THIRD EDITION

CRC Press
Taylor & Francis Group

PLANNING & DESIGN OF ENGINEERING SYSTEMS

GRAEME DANDY
TREVOR DANIELL
BERNADETTE FOLEY
ROBERT WARNER

PLANNING & DESIGN OF ENGINEERING SYSTEMS

THIRD EDITION

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

PLANNING & DESIGN OF ENGINEERING SYSTEMS

THIRD EDITION

GRAEME **DANDY**
TREVOR **DANIELL**
BERNADETTE **FOLEY**
ROBERT **WARNER**

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20170907

International Standard Book Number-13: 978-1-1380-3189-0 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

PREFACE TO THE THIRD EDITION

Planning and design are key activities which, together with management and construction or production, allow any engineering project to be taken from the initial concept stage through to successful implementation. Each engineering project, whether in the traditional or the newer developing fields of engineering, relies for its success on the application of the basic processes of planning, design and management. Our prime aim in this book is to show how the processes of planning and design are carried out.

However, the underlying purpose of the book is somewhat broader: to explain the nature of engineering and to describe the type of work engineers undertake. The book therefore deals with the problems engineers are called on to solve, and, most importantly, the common sense methodologies that are used to solve engineering problems. It also presents some quantitative tools that are useful in undertaking the work of engineering planning and design.

The book has been written for students who are commencing their studies of engineering, and for lecturers who are presenting classes to students in the early semesters of an engineering degree course. It is also intended for non-specialist readers who seek information on the nature of engineering work and how it is carried out. Some of the more advanced material in the later chapters is suitable for presentation in the later years of an engineering degree program.

The purpose of the book has not changed substantially from that of the first edition, which was published in 1989. That edition was co-authored by Graeme Dandy and Robert Warner. Since then, however, the need for introductory courses in undergraduate engineering programs has become more generally recognised by engineering academics. Courses that provide an introduction to engineering and to engineering planning and design are now an integral part of most engineering undergraduate programs.

In the second edition (published in 2008) David Walker and Trevor Daniell were added as co-authors and the contents were extended, revised and updated extensively. In particular, the chapters on creativity, problem solving, and on social and environmental aspects were extensively rewritten and expanded and new chapters on project scheduling, management, communications, law and ethics and risk and reliability were added. In our treatment of management we deal with team work, team building, time management and the management of people.

Following David Walker's retirement, he has been replaced as a co-author of this third edition by Bernadette Foley, who has been teaching engineering planning and design in our school. The other co-authors would like to acknowledge David's significant contribution to this book. He revised the chapters on engineering and society, creativity, economic evaluation and engineering decision-making in the second edition as well as adding valuable interest boxes throughout the book. David also took responsibility for the formatting of the entire second edition. He has been kind enough to allow us to use his second edition material in the third edition.

In the third edition, the material and references have been brought up-to-date. It also expands significantly on the previous section on economic evaluation and includes a new section on intractable problems and systems including a discussion of wicked problems, soft systems methodology and the agile approach to software development. Furthermore, additional problems and interest boxes have been added.

The text emphasises the fact that engineering is an essential part of the framework of our society, and is therefore closely linked with various social, political, legal and environmental questions and problems.

Much of the material in the first and second editions has been included in the first-year syllabus of various engineering degree programs at the University of Adelaide for many years, as well as at other institutions worldwide. Some of the more advanced topics are presented in subsequent years of our civil and environmental engineering degree programs. The extensive rewriting we have undertaken for the third edition reflects our ongoing experience in teaching in this field as well as significant changes that have occurred in engineering itself in recent years.

The majority of the examples presented in the book come from the field of civil and environmental engineering, which is the area of engineering expertise of the authors. Nevertheless, we believe that the ideas and content are generally applicable and relevant to all fields of engineering, and beyond.

The material contained in this new edition can be used in various ways. We have divided the major portion of the book into two parts: the first ten chapters set out a conceptual framework for undertaking engineering projects, while Chapters 11 to 13 provide a range of techniques and tools for solving the sorts of problems that commonly arise in engineering planning and design.

