of staff who attend training courses, or the proportion of work activities that have been subject to detailed risk assessment. In these examples, the effort of the organization rather than outcomes can be measured and reported. The measurement of performance is a necessary activity to conform to the developing public reporting requirements of public and private bodies alike.

The health and hygiene team

The management functions discussed range from health checks on employment and periodically to communication and consultation with the workforce, from monitoring exposures to harmful agents, to the design, installation and maintenance of ventilation systems. The competencies required to advise on and discharge all of these functions are broad and incapable of being found in any one individual. The competencies required often result in teams with several permanent members, and others brought in as required:

- *occupational hygienists* to lead the work on hazard identification, risk assessment, control and monitoring;
- *occupational health and safety advisors* to work with the occupational hygienists for the integration of health risk and safety risk management;
- *occupational health physicians and occupational health nurses* to support the risk assessment process, and engage in pre-employment screening, health surveillance and related work;
- *HR professionals* to lead on organizational structure, staff arrangements and matters such as absence management and stress;
- *directors, line managers, supervisors and staff* (including their representatives) who need to be fully engaged for the health risks to be effectively managed.

In addition, other skills are required to create the appropriate complement to manage specific workplaces. These will range from occupational psychologists, ergonomists and acousticians who are able to diagnose and offer solutions, through to

plumbers and ventilation engineers required to execute plans for welfare facilities or local exhaust ventilation. Modern health risk management requires a multiskilled team to be properly coordinated and managed in order to effect improvements in performance.

Stakeholder involvement

Some stakeholder analyses help to define what a system needs to deliver in order to satisfy the expectations of those with an interest in the outputs of that system. For occupational health and hygiene, some stakeholders also need to be active participants for the system to be effective either because of their special knowledge and/or their leadership role.

Directors/trustees

Directors (charity Trustees) are legally responsible for organizational performance. Traditionally, financial performance indicators are the only ones included in Directors' annual reports but measures of performance in other key areas relevant to occupational health, notably 'Corporate Social Responsibility' (CSR), are increasingly required. UK accounting standards for organizations quoted on the London Stock Exchange and for registered charities require that Directors (Trustees) provide assurances that all significant risks, including occupational health risks, have been identified and appropriate controls are in place. Good performance depends upon the leaders of organizations demonstrating that health and safety results are critical to business goals. This should be reflected in annual business targets and performance should be a standard part of business agendas.

Workforce

Workers are both the group largely exposed to or protected from the health risks in the workplace and also a key component of an effective management system. In the UK, the statutory framework accords certain rights and privileges to trades union-appointed safety representatives and establishes corporate safety committees as an

appropriate forum for developing occupational health and safety strategies. Research has shown that the presence of safety representatives has a major impact on accident rates (although there is the *caveat* that enterprises that encourage worker participation may already have the type of organizational culture which encourages safe working),⁸ and there can be little doubt that workers represent a critical resource of knowledge and experience.

Community

In recent years, in developed countries, large work sites have been subject to local pressure on environmental matters – replacing or supplementing the welcome for employment with a suspicion about adverse impacts from emissions, traffic movements and similar issues. Although the practice of occupational hygiene focuses on worker exposures to health risks, many of the techniques employed to evaluate environmental risks are extensions of occupational hygiene practice. Measuring airborne dust inside a plant may use similar sampling technologies to those required for dust emission monitoring.

However, having an effective technical approach to monitoring the efficacy of environmental protection measures is only part of the response required. The communities which may be affected by the workplace have to be satisfied with the environmental controls. That means that communication has to take place with people with their own interests and needs, which may be quite different from those of the business. Environmental reporting requires a different style, with a different target audience compared with the recipients of occupational hygiene reports. In addition, the environmental standards are different to occupational standards, often adopted to protect amenity or animals and plants rather than human health.

Health-care system

Efforts to relate the damage to workers' health to the financial performance of their employers is intrinsically easier and ultimately more likely to succeed in those countries where employers bear a large burden of the cost – directly, or through

social and commercial insurance. In the UK, the National Health Service incurs most of the costs associated with health maintenance and ill health treatment, and the benefits system largely pays for invalidity. This means that the Government has a direct, although complex and difficult to fully analyse, incentive to reduce occupationally related ill health. To date little evidence of this has been recognized, and it continues to be dominated by treatment rather than prevention.

Regulators

Government bodies, including enforcement authorities, are increasingly being driven towards measuring and monitoring their own performance. This can take the form of statistical tables showing visits and inspections, notices issued and prosecutions taken, but each of these is an 'input' measure. Although not directly under their control, in practice the outcome measures are the accidents and ill health that arise at work. This is encouraging the statutory authorities to set targets for improvements – a process greatly stimulated by policies pursued under the first Clinton administration in the USA, and widely imitated and developed in many other countries. Targets to reduce the incidence of occupational disease, to reduce the numbers of working days lost to work-related sickness absence, to increase the rate of rehabilitation of injured and ill workers – these all reflect a growing sense in which the authorities are stakeholders in the performance of every enterprise.⁹

Formal management systems

It is important to understand the context within which each organization operates, the technical content of an occupational hygiene programme and the stakeholders who depend upon and need to participate for the programme's success. However, the result of such understanding could be an incoherent approach to occupational hygiene unless it is embedded within a systematic approach to managing health risks. This explains why Occupational Health and Safety Management Systems (OSHMS) have developed through national and

international cooperation. Such systems have various origins – the European Union Framework Directive (1989/391), industrial sector developments such as in the chemical and petroleum industries.

The publication of International Labour Organization (ILO) guidance¹⁰ represented the most significant step in establishing a global approach, which has been adopted by major countries, including China. Embodying health and safety arrangements for an enterprise within a formal, documented and auditable system removes the potential arbitrariness of processes developed by a few individuals.¹¹

Effective OSHMS include the following elements.

- *Policy.* A Mission Statement for occupational health (and safety) that displays the organization's commitment and vision. This should create a framework for accountability, which clearly incorporates and is led by Directors and senior managers.
- *Planning.* Effecting the key processes of hazard identification, risk assessment and risk control. Plans will also include emergency preparedness and response, with identification of legal and other standards which apply. The enterprise should set long-term occupational safety and health (OSH) objectives, plan the management targets and actions to achieve them, and break these down into milestones that facilitate monitoring progress.
- *Organizing.* By clearly defining the organizational structure and allocating responsibilities to employees, the crucial link to operational control of risks can be established. At each level, competence of workers with assigned duties should be addressed.
- *Worker participation.* Workers, often through their representatives, represent a key resource capable of making a valuable contribution to occupational health risk management. Consultation with the workforce is a minimum requirement to mobilize that resource.
- *Communicating.* From basic information and work procedures to the details of the system itself, in two-way communications.
- *Implementing and operating.* This is the heart of the system, and what really matters is the practical action taken to protect health. Making the system

come to life requires implementing the management processes and plans. These include practical activities such as designing, constructing, commissioning, maintaining and testing local exhaust ventilation. Above all, it requires the ventilation system to be used; safe working practices are critical to exposure control.

- *Measuring performance.* From data on work-related ill health and diseases to exposure monitoring. Formal audits are used to evaluate the overall performance of the system.
- *Corrective and preventive actions.* Fundamental to the formal system is a consistent approach to identifying opportunities to reduce exposure to health risks, including the investigation of work-related ill health. Inspections and checks are used to identify system non-compliances and to correct them, and to seek ways in which adverse outcomes may be prevented.
- *Management review.* Evaluates the appropriateness of the overall design of the system and its objectives in the light of performance achieved. This is an opportunity to check whether the system is meeting the expectations of the stakeholders.
- *Continual improvement.* Fundamental to an OSHMS is a commitment to improve the management of health risks, so that work-related ill health is reduced (effectiveness) and/or the system achieves good health status with less resource (efficiency).

Many OSHMS have been published over the past 20 years. Some reflect the interests of the sponsoring bodies, for example the American Industrial Hygiene Association system places the industrial (occupational) hygienist centre stage as the crucial competent person.¹² Others such as the International Safety Rating System (ISRS) were developed so that commercial organizations could offer third-party certification. The ILO is a UN agency that influences the development of labour laws across the globe. Its guidance is authoritative and the publication of the Occupational Safety and Health Management System Guidelines in 2001 (following detailed review of over 20 management systems world-wide conducted on ILO's behalf by the International Occupational Hygiene Organization – IOHA) established an international specification. It reflects globalization of

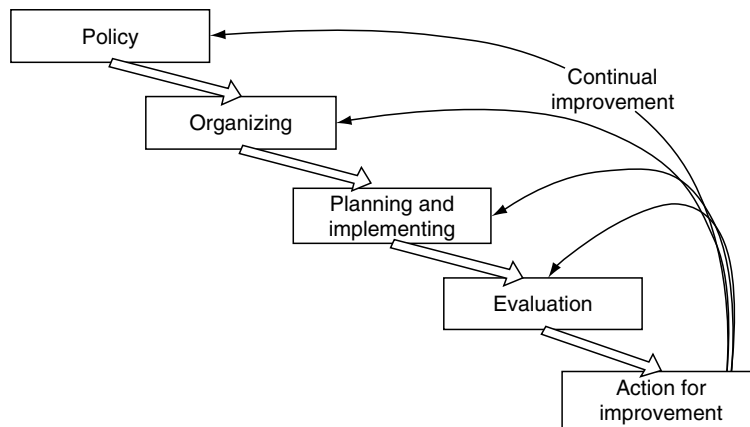


Figure 32.3 ILO system diagram.

organizations and the increase in outsourcing and partnering – an indication of how systems need to continually evolve to reflect new ways in which organizations manage activities (Fig. 32.3).