The conceptual framework provided by Chapters 1–10 is designed to introduce engineering as a profession, to set out what engineering work involves and how it is undertaken, and to describe an approach to engineering problem-solving that is both flexible and formal. We see all ten chapters as being important for any first year engineering course. At the University of Adelaide, material selected from these chapters is also applied in a first year design course for civil and environmental engineering students.

The techniques and tools presented in Chapters 11–13 are designed to allow quantitative assessment of design and planning problems. These chapters contain material that is beyond what would normally be covered in a first year introductory course and is ideal for higher-level courses in an engineering degree program.

Graeme Dandy,
Trevor Daniell,
Bernadette Foley
Robert Warner,

Adelaide,
2017

CHAPTER ONE

Engineering and Society

In this introductory chapter we discuss the nature, history, and scope of engineering work and the role of the engineer in society. In broad terms, engineers are responsible for designing, planning, constructing and maintaining the physical infrastructure that supports modern society and allows it to function effectively. Engineers make extensive use of scientific and mathematical knowledge, but their work is distinguished from the work of scientists and mathematicians by an emphasis on the use of all available relevant knowledge to solve real infrastructure problems as economically and efficiently as possible, in an environmentally and socially responsible way.

1.1 MODERN SOCIETY AND ITS ENGINEERED INFRASTRUCTURE

Modern society functions within the framework of a vast and complex engineered infrastructure that supports, and indeed makes possible, modern everyday life. For example, the large volumes of clean water that are used each day domestically, and by industry and agriculture, are supplied by a complex engineering system that is made up of remote rain catchment areas, reservoirs, pumping stations, pipelines, desalination plants, water treatment plants and local networks of reticulation pipes. The water supply system is one small part of the infrastructure. Another system generates and supplies the energy that is used to heat, cool and light the buildings that we live in, work in and relax in. Enormous amounts of additional energy are consumed by our factories and industries. Yet another part of the infrastructure enables communications over large distances and the sending and receiving of large quantities of information almost instantaneously. We use a transport system to move people and goods within and between our cities and urban regions and between countries. At home, at work and at leisure we have an array of low-cost manufactured goods available, including labour-saving devices, personal computers, smart phones and audio and TV devices that reduce our physical work load and entertain us. These are just a few examples of the component parts of the engineering infrastructure.

This *engineering infrastructure* is made up of large, separate, but interacting, parts, which we call "systems". Each system has been designed to satisfy a specific set of goals that are related to community needs, and consists of many interacting components, or sub-systems. The built environment is a part of the engineering infrastructure. It consists of the many buildings, large and small, that provide shelter for us and for our belongings, and safe venues for our day-to-day activities. These buildings dominate the visual landscape of our cities and urban regions. They include high-rise and low-rise apartment buildings, houses, office buildings, factories, hospitals, schools, sports stadia, entertainment complexes and shopping complexes. They have been created by engineers, working with architects and builders. Of course, not all the engineering infrastructure is highly visible like the built infrastructure; much of it is unseen.

Although the engineering infrastructure has been created to satisfy community and individual needs, its functioning is not well understood or appreciated by many of the people who use it. When we want clean, fresh water we simply turn on a tap. We do not think about the many components and processes that bring the water to us. Likewise, by pulling the plug in a basin or by pressing a button on a toilet cistern we remove unwanted waste water and effluent without thinking of the complex disposal system and its many components that have been carefully planned, designed and constructed. Other disposal systems remove, treat and recycle the rubbish, garbage and other solid wastes that we generate. The processes of removal of liquid and solid wastes, of their transport, treatment, recycling and disposal, are out of sight and out of mind for most people. Lack of concern with, or understanding of, such details is perhaps a positive sign, that our engineering infrastructure is operating successfully.

This infrastructure has been developed progressively over many generations to meet the ongoing and expanding needs and demands of society, but there is also a feedback effect at play. The characteristics of the existing infrastructure, and especially its strengths and weaknesses, affect the way society itself develops and changes over time. A simple example is the way urban expansion is affected by the existing transport system. The transport system in a city develops in response to the need of the community to move people and goods efficiently, but there is inevitably a significant time delay before needed improvements are made to the existing transport corridors, and before new corridors can be constructed. New urban development therefore tends to cluster around the most efficient and most capacious existing transport corridors. The cost and efficiency of vertical construction and transport (in the form of fast elevators), as compared with the cost of land and horizontal transport, also influence the degree to which high-rise development will occur as well as, or as an alternative to, the continued outwards urban development.

In the past, major new steps forward in science and technology have resulted in enhancements to the infrastructure of our society. Over the last fifty or so years, developments have been particularly rapid, due in large part to the development of computer and Internet technologies. This development will continue into the foreseeable future and it will be complemented by the replacement of older parts of the infrastructure by newer and more efficient ones, because of the development of new, disruptive technologies.

It is the role of engineers to plan, design and create the systems, and their components, that constitute our physical infrastructure. They also have the task of managing, maintaining and expanding this infrastructure in response to developing societal needs.

Of course, the engineering infrastructure does not always operate perfectly, but for most of the time it is effective, efficient and economical. It is only on rare occasions, when for example there is a temporary interruption to the supply of water or electricity, or when the Internet is "down", or when there are longer term water restrictions, that most people reflect on the engineering infrastructure that makes our modern life possible.

Unfortunately, there are many countries and regions in the world that are without adequate engineering infrastructure. This is a challenge to engineers, but it is an equal or even greater challenge to politicians and civil servants. There are always technical and organisational problems for engineers to solve, but poor infrastructure within a region or city or nation is usually indicative of broader issues and problems that do not have engineering solutions. It is too often a sign that fundamental social and political problems must be addressed by the people and their political representatives.

City growth and transport technology

In the past there has been a progressive increase in the size and population of cities around the world, which has resulted from the development of new transport technologies.

In ancient city states such as Athens, people relied on the muscle power of humans and animals for transport. Although heavy goods could be delivered using animal-drawn carriages, most citizens walked wherever they needed to go. The city size was thus limited by the ability of most people to walk to the various parts of their city, and home again, comfortably in a day. This limitation applied to ancient Rome, and to cities up to and beyond the middle ages.

It was not until the 17th and 18th Centuries that mass transport began, initially by means of large horse-drawn vehicles. Cities such as London and Paris then increased rapidly in size and population. Before this, the size and layout of a city were governed by the need for workers to be within walking distance of their places of work (Richards, 1969; White, 1978). With the invention of the steam engine and then the advent of city and suburban railways in the 19th Century, large cities with a commuting workforce became possible.

In the 20th Century, the internal combustion engine and the automobile, together with upgraded roads and freeways, allowed city conglomerates to develop with a very large population spread over an enormous urban area. The way of life in such cities is necessarily very different to that in a small traditional city, such as those still found in Europe.

In a discussion of the relative merits of compact cities, Gordon and Richardson (1997) state that in 1890 the effective radius of US cities was approximately 2 miles (3.2 km). By 1920 it had grown to 8 miles (12.9 km) following the development of public transport. By 1950 the figure had risen to 11 miles (17.7 km) with the widespread use of the automobile. By 1970, with the construction of urban freeways the average radius had increased to 20 to 24 miles (32.2 to 38.6 km).

1.2 HISTORICAL NOTE

In this short note, we can only provide a superficial view of engineering in history. However, society and its engineering infrastructure have developed together in step, and are closely interwoven. One practical lesson from history is that engineering work must be based on an understanding of society and how it functions, and on a careful and impartial evaluation of the sometimes-conflicting demands that come from individuals and from the community.

The first significant engineering works were probably undertaken around 10000 to 12000 years ago, when our human forebears started to live together in permanent settlements. This was a time of great change: humans were giving up the life of nomadic hunter-gatherers to adopt a more sedentary, agricultural-based way of life. The settlements needed basic infrastructure, such as permanent dwellings, all-weather pathways, and outer defensive walls to protect against wild animals and marauding humans. Archaeologists tell us that the agricultural revolution probably began in the river valleys of the Fertile Crescent in the Middle East, which includes parts of modern day Iraq, Turkey, Syria, Lebanon, Jordan, Israel, and Egypt.