Continual improvement

Continual improvement is vital if management systems are to be effective and efficient, particularly for organizations operating in a changing environment. The target for occupational health and safety systems is to reduce accidents and occupational ill health – so intrinsically the arrangements have to be designed to progressively improve the situation. That is why continual improvement has to be the first and foremost characteristic of a formal OSHMS.

Continual improvement in an OSHMS can have four different aspects:

- 1 results that are better year on year, as measured by falling rates of ill health, sickness absence or other indicators such as reducing exposure levels;
- 2 achieving acceptable results, but more efficiently with fewer resources;
- 3 improvements in the system itself;
- 4 culture change driven by results, which represents a breakthrough in performance to a new state of effectiveness and efficiency.

The formal mechanism for improvement is the use of audits and the actions they prompt. This is the ‘Check–Act’ part of Plan–Do–Check–Act. In addition to formal audits there are other mechanisms available – ranging from the equiva-

lent of quality circles, where teams of workers and their managers look for ways to make work healthier, to discussions with customers or suppliers on where there may be opportunities to improve.

One way of leading the search for improvement is to set high-level objectives for the system, and then ensure that activities within the OSHMS coherently support the meeting of those objectives. Appropriate objectives include:

- clear policy-making with written commitment to good, clearly defined standards at the highest level in the organization, supported by visible leadership, involvement and regular monitoring of performance;
- employment of or access to competent personnel, with adequate resources and time for their training and development;
- effective arrangements for involvement of, and consultation with, key stakeholders such as: employees, trades unions, customers, suppliers, regulators and other statutory consultees, partners and neighbours;
- ensuring purchased materials, equipment and services are selected using appropriate occupational health criteria as well as price;
- ensuring technical and operational data and records are available, updated and retained as necessary to meet changing business needs and regulatory requirements;
- regular monitoring of all parts of the OSHMS by those responsible for comparing actual performance with expected results and/or goals;

- in addition to monitoring, a system for pre-planned audits verifying how effective the OSHMS is in practice;
- systems for identifying and reporting instances of failure to meet required standards, including external reporting where required, investigation of root causes, with corrective actions applied to improve the OSHMS and prevent repeated failures;
- emergency systems, including plans and competent people to implement them, for containing and controlling serious system or business failures and minimizing effects on all those involved.

Auditing/verification

Auditing means competent, independent people sampling a business process and reporting on its effectiveness, with emphasis on the inputs, outputs and testing of internal controls. 'Verification' has a similar meaning to 'audit', but verification is typically by an accredited body to an external, recognized standard. If this results in the issue of a compliance certificate, it is 'certification'. Key features are:

- Auditor independence, which gives the audit findings credibility.
- Company procedures and local documents are sampled and evidence is collected to ensure that operational practice is consistent with documented practice, i.e. 'Say what you do; Do what you say you do; Prove it'.
- Because of sampling, however well judged, no audit results in a 'perfect' view of true facts; also findings are truly valid only at the time of audit. Audit evidence, though very powerful, should be reviewed together with other data on system performance when planning improvements.
- Evidence is collected from documents, interviews and inspections of workplaces. Auditor(s) should select some samples, not just accept what is put before them.

Linking health and hygiene management to safety, environment and other issues

When any organization develops a formal and systematic approach to the management of occupational health and hygiene, it soon becomes ap-

parent that the approach taken has many similarities to the management of quality, environment and particularly safety. This applies whether the management is compliant with one of the published occupational health and safety management systems or has developed the system from first principles in a manner unique to the enterprise.

Similarities

Management systems are constructed on a common framework – planning, doing, checking and acting on the checks. Thus, there are significant overlaps between occupational health management and other topic areas such as quality and environment. Management systems are increasingly harmonizing, and standard-setting bodies encourage this. For example, a substantial element of the guidelines on occupational health and safety management systems published in 1996 by the British Standards Institute¹³ is based on the earlier international standard on environmental management.¹⁴ Deciding how health is to be managed, developing a policy and an organizational structure with assigned responsibilities and authorities to implement that policy – these are all common elements for any management system. This means that many of the practical techniques that are developed for one, such as consultation and communication, may be used for others, and can also result in some of the same people being involved and bringing their experience to bear. Many enterprises have found that staff who had been instrumental in developing quality management systems are very helpful in introducing a formal approach to health and safety management.

Differences

There are crucial differences, however, between the targets for each system. Quality management is rarely a legal compliance issue, whereas in occupational health this is often the starting point for any systematic approach. Because failure to achieve legal compliance places the organization at risk, the timescale for implementing new control measures may be under considerable pressure. Finally, despite the overlaps, there are specific

technical skills that are required to effectively manage occupational health risks – occupational hygiene, engineering and medical competencies are clearly central to any programme.

Health and hygiene in the boardroom

Managing business risk

Organizations are taking a holistic approach to risk management, seeking to remove or minimize the effects of risks intrinsic to the organization's activities and those engendered by speculative actions. This is developed in a logical sequence as shown in Fig. 32.4.

Corporate social responsibility

The business community has been rocked by periodic scandals, from speculative booms and busts such as the eighteenth century South Sea Bubble¹⁵

and the more recent dot com years to corporate fraud and mismanagement such as Mirror Group Newspapers, BCCI, Polly Peck, Enron and WorldCom. This has recently fuelled the growing movement for companies to make a clear commitment to socially responsible behaviour, and to use techniques of auditing and reporting to verify their performance. Health and safety is a critical area, as a European consumer survey has demonstrated. There is evidence that occupational health risk management is also being considered by investors, with consideration of ethical values and sustainability having an impact beyond the formally 'ethical investment' trusts.¹⁶ Public companies are being pressed to conform to a new model – whereby they set ethical standards for social responsibility, which includes looking after the health of their own workforces – and to publish the results of their efforts including the ill health levels (working days lost through illness) as a mark of transparency. This is a power-

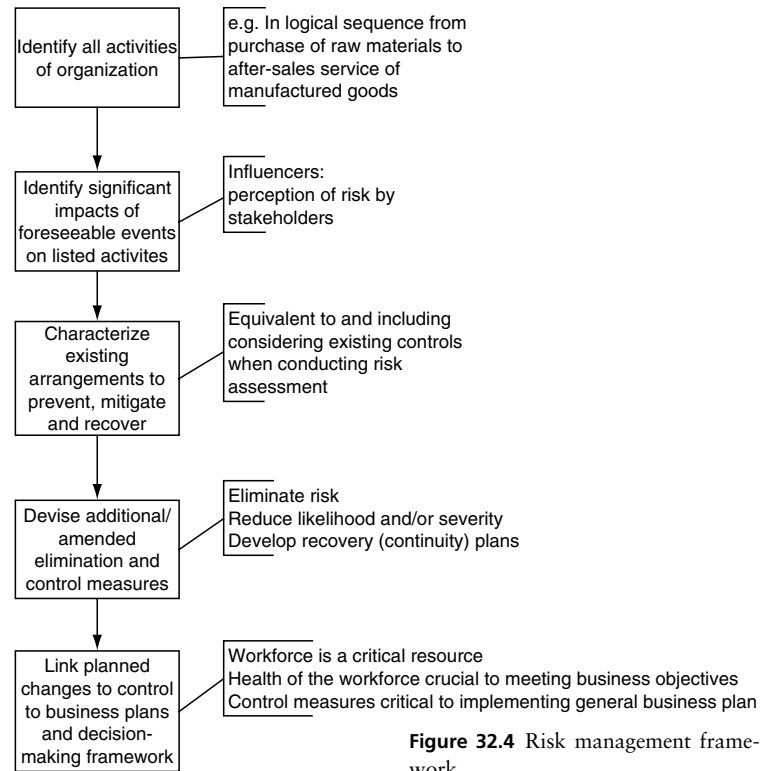


Figure 32.4 Risk management framework.

ful driver that occupational health professionals need to respond to, helping the organizations meet the growing expectations of their stakeholders.

Setting organizational objectives and targets

With risk management developed as an integrated practice, and the results in terms of gains and claims being published, it is clear that for the enterprise to seize the agenda it has to have unambiguous targets to which it aspires and is working. These need to be in the form of SMART objectives – that is they need to be ‘specific’ (not vague and woolly, which would prevent certainty as to their achievement), ‘measurable’, ‘agreed’ (with the management team and, through consultation, the workforce as a whole), ‘realistic’ (to avoid demotivated, disappointed people who work hard for something that could not be attained) and ‘time-tabled’ (indicating when the target, or milestones, is to be met en route).

Leadership

In general, organizations are shaped by their history, their context and their leaders. How the enterprise or institution has developed over time will embed certain cultural characteristics, and the expectations of stakeholders – staff, customers, neighbours, regulators – will have a huge influence on how it operates currently. However, the example set by directors and senior managers should not be underestimated – they can encourage or undermine a sound approach to minimizing the risks to health. There are many examples of such claims for the role of leaders within organizations, including its primacy in the Business Excellence Model (©EFQM)¹⁷ (Fig. 32.5).