This new communal way of life, based on agriculture and the domestication of animals, led to greater efficiency in the production of food, which in turn allowed the settlements to grow and to become more efficient. As the settlements grew, there was need for more complex infrastructure, for large, communal buildings and for permanent and reliable sources of water, both for domestic use and for agriculture. Archaeological evidence of early irrigation systems and large buildings abound in the Fertile Crescent. Ruins have been excavated in Turkey of monumental buildings that predate the Egyptian pyramids by millennia. Extensive engineering works were carried out by the Sumerians in present-day southern Iraq, in the period from the fifth to the second millennium BCE. According to Kramer (1963), the land they inhabited was “hot and dry, with soil that was arid and wind-swept and unproductive”. But by 3000 BCE the Sumerians had turned the land ... “into a veritable Garden of Eden and developed what was probably the first high civilization in the history of man". To transform their environment, they had to use both technological and organisational skills. As Kramer points out, “the construction of an intricate system of canals, dykes, weirs, and reservoirs demanded … engineering skill and knowledge. Surveys and plans had to be prepared which involved the use of levelling instruments, measuring rods, drawing, and mapping." The creation of such buildings and irrigation canals required both the organisational and technological skills that are basic to engineering work.

Continuing improvements in the agricultural economy allowed people in larger settlements to specialise in their work, and, with increased leisure time, to put more effort into handicrafts and art forms. However, these developments also led to an ever more complicated way of life, and to a stratification of society with inequalities that are still with us. With everyday life and work becoming more and more complicated, it was necessary to create and maintain permanent records, and so the first systems of writing were developed, using the mediums of stone, wax, papyrus and, somewhat later, paper.

There was an increasing demand for engineering works as permanent settlements expanded into cities, supported by surrounding agricultural regions, and then as the cities and regions came together to become countries. Of course, these developments were not restricted to the Middle East. By the beginning of the common era, a little over 2000 years ago, the results of the agricultural revolution were to be seen in many parts of Europe, Asia and the Middle East. Engineering works of great complexity and magnitude had already been undertaken in ancient Egypt, the Roman Empire, India and China. Comparable developments in the Americas came somewhat later, because of the later migration of peoples into that part of the world.

The ever-increasing demands for an improved and expanded infrastructure led to increased efficiency and increased capacity in the irrigation systems for agriculture, the water supplies for cities, the roads for transport and for communication, and the walls and fortifications to protect cities against attack. In undertaking large, monumental engineering works, organisational skills were needed, as well as technological skills. As Rivers (2005) has pointed out, the Great Wall of China, which was started around the 6th or 7th Century BCE, "could not have been constructed solely with the technology of cutting stone from quarries and then transporting those stones to the designated sites. Unless an elaborate organisation existed in Chinese society, the Great Wall would never have been built, then or now."

An ancient engineer's life and its rewards

Sprague de Camp (1963) gives an account of an engineer who made good in ancient Egypt, a land of rigid class lines not easily crossed. However, engineers succeeded in crossing them, since engineering ability was not a common gift. The builder-engineer-architect Nekhebu (from 24th Century BC) told on his tomb the story of his rise from humble beginnings:
"His Majesty found me a common builder; and His Majesty conferred upon me the offices of Inspector of Builders, then Overseer of Builders, and Superintendent of a Guild. And His Majesty conferred upon me the offices of King's Architect and Builder, then Royal Architect and Builder under the King's Supervision. And His Majesty conferred upon me the offices of Sole Companion, King's Architect and Builder in the Two Houses."
Nekhebu also received other titles and was rewarded with gold, bread and beer.

Early engineering was usually either military or non-military in nature. The term "civilian" engineering was used to distinguish between the two fields, and is now referred to as civil engineering. Armed conflict makes up a large part of human history; it seems that it has always been with us. Skirmishes surely occurred between small competing groups of hunter-gatherers. This would have escalated to conflict between permanent settlements, and out-and-out warfare between city states, and then between countries and even civilisations.

Two important aspects of military engineering are defence and attack. The engineering works used in the defence of ancient cities included high outer walls and fortifications, and large underground cisterns within the city to store water. To attack and besiege a city, military engineers could attempt to undermine the outer walls and fortifications, and use mechanical catapults to bombard the walls and the interior of the city with large rocks. Movable towered structures with protective armour assisted in the breaching of walls. Catapults and cranes were used by the defenders to bombard the attacking army and to drop rocks and burning oil on those attempting to scale the walls. There were also non-engineering options for the besieging army: to attempt to poison the water supply and to cut off the supply lines to the city that brought food and other essential goods.

Early American civilisations and their engineering infrastructure

When Juan de Grijalva set sail from Havana on April 8, 1518, he was in search of an ancient civilisation, one that Spanish sailors had discovered quite by accident the previous year. Three sailing ships, blown off course by a severe storm, had happened upon ancient ruins of a splendid city. There, according to Gallenkamp (1960) "an astonishing sight was visible in the distance: rising up as though an outgrowth of the native limestone were a high wall enclosing a series of terraced pyramids and palace-like buildings constructed of carefully fitted stones".

The constructions were more than just the work of skilled stonemasons: they showed a high level of planning, design and organization, and the Spaniards were convinced they had found a lost civilisation. This assessment came well before they saw any of its population, before they had studied its language, its art, and its legal and political systems. The evidence of technological expertise, engineering skill and organisation on a significant scale was all that was required.

The Maya civilisation was not unique. There were also the Incas in South America and the Aztecs in central and southern Mexico. In each case an advanced civilisation could be recognised by the architecture and engineering infrastructure. On further study, the characteristics of each civilisation were uncovered. Sometimes there were surprises: the Incas had no written language, nor had they developed the wheel, yet the fact that they had achieved a state of civilisation was beyond question on the basis of their engineering prowess alone.

Figure 1.1 Ruins of the Inca Settlement at Machu Picchu, constructed in the mid-15th Century CE (© G C Dandy).

Archaeological evidence indicates that mechanical, mining and chemical engineering works were also undertaken in the distant past, when humans started to live together in settlements. Around 9800 BCE, gypsum plaster was produced in

Mesopotamia by heating gypsum in wood-fired kilns. This material is also known as "plaster of Paris" because there was a large deposit of gypsum in Montmartre in Paris, France. In the time between 8000 and 7000 BCE, people inhabiting what is now modern day Turkey mined copper ore. They beat it into beads, hooks and metal sheets. Somewhat earlier, the people of Mesopotamia had formed copper plates into small tubes as a form of jewellery (Mithen, 2003).

Later in ancient Rome, multi-storey apartment buildings up to six stories high were constructed in the cheap parts of town to house the ordinary people, the plebeians. As the building work was often shoddy, special building ordinances were created to try to improve the quality of construction.

Water was supplied to the city of Rome (though not to individual apartment buildings) by aqueducts. These traversed the plains, and, where necessary, linked up with tunnels that were dug through the hills. Of the 11 aqueducts for the city of Rome, one was over 100 km in length. In the city, water was reticulated by means of pipes made of clay and lead. Although engineering has been attributed with improving the health of populations by the provision of clean potable (drinkable) water, it has been speculated that the use of lead water pipes may have been responsible for lead poisoning among the population of Rome before its fall in the 5th Century CE.

Spectacular multi-level stone Roman aqueducts still survive in various places in Europe, such as Segovia in Spain, and Nîmes in France (Figure 1.2).