Occupational health and hygiene issues therefore need to be explained and communicated at board and senior management level, and also to form part of the leadership’s sets of business plans and targets. To have occupational health objectives as an integrated component within business planning represents a critical step in ensuring that the

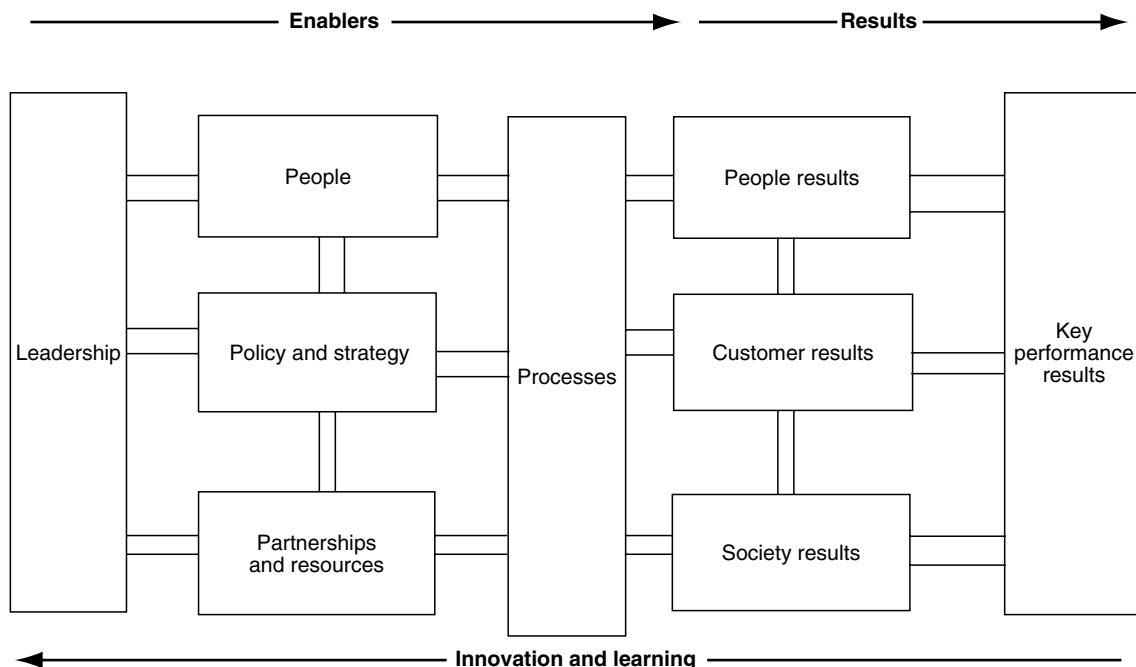


Figure 32.5 Business Excellence Model.

whole organization commits resources to protecting health. Such objectives could be:

- measurable annual reductions in rates of sickness absence;
- reduction in reported occupational disease/ill health;
- measurable increase in attendance at health-related briefings and training courses;
- reduction of exposures to hazardous substances to no greater than 50% of published occupational exposure limits;
- development of staff 'climate survey' and its annual use to evaluate culture and related psychosocial issues;
- achieving an agreed percentage of safe behaviours.

The future

There are some discernable trends that may be taken into account by occupational hygiene practitioners when developing policies and approaches within their organizations; how far each of them will go and at what pace is open to speculation.

- Formal management systems are set to become more common, in every region. More countries will join Norway and China in making them obligatory.
- Occupational health and hygiene will gain in prominence in developed countries as the nature of work changes and accident risks are reduced.
- Health risks will be increasingly identified as an important aspect of business risk management, and a necessary component of corporate social responsibility.
- Governments will encourage these developments as part of the strategy to reduce the costs of occupationally related ill health, and to improve access to work for people with disabilities and older people.
- Occupational hygiene is a developing professional practice that will operate in multidisciplinary teams with safety practitioners, occupational physicians, nurses, ergonomists, psychologists, engineers and managers.

How occupational hygienists participate in and even lead such developments will depend largely

on their individual abilities. Critical will be the willingness of the whole profession to recognize that the technical components of hygiene practice are vital, but that without management systems they are limited to local and immediate impacts on occupational exposures. Understanding and developing management systems with a significant health protection component is a crucial part of securing the well being of whole working populations.¹⁸

Notes

1 Declaration of Occupational Health for All, Meeting of WHO Collaborating Centres in Occupational Health, Beijing, China, October 1994. This declaration was based on Health for All, adopted as the cornerstone of WHO strategy in 1977, updated as Health for All in the 21st Century, May 1998.

2 Internal Control: Guidance for Directors on the Combined Code (Turnbull Report), 1999, Institute of Chartered Accountants in England and Wales, London.

3 Health and Safety Indicators for Institutional Investor, Health and Safety Executive, CRR.

4 MORI Poll reference.

5 Guidelines on Setting In-House Occupational Exposure Limits for Airborne Therapeutic Substances and their Intermediates, ABPI.

6 Ventilation and Air Conditioning, Guide B2, CIBSE, 2001, SBN 1 903287 16 2.

7 Guide to Occupational Health and Safety Management Systems, BS 8800: 1996, BSI London. Annex D provides a very helpful guide to risk assessment, which has been drawn upon for this discussion.

8 Reference regarding TUC Research.

9 Securing Health Together, MISC 225, HSE 2000, represents a long-term (10-year) occupational health strategy for England, Scotland and Wales, which set targets to reduce incidence of work-related ill health and help people enter/re-enter the workforce despite ill health or disability or following illness.

10 International Labour Office. *Guidelines in Occupational Safety and Health Management Systems*. ILO-OSH 2001. ILO, Geneva.

11 The International Standards Organization (ISO) has published guidelines that encourage coherence between management systems. ISO Guide 72:2001 Guidelines for the Justification and Development of Management System Standards includes information on guidance for: justifying and evaluating a proposed management system standard project with a view to assessing market relevance; the methodology (process) of developing and maintaining (i.e. reviewing and revising) management system standards with a view to ensuring compatibility and enhancing alignment; and the terminology, structure and common elements

of management system standards with a view to ensuring compatibility as well as enhancing alignment and ease of use.

12 American Industrial Hygiene Association. (1996). *Occupational Health and Safety Management System – An AIHA Guidance Document*. AIHA, Virginia.

13 Guide to Occupational Health and Safety Management Systems BS 8800:1996, BSI 1996.

14 Environmental Management Systems: Specification with Guidance for Use, BS EN ISO 14001:1996, BSI 1996.

15 The bursting of the South Sea Bubble in 1720 represented the first stock market collapse (a very readable account may be found in *A Very English Deceit: the Secret*

History of the South Sea Bubble by Malcolm Balen, Fourth Estate 2002).

16 Health and Safety Indicators for Institutional Investors, Mark Mansley, HSE 2002.

17 European Foundation for Quality Management, EFQM, has published a full guide to this model on the website www.efqm.org.

18 There is extensive published material on management systems. Useful reviews include: Frick, K. *et al.* (2000). *Systematic Occupational Health and Safety Management – Perspectives on International Development*. Pergamon; Noble, M.T. (2000). *Organisational Mastery with Integrated Management Systems*. Wiley.

Index

Note: page numbers in bold refer to tables, those in *italic* refer to figures.

- absorption coefficient 234–5
- accident investigation 412–13
- accident prevention 403–17
 - errors and violations 413–16, **416**
 - behavioural safety 414–15
 - classification of human errors **413**, *414*
 - error reduction strategies 414
 - safety culture 415–16
 - legislative controls of safety 404
 - measurement of safety performance
 - key performance indicators 409
 - leading and lagging performance indicators 409
 - proactive and reactive monitoring 408–9
 - reactive safety management 405–6, *405*
 - risk assessment 409–812
 - consequences 410–11
 - continuing hazards 410
 - establishing objectives 410
 - failure hazards 410
 - failure mode and effect analysis 412
 - HAZOPS studies 411–13
 - system boundaries 410
 - techniques 410
 - tolerability/acceptability of risk 411
 - safety management systems 406–8, *407*, *408*
 - workplace safety management 404
- accommodation 273
- acid anhydrides, occupational asthma 51
- acinus 14
- acoustic pressure 223
- acoustic trauma 60–1
- acoustics 222–4
- acrylamide neurotoxicity 55
- acrylic monomers, contact dermatitis 33
- activated charcoal 91, 211–12
- acute toxicity 68–9
- ad hoc records 172–3
- adaptation 273
- adsorption 91–2
- aerosol sampling
 - bioaerosols 203–4, *205*
 - criteria 186–8, *187*
 - errors 198, 201–2
 - chemical 202
 - flow rate setting and control 201
 - gravimetric 201–2, *201*
 - performance of sampling head 201
 - variability of exposure 202
 - fibrous aerosols 202–3, *203*
 - investigational instruments 195–8, *197*, **199–200**
 - multifraction samplers 195, *196*
 - personal samplers 192, 193–4, *193*, *194*, 195
 - respirable aerosols 193–5
 - static (area) samplers 192–3, 194, 195
 - strategy 189
 - system components 189–90, *189*
 - filters 190–1
 - pumps 191–2
 - sampling head 190
 - size selectors 190
 - transmission section 190
 - thoracic aerosols 195
 - two-phase compounds 204–5
- aerosols 92–3, *92*, 185–207
 - biological 93, 188, 203–4
 - elementary particle size 94–6
 - evolution of 93
 - fibrous 188
 - generation of 93
 - inhalable fraction 186
 - interaction with electromagnetic radiation 102–3, *104*
 - particle shape 93–4, *94*
 - particle size 94
 - respirable fraction 186, 187
 - sampling *see* aerosol sampling
 - thoracic fraction 187
- after-work creams 35
- age, and lung function 22
- agricultural workers, health hazards 344–50
 - anthrax 349–50
 - bovine tuberculosis 349
 - brucellosis 345
 - cowpox 346–7
 - diarrhoeal disease 344–5
 - erysipelothrix 347

- hydatid disease 348
leptospirosis 350
Lyme disease 347–8
orf 346
psittacosis 349
Q fever 345–6
ringworm 347
Streptococcus suis 348–9
- air cleaning 458
air conditioning 303
air density 440–41
air ducts 452
 sizing and loss calculation 453
air movement 303
air pollution 87
air temperature 298
air-conditioner disease 54
airborne contaminants 85–104
 aerosols 92–3
 elementary particle size 94–6
 evolution of 93
 generation of 93
 particle shape 93–4, 94
 particle size 94, 95, 96
 gases and vapours 86–7
 adsorption 91–2
 density 88
 diffusion 90–1
 humidity 88–9
 ideal gas laws 89
 partial pressure 89–90
 vapour pressure 87–8
 interaction with electromagnetic radiation
 aerosols 102–3, 104
 gases 103
 particle motion 96–102
 aspiration 99–101, 100
 diffusion 101–2, 101
 drag force 96–7, 97, 98
 elutriation 99
 gravitational force 97–8
 impaction and interception 99, 100
 physical properties of matter 86
airflow measurement 442
airways 13–14, 14
 cross-sectional area 15
 disease 50–3
 irritation 49–50
 lining 16–17, 16
aldehydes, occupational asthma 51
aliphatic polymamines, contact dermatitis 33
allergens 54
allergic contact dermatitis 33–4, 33, 34
alpha radiation 331
 properties of 333
altitude
 flying at 66
 working at 65
aluminium, occupational asthma 51
alveolar clearance 49
alveolar macrophages 16
alveolus 15–16, 15
ambient saturation concentration 88
aminothiazole, contact dermatitis 33
ammonia
 airways irritation 23, 50
 contact dermatitis 33
ammonium persulphate, contact dermatitis 33
amylase, occupational asthma 51
Andersen microbial sampler 204, 205
anemometers 301
animal experiments 72–3
animal hair in airways irritation 23
animal handler's lung 54
anthrax 349–50
anthropometric data 384
antibiotics, occupational asthma 51
antimony, neurotoxicity 55
apocrine glands 28
arousal curve 365
arsenic, neurotoxicity 55
arsine, cardiovascular toxicity 58
as low as reasonably practical (ALARP) 120
asbestos 152
 airways irritation 23
 malignant disease 53–4
 respirable fibres 188
asbestosis 53
aspiration 99–101, 100
asthma 51–2
atomic structure 329–30
attitudes 362, 363
attributable risk 176
audiometric notch 60
auditing 487
aural comfort 234
axial flow fan 455–6, 455
- back pain 38–45
 biopsychosocial approach 41–2
 epidemiology 42
 occupational management 42
 physical treatments 43
 primary prevention 43

- back pain (*cont'd*)
questionnaire 379
'red flags' 40
secondary prevention 43
tertiary prevention 43–5
triage 39
'yellow flags' 40–1
- background noise 226–7
- background radiation 342
- bagassosis 54
- balance/vision defects 55, 56–7
- Bangladesh, child labour 7
- barium, airways irritation 23
- barley, airways irritation 23
- barotrauma 64–5
- barrier creams 34–5
- becquerels 335
- behaviour 480
safety-related 414–15
stress effects on 367
type A pattern 363
- bel scales 225
- benchmark dose 74
- benzene
cardiovascular toxicity 58
narcosis 55
- beryllium, airways irritation 23
- beta radiation 331–2
properties 333
- bioaerosols 93, 188
sampling 203–4, 205
- biological effect monitoring 161
- biological exposure indices 110
- biological hazards 344–58
agricultural workers 344–50
anthrax 349–50
bovine tuberculosis 349
brucellosis 345
cowpox 346–7
diarrhoeal disease 344–5
erysipelothrix 347
hydatid disease 348
leptospirosis 350
Lyme disease 347–8
orf 346
psittacosis 349
Q fever 345–6
ringworm 347
Streptococcus suis 348–9
health-care workers 352–7
blood-borne viruses 352–3
hepatitis B 353
hepatitis C 353–4
HIV 354
tuberculosis 355, 356
microbiology laboratory workers 355–7
occupational travellers 350–2
hepatitis A 352
malaria 351
travellers' diarrhoea 351–2
- biological monitoring 160–9
choice of container 164–5
contact with laboratory 165
definitions and terminology 160–1
guidance values 110
indications for 161
interference 167
interpretation of results 165
notification of results 167–8
reference values 164, 165–6
samples
blood 162–3
breath 163
fat 163
hair and nail 163
urine 161–2
specificity of metabolites 166
staff training 167
storage of results 168
timing of sample collection 163–4
units of expression 166–7
- biological tolerance values 166
- bird fancier's lung 54
- birth certificates 172
- blood samples 162–3
- blood–gas barrier 16
- blood–gas transport 22
- body map 378
- body plethysmography 21
- Borrelia* spp. 347
- botsball 291
- bovine tuberculosis 349
- breath samples 162, 163
- bremsstrahlung radiation 332
- British Standard *see* BS
- bronchioles 13
- brucellosis 345
- BS 6841 261
- BS 6842 262–3
- bullying 365
- business risk 474–5, 488
- byssinosis 23, 52

- cadmium oxide, airways irritation 23
cancer 367
candela 269
canister samplers 210, 217
capture velocity 446, 448, 450–451
carbamate pesticides, neurotoxicity 55
carbon dioxide narcosis 55
carbon disulphide
 as extraction solvent 218
 mental changes 55
 neurotoxicity 55
 parkinsonism 55
carbon monoxide
 end-exhaled air, reference value 164
 narcosis 55
 parkinsonism 55
carbon tetrachloride narcosis 55
carboxyhaemoglobin, blood reference value 164
carcinogenicity 69, 76
cardiovascular disorders 366–7
cardiovascular system toxicity 58
career development 364
carpal tunnel syndrome 254
case management 482
case–control studies 156, 178
centrifugal fan 456–7, 456, 457
cerebral arterial gas embolism 65
charcoal tubes 218
cheese washer's lung 54
child labour 7–8
chilling temperature 297
Chlamydia psittaci 349
chlorinated naphthalenes, epilepsy 55
chlorine, airways irritation 23, 50
chloroform narcosis 55
chronic obstructive pulmonary disease 52
CIP10 personal sampler 194
classification of hazardous substances 113
clothing insulation 288, 297
coal, airways irritation 23
coal-miners' pneumoconiosis 53
coated sorbents 214, 220
cobalt, airways irritation 23
cobalt chloride, contact dermatitis 33
cognitive ergonomics 374
cohort studies 73, 153
cold conditions 303, 304–5
 effects of 63
 hypothermia 64
cold injuries
 freezing 63–4
 non-freezing 63
cold stress indices
 required clothing insulation 297
 still shade temperature 296
 wind chill 296–7
cold trap samplers 211
colophony, occupational asthma 51
colour rendering index 269, 277
comfort indices
 analytical 295–6
 direct 296
 empirical 295
Commission Regulation No. 1488/94 107
communication of risk 111
composite partitions 237
computer modelling 375
conduction 287–8
confounding factors 177
 control of 179
continuing hazards 410
continuous active samplers 210–15
 calculations 214–15
 coated sorbents 214, 220
 cold traps 211
 sampling bags 211
 sampling train 214
 solid sorbents 211–13
 sorbents 210–11
 thermal desorption 213–14, 219–20
 wet chemistry and spectrophotometry 214
Control of Asbestos at Work Regulations (1985) 106
Control of Major Accident Hazard (COMAH)
 Regulations (1999) 106
control programmes 479–80
Control of Substances Hazardous to Health
 (COSHH) Regulations (1989) 106, 465
convection 288–9
convective diffusion 102
coping strategies 362–4
corporate social responsibility 475, 488–89
corrected effective temperature 291, 292
cost–benefit analysis 110
cost–benefit assessment 110
cotton dust, airways irritation 23
Council Regulation No. 9793/93 107
cowpox 346–7
cross-sectional studies 73
cumulative distribution functions 132
cumulative exposure 146
Cunningham correction factor 97
Dalton's partial pressure law 90
Dangerous Substances Directive (67/548/EEC) 77, 113

- daylight 277–8
daylight factor 269, 278
death certificates 171–2
decay constant 330
decibel scale 226
decompression illness 65
degreaser's flush 167
density
 air 440–1
 gases and vapours 88
dermal exposure assessment 389–98
 conceptual model 391–4, 392
 dermal uptake 390–91
 measurement of exposure 394–6, 394
 modelling exposure 396–7
 risk evaluation 397–9
dermis 27
diabetes 367
diarrhoeal disease
 agricultural workers 344–5
 travellers' diarrhoea 351–2
diffusion 22, 49
 gases and vapours 90–1
 particles 101–2, 101
diffusive samplers 215–18, 220
 bias 216
 diffusive uptake rate 215–16
 factors affecting
 air velocity 217
 humidity 216
 pressure 216
 temperature 216
 transients 217
 principles of 215
dilution ventilation 451
 volume flow rate for 451–2
Directive 93/67/EEC 107
disability 37
Disability Discrimination Act (1995) 45
disability glare 275
discharge lamps 276–7
discomfort glare 275–6
Display Screen Equipment Regulations (1992) 377
dose index 146
dose–response assessment 115–16, 117
dose–response curves 75, 116
downsizing 6
drag force 96–7, 97, 98
dry bulb temperature 292, 299, 300
dry resultant temperature 296
dusts 23
 aerosols 92
 ear defenders 468–9
 ear protection 246–9
 earmuffs 246, 247
 earplugs 246
 received noise level 249
 earmuffs 246, 247
 earplugs 246
 eccrine glands 28
 ED₅₀ 116
 effective temperature 291, 292
 effusion 90
 electromagnetic fields 8
 electromagnetic radiation 307–27
 exposure and emission standards 317–18
 lasers 323–6, 323
 classification 324–5, 325
 hazards 324
 measurements 326
 standards 325–6
 uses 324
 low frequencies 309–15, 310
 control 313
 low-level exposures 312–13
 measurements 313–14
 static electric fields 311
 static magnetic fields 311–12, 312
 optical 318–23
 biological effects 320–1
 control 322
 measurements 322–3
 physical interactions 319–20
 standards 322
 radiofrequencies 314–17, 314
 control 317
 low-level exposures 315–17
 physical interactions and physiological effects
 315, 316
 uses of 310
 electromagnetic spectrum 268–9, 269, 308
 elutriation 99
 emergency lighting 280
 emphysema 23
 enclosures 446, 448
 enzymes, causing occupational asthma 51
 epidemiology 132–4, 170–81
 causation or association 176
 data sources 171–3
 definition 170–1
 exposure 173
 field studies 181
 goals 177
 cost 178

- precision 178
- validity 177
- health outcome 173–6
 - measures of frequency 174–5
 - measures of occurrence 173–4
 - measures of risk 176
 - time, place and person 175–6
- options 179
 - control of confounding factors 179, 180
 - cross-sectional studies 178
 - data handling 180
 - longitudinal studies 178–9, 180
 - subject allocation 179–80
 - timing 178
- type of study 180–2
- uses 171
- epidermis 25–7
- epilepsy 55
- epoxy resins, occupational asthma 51
- equivalent continuous sound level 230
- ergonomics 373–88
 - control measures 382–7
 - tool design 386–7
 - work design/pacing 387
 - workplace changes 387
 - workplace design and human variability 382–5
 - workplace layout 385–6
 - current issues 376–7
 - definition of 374
 - ill health and injury data 378
 - methodologies 374–6
 - and occupational hygiene 374
 - problem evaluation 378–82
 - numbers involved 378–9
 - workplace evaluation 378–82, 380, 381
 - risk assessment 377–8
- errors and violations 413–16, 416
 - aerosol sampling 198, 201–2
 - behavioural safety 414–15
 - classification of human errors 413, 414
 - error reduction strategies 414
 - safety culture 415–16
- erysipelotheix 347
- Estimation and Assessment of Substance Exposure (EASE) 119, 397
- EU *see* European Union
- EU Machinery Safety Directive 260–1, 264
- EU Physical Agents Directive 261, 264–5, 265
- European Inventory of Existing Commercial Chemical Substances (EINECS) 109, 113
- European Union (EU) 8
- evacuated flask samplers 210
- evaporation 289
- evaporative cooling 303
- event tree analysis 108
- exposure assessment 116–19
 - exposure modelling 119
 - exposure variation 118–19
 - sampling and analysis methods 118
 - workplace 117–18
- exposure determinants 134–9
 - control for variability 138–9
 - experimental studies 135
 - multiple linear regression analysis 136–8, 138
 - observational studies 135–6
 - selection of 134–5, 135
- exposure limit values 265
- exposure limits, compliance with 130–1
- exposure measurement surveys 124–44
 - compliance with exposure limits 130–1
 - determinants of exposure 134–9
 - epidemiological studies 132–4
 - exposure variability 125–8, 127, 128
 - hazard control 131–2
 - mixtures 139–40
 - non-detectable exposures 129
 - priority setting 125
 - risk assessment models 140–2
- exposure measures 173
- exposure modelling 119
- exposure variability 125–8, 127, 128, 150–1
 - interindividual 150–1
- exposure variation 118–19
- extent of risk 106
- exterior lighting 280–1
- external radiation 334
- extraction ventilation 445–6, 447
- extremely low frequency radiation 309, 310
- extrinsic allergic alveolitis 53
- eye and vision 270–6, 270
 - accommodation 272
 - acuity 272–4
 - adaptation 272
 - construction of eye 270–1
 - optical radiation effects 320–1
 - retinal fatigue 271
 - sensitivity of eye 271
 - visual environment 274–6
 - visual fatigue 274
 - visual perception 271–2, 273
 - visual performance 274
 - visual task 274

- factor analysis 140
failure hazards 410
failure mode and effect analysis (FMEA) 115, 412
fan static pressure 454
fan total pressure 454
fan velocity pressure 454
Fanger analysis 295–6
fans 454–8
 axial flow 455–6
 centrifugal 456–7, 456, 457
 characteristic curves 454–5
 matching to system 457–8, 458
 power and efficiency 454
 propeller 455
far-field conditions 308
farmer's lung 54
fat samples 162, 163
fatigue-decreased proficiency limit 265
fault tree analysis 108
Feret's diameter 94
fibrous aerosols 188
 sampling 202–3
Fick's law of diffusion 90–1
field surveys 181
filament lamps 276
film badges 337
financial context 5
fixed dose procedure 74
Fletcher method 449–10
flexible plastic container samplers 210
flour, occupational asthma 51
follow-up studies 178–9
forced vital capacity 20, 21
Framework Directive on the Introduction of
 Measures to Encourage Improvements in the
 Safety and Health of Workers at Work 426
frequency measures 174–5
frequency-of-use principle 386
frostbite 64
frostnip 64
fumes 93
function principle 386

gamma radiation 332
 properties 333
Garrison method 450
gas toxicity 65
gas transfer 21–2
gas and vapour sampling 208–21
 continuous active samplers 210–15
 calculations 214–15
 coated sorbents 214, 220
 cold traps 211
 impinger 214
 sampling bags 211
 sampling train 214
 solid sorbents 211–13
 sorbent tubes 214–15
 sorbents 210–11
 thermal desorption 213–14, 219–20, 219
 volume fraction 215
 wet chemistry and spectrophotometry 214
diffusive samplers 215–18
grab samplers 209, 210
 evacuated flasks 210
 flexible plastic containers 210
 passivated canisters 210, 217
measurement techniques 208–9
practical applications 218–20
quality systems and quality control 221
sampling devices 208–9
gases and vapours 23, 86–7
 adsorption 91–2
 density 88
 deposition/uptake 49
 diffusion 90–1
 humidity 88–9
 ideal gas laws 89
 interaction with electromagnetic radiation 103
 partial pressure 89–90
 sampling *see* gas and vapour sampling
 vapour pressure 87–8
gastrointestinal disorders 365
 see also diarrhoeal disease
Geiger–Müller tube 336
genotoxic potential 74–6
genotoxicity 69
glare 275–6
glare index 269, 276
global trends 6–9
globe thermometer 298–9
grab samplers 209, 210
 evacuated flasks 210
 flexible plastic containers 210
 passivated canisters 210, 217
Graham's law 90
grain dust, occupational asthma 51

hair 27
 samples 162, 163
hand–arm vibration syndrome 45, 253
hand-transmitted vibration 252–61
 causes of 257
 EU Machinery Safety Directive 260–1

- EU Physical Agents Directive 261
- musculoskeletal effects 255–6, 257
- national and international standards 258–60
- neurological effects 255
- prevention of 257–8
- vibration-induced white finger 253–5
- harassment 365
- hardiness 362–3
- hay, airways irritation 23
- hazard banding 79
- hazard control 131–2, 433–9
 - after design stage 434–7
 - elimination 434
 - isolation or segregation 436
 - maintenance and housekeeping 436–7
 - process changes 435–6
 - substitution 434–5, 425
 - ventilation 436
 - at design stage 433–4, 434
 - education and training 437
 - personal protective equipment 374, 437–8
 - source, transmission and the individual 438–9, 438
- hazard elimination 434
- hazard identification 77, 106, 112–14, 410, 476
 - continuing hazards 410
 - failure hazards 410
 - occupational hygiene instrumentation 115
 - organizational hazards 423–4, 423, 425
 - psychosocial hazards 363–4, 423–4, 425
 - toxicological basis of 112–14
 - workplace hazards 114
 - see also* biological hazards
- hazard and operability studies (HAZOPS) 114, 411–2
- health hazards 476–7
- health and hygiene team 483
- health outcome 173–6
- Health and Safety at Work Act (1974) 106, 110, 338, 426
- health surveillance 161, 482
- health-care workers, health hazards 352–6
 - blood-borne viruses 352–3
 - hepatitis B 353
 - hepatitis C 353–4
 - HIV 354
 - tuberculosis 355, 356
 - Varicella zoster 354
- hearing assessment 240–3
 - pure tone audiometer 240–3, 241, 242
 - standard hearing 240
 - test conditions 241
- hearing conservation 244–5
- hearing mechanism 59–60
- auditory ossicles 59
- cochlea 60
- hair cells 60
- heat cramps 62
- heat disorders 62
- heat exhaustion 62
- heat homeostasis 286
- heat stress conditions 302
- heat stress indices 292
 - effective temperature 292
 - rational/analytical 294
 - use and application 294–5, 295
 - wet bulb globe temperature 293
 - wet bulb temperatures 292
- heat transfer 287–9
 - conduction 287–8
 - convection 288–9
 - evaporation 289
 - radiation 289
- heated head anemometers 444
- heatstroke 63
- heavy metals, renal toxicity 57
- hemp dust, airways irritation 23
- Henry's law 65
- hepatitis A 352
- hepatitis B 353
- hepatitis C 353–4
- n*-hexane, neurotoxicity 55
- hierarchical task analysis 375–6, 376
- HIV 354
- home–work interface 365
- hoods 449
- hot conditions 301–3
- housekeeping 436–7
- human errors 413
- human immunodeficiency virus *see* HIV
- humidity 88–9, 299
- hydatid disease 348
- hydrogen sulphide narcosis 55
- hypersensitivity reactions 32–3
- hypothermia 64
- ICNIRP guidelines 317, 322
- ideal gas laws 89
- illuminance 269
- illuminance levels 284
- illuminance meters 282
- immersion foot 63
- immune system disturbance 367
- impaction 48, 99
- impairment 37
- impedance of free space 308

- importance principle 386
in vitro systems 71
incidence 173, 174
individual risk 109
industrial chemicals, regulatory framework 79–80, 80
infrared radiation 269
inhalable fraction 186
inhaled materials, effects of 47–58
 cardiovascular system 58
 kidney 57–8
 liver 57
 lung
 airways disease 50–3
 airways irritation 49–50
 extrinsic allergic alveolitis 53
 malignant disease 53–4
 particle deposition 47–9, 48, 49
 nervous system 55–7
insects, airways irritation 23
inspection lighting 280
integrated occupational health 8
integrating meters 301
interception 99, 100
International Agency for Research on Cancer 8
International Commission on Radiological Protection 338
International Standards *see* ISO
interpersonal relationships 364
interstitial fibrosis 23
Ionising Radiation Regulations (1985) 106, 338
ionization chamber 336
ionizing radiation 328–43
 alpha radiation 331, 333
 atomic structure 329–30
 background sources of radiation 342
 beta radiation 331–2, 333
 biological effects 337–8
 bremsstrahlung radiation 332
 control
 administrative measures 339
 distance 340
 dose limits 338, 339
 of external exposure 340
 internal radiation 341–2
 legislative 338–9
 practical measures 339–340
 radiation monitoring 341
 shielding 340
 time 340
 transport of radioactive material 342
 external radiation 334
 gamma radiation 332, 333
 instrumentation 336–7
 measurement 334–6
 units of radiation dose 335
 units of radiation dose rate 335–6
 units of radioactivity 335
 neutron radiation 332, 333–4, 333
 properties of 332–4
 radioactive contamination 334
 radioactivity 330–1
 sealed/closed sources 334
 unsealed/open sources 334
 X-rays 329, 332, 333
iron, airways irritation 23
irritant contact dermatitis 30–2, 31
 clinical course 31–2
irritation 69
ISO 2631 261, 263–4, 263
ISO 5349 258, 259–60, 261
ISO 7423 293
ISO 7730 291, 295
ISO 7933 291, 294
ISO 8662 258
isocyanates
 airways irritation 23
 occupational asthma 51
job title 146
katathermometer 299
key performance indicators 407
kidney, toxicity 57–8
labelling 113
lagging performance indicators 407
Lambert–Beer law 103
lasers 323–6, 323
 classification 324–5, 325
 hazards 324
 measurements 326
 standards 325–6
 uses 324
latent heat of vaporization 288
latent variables 140
latex
 allergy 33
 occupational asthma 51
LD₅₀ 74
lead
 blood reference value 164
 cardiovascular toxicity 58
 neurotoxicity 55

- leading performance indicators 409
- legal matters
- occupational health management 475–6
 - psychological issues 370
 - work-related stress 426
- leptospirosis 350
- light and lighting 268–85
- daylight 277–8
 - discharge lamps 276–7, 277
 - electromagnetic spectrum 268–9, 269, 308
 - emergency lighting 280
 - exterior lighting 280–1
 - filament lamps 276
 - illuminance levels 284
 - infrared radiation 269
 - inspection lighting 280
 - lighting systems 278
 - lumen method of lighting design 279–80
 - luminaires 277
 - surveys and survey techniques 282–4
 - illuminance meters 282
 - interpretation of data 282, 284
 - luminance meters 282
 - minimum number of measuring points 282
 - preliminary report sheet 282, 283
 - terms and definitions 269–70
 - ultraviolet radiation 269
 - visible radiation 268–9
 - visual display units 281–2
 - general lighting requirements 281
 - luminaires 281
 - reflections and glare 281
 - uplighters 281–2
 - visual task lighting 278–9, 279
 - see also* eye and vision
- lighting systems 278
- lindane, contact dermatitis 33
- liver, toxicity 57
- local exhaust ventilation 115, 148
- local records 172
- locus of control 362
- longitudinal studies 175
- loudness 227
- low-frequency radiation 309–15
- control 313
 - low-level exposures 312–13
 - measurements 313–14
 - ranges and uses 310
 - static electric fields 311
 - static magnetic fields 311–12
- low-frequency electric fields 311
- low-frequency magnetic fields 312
- lowest observed adverse effect level (LOAEL) 74, 108
- lumen 269
- lumen method of lighting design 279–80
- luminaires 269, 281
- luminance 269
- luminance contrast 270
- luminance meters 282
- luminous efficacy 270
- luminous flux 270
- luminous intensity 270
- lungs 13–24
- acinus 14
 - airway lining 16–17, 16
 - airways 13–14, 14, 15
 - alveolus 15–16, 15
 - blood supply 17
 - blood-gas barrier 16
 - blood-gas transport 22
 - compliance 21
 - consequences of branching 14–15
 - function 18–19, 18
 - and age 22
 - smoking 22–3
 - gas transfer 21–2
 - inhaled materials
 - airways disease 50–3
 - airways irritation 49–50
 - extrinsic allergic alveolitis 53
 - malignant disease 53–4
 - particle deposition 47–9, 48, 49
 - lymphatics 17
 - occupational lung disease 23
 - pleura 17–18
 - structure 13
 - ventilation 19–21
- Lyme disease 347–8
- lymphatics
- lungs 17
 - skin 28
- magnetophosphenes 312
- maintenance 436–7, 480–1
- personal protective equipment 465
- maintenance factor 270
- MAK values 166
- malaria 351
- malt worker's lung 54
- man-made mineral fibres, respirable 188
- Management of Health at Work Regulations (1992) 106
- mandelic acid, urinary reference value 16

- manganese
 - mental changes 55
 - parkinsonism 55
- manifest variables 140
- Manual Handling Operations Regulations (2004) 377
- Manual Handling Regulations (1992) 39
- maple bark stripper's lung 54
- mass flow rate 442
- materials
 - absorption coefficient 234–5
 - composite partitions 237
 - sound insulation 236–7
 - transmission coefficient 236
- maximum exposure limit 110, 462
- melanocytes 28–9
- mental changes 55, 56
- mental health 366
 - see also psychological issues
- mercury
 - airways irritation 50
 - mental changes 55
 - urine reference value 164
- mercury compounds, neurotoxicity 55
- metabolic rate 288
- method limit of detection (MLOD) 129
- methyl mercury, balance/vision defects 55
- methylbutylketone, neurotoxicity 55
- 4,4'-methylene dianiline, urine reference value 164
- microbiology laboratory workers, biological hazards 355–7
- Microsporum canis* 347
- mists 93
- mixtures 139–40
- monitoring 480–1
- Monte Carlo simulation 119
- morbidity 172
- mucociliary escalator 16
- multifraction samplers 195
- multiple linear regression analysis 136–8
- musculoskeletal disorders 36–46
 - back pain 38–45
 - hand–arm vibration syndrome 45
 - stress-induced 367
 - terminology 36–7
 - upper limb pain 37–8
- mushroom worker's lung 54
- nail 27
 - samples 162, 163
- narcosis 55, 56
- national records
 - ad hoc records 172–3
 - birth certificates 172
 - death certificates 171–2
 - local records 172
 - morbidity 172
- near-field conditions 309
- negative affectivity 363
- neurological disorders, vibration-induced 255–6
- neurotoxins 55
- neutron radiation 332
 - properties 333–4, 333
- new substances 113
- nitrogen dioxide, airways irritation 23
- nitrogen oxides, airways irritation 50
- no observed adverse effect level (NOAEL) 74, 107, 116
- nociception 37
- nodular fibrosis 23
- noise 59–61, 222–49
 - aural comfort 234
 - background 226–7
 - basic acoustics 222–4
 - bel scales 225–6
 - composite partitions 237
 - control of exposure levels 245–6, 245
 - decibel scale 226
 - dosemeter (dosimeter) 233–4, 233
 - ear protection 246–9
 - ear defenders 466–7
 - earmuffs 246, 247
 - earplugs 246
 - received noise level 249
 - estimation of dB(A) level 229, 230
 - excessive
 - auditory effects 60–1
 - non-auditory effects 61
 - frequency analysis 228
 - hearing assessment 240–3
 - pure tone audiometer 240–3, 241, 242
 - standard hearing 240
 - test conditions 241
 - hearing conservation 244–5
 - immission level 243–4, 244
 - instrumentation 228
 - loudness 227
 - materials affecting 234–6, 235, 236
 - absorption coefficient 235–6
 - transmission coefficient 236
 - measurement of fluctuating levels 229–33
 - equivalent continuous sound level 230, 231
 - single event noise exposure level 232
 - statistical levels 232–3, 232
 - weekly average noise exposure 230–1

- mechanism of hearing 59–60
- reverberation time 239
- sound in enclosed spaces 237–9, 238, 239
- sound insulation 236–7, 237
- sound quantities 224–5, 224, 225
- survey report 248, 249
- Noise at Work Regulations (1989) 106
- noise criteria curves 234, 235
- noise dosimeter (dosimeter) 233–4, 233
- noise immission level 243–4
- noise rating curves 234
- noise reduction
 - at source 245
 - control of transmission path 245–6
 - ear protection 246–9
- non-detectable exposures 129
- non-ionizing radiation *see* electromagnetic radiation
- Nordic Musculoskeletal Questionnaire 379
- Notification of New Substances (NONS) Regulations (1993) 113
- Nottingham model 427

- occupational exposure level 110, 462
- occupational exposure standards 109–10
- occupational health management 473–91
 - behaviour 480
 - business risk 474–5, 488
 - control programmes 479–80
 - corporate social responsibility 475, 488–9
 - formal management systems 484–7, 486
 - auditing/verification 487
 - continual improvement 486–7
 - health hazards 476–7
 - health and hygiene team 483
 - international trends 475
 - leadership 489–90, 489
 - legal compliance 475–6
 - links to safety and environment 487–8
 - monitoring and maintenance 480–1
 - occupational health support 481–3
 - case management 482
 - health surveillance 482
 - measurement of system performance 482–3
 - pre-employment health status checks 481–2
 - rehabilitation 482
 - well-being 482
 - organizational objectives and targets 489
 - record-keeping 481
 - risk 476
 - categorization of 478
 - qualitative assessment 477–8
 - quantitative assessment 478–9
 - staff training/consultation 480
 - stakeholder involvement
 - community 484
 - directors/trustees 483
 - health-care system 484
 - regulators 484
 - workforce 483–4
- occupational health nurse 483
- occupational health physician 483
- Occupational Health and Safety Administration (OHSA) Act 106–7
- occupational health and safety advisor 483
- occupational hygienist 483
- occupational neurosis 37
- Occupational Stress Checklist 369
- Occupational Stress Measure 369
- optical particle samplers 198
- optical radiation 318–23, 318
 - biological effects 320–1, 320
 - eye 320–1
 - skin 321
 - control 322
 - measurements 322–3
 - physical interactions 319–20
 - sources of 319
 - standards 322
- orf 346
- organ of Corti 60
- organic dusts, airways irritation 23
- organic solvents, narcosis 55
- Organization for Economic Co-operation and Development 113
- organizational effects of stress 367–8
- organizational ergonomics 374
- organizational hazards 423–4, 425
- organophosphate pesticides, neurotoxicity 55
- osmium tetroxide, airways irritation 50
- Ovaka Working Analysis System 380, 382
- ozone, airways irritation 50

- pain 37
- parkinsonism 55
- partial pressure 89–90
- particle motion 96–102
 - aspiration 99–101, 100
 - diffusion 101–2, 101
 - drag force 96–7, 97, 98
 - elutriation 99
 - gravitational force 97–8
 - impaction and interception 99, 100
- particle shape 93–4, 94

- particle size 94, 95, 96
peripheral neuropathy 55–7
permanent threshold shift 60
permissible exposure limits (PELs) 107
personal monitors, thermal environment 298
personal protective equipment 374, 437–8, 460–71
 awareness of consequences of exposure 462
 compatibility of 463–4
 ear protection 246–9
 ear defenders 468–9
 earmuffs 246, 247
 earplugs 246
 received noise level 249
 free provision of 465–6
 inspection of 465
 maintenance of 465
 matching to wearers 462
 minimization of wear period 464
 monitoring of usage 465
 objective fit tests 463
 personnel at risk 461
 programme set-up 461–6
 protective clothing 304–5, 469–40
 record keeping 465
 respirators 118, 467–8
 risk assessment 461
 risk creation by 463
 risk reduction 461
 selection of 462
 standards 466–7
 storage facilities 465
 supervision of wearers 465
 training in use of 464
personal samplers 192
 respirable aerosols 193–4
 thoracic aerosols 195
personality 362
phosgene, airways irritation 23, 50
physical environment 365–6
physical ergonomics 374
physicochemical properties 70–1
physiologically based pharmacokinetic modelling 71
pigeon droppings, airways irritation 23
pitot-static tube 443–5, 443
pituitary snuff-taker's lung 54
platinum salts
 airways irritation 23
 contact dermatitis 33
 occupational asthma 51
pleura 17–18
pneumoconiosis 23, 52
 benign 53
 coal-miners' 53
political context 5
pre-employment health status checks 481–2
predicted heat strain 294
presbycusis 61
prescribed diseases 59
pressure 64–6, 441–2
 barotrauma 64–5
 decompression illness 65
 gas toxicity 65
 high altitude
 flying at 66
 working at 65
 increased 65
 loss of 452–3
 measurement of 442–3
Pressure Management Indicator 369
prevalence 173, 174
principal component analysis 140
proactive monitoring 408–9
propeller fan 455
proportional counter 336
proportionate mortality ratio 175
protective clothing 469–70
 cold environments 304–5
 hot environments 304
 see also ear protection
psittacosis 349
psychological issues 360–72
 assessment 368
 attitudes 362, 363
 coping strategies 362–3
 evaluation 670
 hardiness 362–3
 home–work interface 365
 intervention and control 369–70
 legal considerations 370
 locus of control 362
 management 368
 negative affectivity 363
 personality 362
 physical environment 365–6
 psychosocial hazards 363–4
 recognition 368
 stress effects 361–2, 366–70
 mental health 366
 physical health 366–8
 type A behaviour pattern 363
 work organization 364–5

- bullying and harassment 365
- career development 364
- change 364–5
- interpersonal relationships 364
- role 364
- violence and trauma 365
- working hours 364
- psychosocial hazards 363–4, 423–4, 425
- pulmonary oedema 23
- pure tone audiometer 240–3, 241, 242

- Q fever 345–6
- qualitative risk 108
- quality control in gas and vapour sampling 221
- quantitative risk 108
- quantitative SAR 71
- quick exposure checklist 380

- radiant temperature 298
- radiation 23, 289
- radiation dose rate 335–6
- radiation dose units 335
- radiation monitoring 341
- radiation shields 303–4
- radioactive contamination 334
- Radioactive Substances Act (1993) 338
- radioactivity 330
 - alpha radiation 331, 333
 - beta radiation 331–2, 333
 - bremsstrahlung radiation 332
 - gamma radiation 332, 333
 - neutron radiation 332, 333–4
 - nuclear particles 330
 - units of 335
 - X-rays 329, 332, 333
- radiofrequencies 314–17, 314
 - control 317
 - low-level exposures 315–17
 - physical interactions and physiological effects 315, 316
 - uses of 310
- rads 335
- Rapid Upper Limb Assessment (RULA) 382–4, 383, 384
- reactive monitoring 408–9
- reactive safety management 405–6
- received noise level 249
- record-keeping 481
 - personal protective equipment 465
- recruitment 60
- reference values 164, 165–6
- rehabilitation 482

- relative risk 176
- reliability 156–7
- rems 335
- repeated-dose toxicity 69
- repetitive strain injury 37, 38
- reproductive toxicity 70
- required sweat rate 294
- RESPICON sampler 195, 196
- respirable fraction 186, 187
- respirators 118, 467–8
- respiratory tract 19
 - deposition/uptake of gases 49
 - particle deposition 47–9, 48, 49
 - airflow 48
 - alveolar clearance 49
 - diffusion 49
 - impaction 48
 - sedimentation 48–9
 - see also* lungs
- retinal fatigue 271
- retrospective exposure assessment 145–59
 - case–control studies 156
 - dose and exposure indices 146
 - extrapolation of past occupational exposures 147–9
 - source–receptor model 147–8
 - task-specific TWA model 149
 - interindividual exposure variability 150–1
 - reliability and validity 156–7
 - traditional methods
 - calibrated expert judgement 153
 - cohort studies 153
 - deterministic physical exposure models 155
 - expert judgement 152
 - extrapolation of exposure over time 155
 - grouping of similar jobs 153–4
 - mean exposure estimation 15
 - semiquantitative expert estimates 152–3
 - statistical extrapolation 155
 - work history 146–7, 147, 149–50
- reverberation time 239
- Reynolds number 441
- ringworm 347
- risk 106, 476
 - attributable 176
 - categorization of 478
 - extent of 106
 - relative 176
- risk assessment 77–8, 105–23, 377–8
 - accident prevention 409–812
 - consequences 410–11
 - continuing hazards 410

- risk assessment (*cont'd*)
 establishing objectives 410
 failure hazards 410
 failure modes and effects analysis 412
 HAZOPS studies 411–2
 system boundaries 410
 techniques 410
 tolerability/acceptability of risk 121, 411
dose–response 115–16
exposure assessment 116–19
hazard identification 112–15
hazards and risks 106
and legislation 106–11
 communicating risks 111
 cost–benefit analysis 110
 health effects 107
 human populations 107–8
 individual and societal risk 109
 occupational exposure standards 109–10
 risk rating 108–9, 108
models 111–12, 111, 140–2
personal protective equipment 461
qualitative 477–8
quantitative 478–9
recording 122
work-related stress 427–30, 428, 429
risk characterization 119–22
 acceptability and tolerability 120–1, 121
 exposure data versus exposure limits 121–2
risk phrases 113
risk rating 108–9, 108
röntgens 335
rotating vane anemometers 444
safety
 legislative control 404
 management systems 406–8
 performance management 408–9
 key performance indicators 409
 leading and lagging performance indicators 409
 proactive and reactive monitoring 408–9
 reactive safety management 405–6
 workplace safety management 404
 see also accident prevention
safety culture 415–16
sampling bags 211
scintillation detector 336–7
sebaceous glands 27–8
sedimentation 48
semiquantitative risk 108
sensitization 69
sequence-of-use principle 386
short-term exposure limit (STEL) 121
sick building syndrome 363
sieverts 335
silica, airways irritation 23
silica gel 213, 218–19, 219
silicosis 52
single event noise exposure 232
sisal dust, airways irritation 23
skin 25–35, 26
 after-work creams 35
 allergic contact dermatitis 33–4, 33, 34
 as barrier 30
 barrier creams 34–5
 blood flow 29
 blood and lymphatic vessels 28
 cleansing 35
 defence mechanisms 29–30
 dermis 27
 epidermis 25–7
 functions 26
 hair 27
 immunology 32–3
 irritant contact dermatitis 30–2, 31
 melanocytes 28–9
 nails 27
 nerve supply 28
 optical radiation effects 320–1
 sebaceous glands 27–8
 subcutaneous layer 27
 sweat glands 28
 sweat production 29
 thermoregulation 29
 see also dermal exposure assessment
slots 449
small airways disease 21
smoke 93
smoke tracers 443
smoke tubes 115
smoking, and lung function 22–3
Snellen chart 273
societal risk 109
solvents
 airways irritation 50
 balance/vision defects 55
 dermatitis 32
 renal toxicity 57–8
sorbents
 activated charcoal 91, 211–12
 coated 214, 220
 silica gel 213, 218–19, 219
sorber samplers 210–11

- sound insulation 236–7
 sound intensity level scale 225
 sound power 224
 sound pressure level scale 225
 source–receptor exposure model 147–8, 148
 spacing-to-height ratio 270
 specific absorption rate of energy 315
 SPEED program 131, 132, 133
 sprays 92
 standard-setting 78
 standardized mortality rate 175
 standards
 British Standard *see* BS
 electromagnetic radiation 317–18
 hand-transmitted vibration 258–60
 hearing 240
 international *see* ISO
 lasers 325–6, 323
 occupational exposure 109–10
 optical radiation 322
 personal protective equipment 466–7
 toxicology 78
 see also individual standards and directives
 static (area) samplers 192–3
 respirable aerosols 194–5
 thoracic aerosols 195
 static electric fields 311
 static magnetic fields 311–12
 still shade temperature 296
 Stokes' law 97
 straw, airways irritation 23
Streptococcus suis 348–9
 stress *see* work-related stress
 suberosis 54
 suffering 37
 sulphur dioxide
 airways irritation 23, 50
 contact dermatitis 33
 surface absorption range of energy 315
 sweat 29
 sweat glands 28

 talc, airways irritation 23
 task analysis 115, 375, 376, 386
 hierarchical 375–6
 task-specific time-weighted average (TWA) model
 147, 148–9
 temperature 61–4
 cold effects 63
 hypothermia 64
 cold injuries
 freezing 63–4
 non-freezing 63
 heat disorders 62
 heat effects 62
 cramps 62
 exhaustion 62
 heatstroke 63
 see also thermal environment
 temporary threshold shift 60
 tenosynovitis 38
 tetrachloroethylene, end-exhaled air, reference value
 164
 tetraethyl lead, mental changes 55
 thallium, neurotoxicity 55
 thermal balance 287–90
 thermal desorption 213–14, 219–20
 thermal environment 286–306
 cold stress indices
 required clothing insulation 297
 still shade temperature 296
 wind chill 296–7
 comfort indices
 analytical 295–6
 direct 296
 empirical 295
 control of
 cold conditions 304–5
 comfort 305
 hot conditions 301–3
 radiation shields and barriers 303–4
 risk assessment 301
 heat stress indices 292
 effective temperature 292
 rational/analytical 294
 use and application 294–5, 295
 wet bulb globe temperature 293
 wet bulb temperatures 292
 survey
 air temperature 298
 air velocity 299–301
 globe thermometer 298–9
 humidity 299
 integrating meters 301
 objective measurements and instrumentation
 297–8
 personal monitoring 298
 radiant temperature 298
 wet and dry bulb methods 299
 thermal balance 287–90
 heat transfer mechanisms 287–9, 288, 290
 overall balance 289–90
 thermal indices 290–1
see also temperature

- thermal indices 290–1
 direct 291
 empirical 291
 rational 291
 selection of 291–2
- thermoluminescent dosimeters 337
- thermoregulation 29
- thoracic fraction 187
- threshold limit values 166
- tin, airways irritation 23
- tinnitus 61
- tools
 design 386–7, 387
 vibration injuries 257, 258
- toxicity 67–8
 acute 68–9
 carcinogenicity 69
 genotoxicity 69
 irritation and sensitization 69
 repeated-dose 69
 reproductive toxicity 70
- toxicodynamics 67–8
- toxicological endpoints 68
- toxicology 67–82
 animal experiments 72–3
 hazard banding 79
 hazard identification 77
 human experience 73
 information obtained 74–6
 physicochemical properties 70–1
 regulatory framework for industrial chemicals 79–80
 risk assessment 77–8
 standard-setting 78
 in vitro systems 71–2
 see also toxicity
- training 480
 biological monitoring 167
 biological sample collection 167
 hazard control 437
 personal protective equipment use 464
- transport velocity 446, 448, 450–1
- trauma 365
- travel-related infections 350–2
 hepatitis A 352
 malaria 351
 travellers' diarrhoea 351–2
- travellers' diarrhoea 351–2
- trench foot 63
- trichloroethylene narcosis 55
- Trichophyton verrucosum* 347
- triethylene tetramine, occupational asthma 51
- triorthocresyl phosphate, neurotoxicity 55
- tuberculosis 354, 355
- type A behaviour 363
- ultraviolet radiation 269
- unemployment 6
- universal gas constant 89
- unsafe acts/conditions 406
- uplighters 281–2
- upper limb pain 37–8
 clinical types 44
- urine samples 161–2
- utilization factor 270
- validity 156–7, 177–8
- vanadium pentoxide, airways irritation 50
- vapour pressure 87–8
- vapour sampling see gas and vapour sampling
- vapour-hazard ratio 87–8
- vapours 23
- veiling reflections 279
- ventilation 303, 440–59
 air cleaning and discharge to atmosphere 458
 air density 440–1
 airflow measurement 442
 capture and transport velocities 450–1
 design features and volume flow rates
 enclosures 446, 448
 hoods 449
 slots 449
 dilution 451
 volume flow rate 451–2
 ducts and fittings 452
 duct sizing and loss calculation 453, 454
 extraction 445–6, 447
 fans 454–8
 axial flow 455–6, 455
 centrifugal 456–7, 456, 457
 characteristic curves 454–5
 matching to system 457–7
 power and efficiency 454
 propellor 455
 in hazard control 436
 instrumentation
 calibration 444
 heated head anemometers 444
 pitot-static tube 443–4, 443
 rotating vane anemometers 444
 techniques for using 444–5, 445, 446
 lungs 19–21, 20, 21
 make up air 458–9
 multibranching systems 453–4

- performance prediction
 - Fletcher method 449–50, 450
 - Garrison method 450
- pressure 441–2
- pressure losses 452–3
- pressure measurement 442–3
- Reynolds number 441
- smoke tracers 443
- units used 440, 441
- volume and mass flow rate 442
- zones of influence at terminals 442
- ventilation–perfusion 22
- ventilatory capacity 20
- vibration 250–67
 - assessment of 262–4
 - in buildings 262
 - direction 251–2
 - duration 252, 262
 - evaluation of 260, 262
 - frequency 251, 262
 - hand-transmitted 252–61, 252
 - causes of 257, 258
 - EU Machinery Safety Directive 260–1
 - EU Physical Agents Directive 261
 - musculoskeletal effects 255–6, 257
 - national and international standards 258–60
 - neurological effects 255, 256
 - prevention of 257–8, 259, 266
 - vibration-induced white finger 253–5, 253
 - magnitude 250–1, 261
 - whole-body 261–6
 - control of 265–6
 - discomfort caused by 261–2
 - health effects 262–5
 - interference with activities 265
- vibration dose value 262
- vibration syndrome 252–3
- vibration-induced white finger 253–5, 253
 - diagnosis 254–5, 255
 - signs and symptoms 253–4
 - Stockholm Workshop scale 254
- vibrational mass balances 198
- violence 365
- visible radiation 268–9
- visual acuity 272–4
- visual display units 281–2
 - general lighting requirements 281
 - luminaires 281
 - reflections and glare 281
 - uplighters 281–2
- visual fatigue 274
- visual perception 271–2
- visual performance 274, 275
- visual task lighting 278–9, 279
- vital capacity 21
- water, as skin irritant 31–2
- Weber–Fechner law 225
- weekly average noise exposure 230, 232
- wet bulb globe temperature 293
- wet bulb temperature 292, 299, 300
- wheat weevil disease 54
- whole-body vibration 261–6
 - control of 265–6
 - discomfort caused by 261–2
 - health effects 262–5
 - interference with activities 265
- Wien’s law 319
- wind chill 296–7
- women at work 7
- wood
 - contact dermatitis 33
 - occupational asthma 51
- wood pulp worker’s lung 54
- work design 387
- work evaluation 379–82
- work history 146–7, 149–50
- work life changes 6
- work organization 421–32
 - failures of 423–4
 - see also* work-related stress
- work pacing 387
- work role 364
- work-related stress 6, 421–32
 - coping strategies 362–4
 - definition of 361–2, 361
 - evidence for 422–3
 - health effects 366–70, 424–5
 - mental health 366
 - physical health 366–8
 - legal issues 426
 - Nottingham model 427
 - risk assessment 427–30
 - risk management paradigm 426–7
 - solutions for 425–6
 - translation and risk reduction 430
 - see also* psychological issues
- work-related upper limb disorders 376–7, 377
- working hours 364
- working plane 270
- workplace
 - changes in 387

design 382, 385
evaluation 379
hazard identification of 114
layout 385–6, 385
safety management 404

X-rays 329, 332
 properties 333

zinc chloride, airways irritation 50