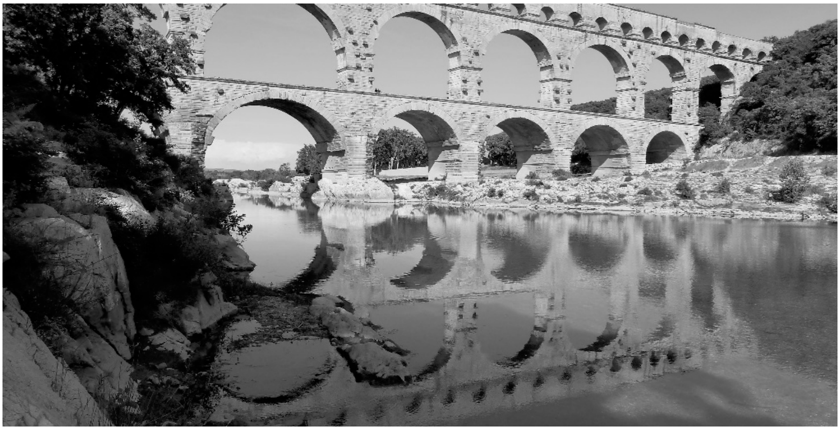


Figure 1.2 The Roman Pont du Gard, part of the Nîmes aqueduct crossing the Gardon river in southern France (© R F Warner).

Engineering in the ancient world was based on experience and empirical rules. According to Straub (1952) "Despite the remarkable scientific standard, especially of the Greeks in the spheres of mechanics and statics, there was hardly any connection between theory and practice, and hardly any attempt to apply the scientific knowledge to practical purposes, in the sense of modern engineering." Engineering was practised as a craft, with reliance on common sense and time-honoured experience rather than on the application of scientific principles.

Nevertheless, the Romans used arches and domes to construct beautiful and almost ever-lasting buildings and bridges.

With the development of scientific knowledge during the Renaissance, the usefulness of scientific principles was increasingly recognised, and science and mathematics became important in the education of engineers. In the last decade of the 16^{th} Century, Galileo Galilei began teaching mechanics, geometry and astronomy at the University of Padua. In his 18 years there, he published important works on the motion of falling objects and on hydrostatics. In his book "Dialogues concerning two new sciences", he investigated the foundations of statics and dynamics. This is regarded by some scholars as the birth of engineering science.

In France, a formal school for military engineers was established in 1675. Members of the Corps du Génie were instructed by military engineers of the calibre of Vauban, whose star-shaped designs for defensive fortifications were known and used throughout Europe. This school was so successful that a French civil corps was formed in 1716, the Corps des Ingénieurs des Ponts et Chaussées. The École des Ponts et Chaussées was created in 1747 to train civil engineers. This school was unique at the time because mathematics, geometry, mechanics (statics) and testing were considered to be the basis for engineering works of all kinds. According to Kirby et al. (1990), the rapid development of this style of formal training came about when it was found that designs based on scientific principles were more economical than those based purely on experience.

Initially, the French emphasis on using scientific knowledge and experimentation was not followed universally. In England in the 18^{th} century, engineering was still regarded as an art which was learnt by working as an apprentice with a master. Among the successful, and indeed excellent, British engineers there tended to be a mistrust of mathematics, and theory was thought to be of little value to the practitioner. Thomas Telford, one of the most innovative and successful engineers of this time, made distinguished contributions to the design of aqueducts, cast-iron bridges, harbours, canals and earthworks, using, in his own words, "no theory".

The first technical university was established in Paris in 1794, to deal with a dearth of engineers and administrators in France as a consequence of the revolution. It was originally called the "École Centrale des Travaux Publics", but in 1795 became the "École Polytechnique". This set a pattern in Europe, with the establishment of polytechnics and technical universities in Zürich, Delft, and then in many other cities. The Massachusetts Institute of Technology was the first technological institute to be set up in North America. It was founded in 1861. In Scotland, and then in England, engineering was introduced as a discipline in the traditional universities. At Cambridge University, for example, the present Engineering Department was established in 1875.

The English word *engineer*, together with its equivalent in other European languages, seems to date from the 17^{th} Century and to come from the much earlier Latin word *ingenium*, which was a type of Roman siege engine. In France and in Britain the term "civil engineering" was used in the 18^{th} Ccentury to refer to non-military, or civilian engineering. John Smeaton is thought to be the first person to refer to himself as a "civil engineer".

In the late 18^{th} Century, engineers began to establish professional societies. A group in England, including John Smeaton, formed a society of engineers in 1771.

In 1828 that society became the Institution of Civil Engineers, which functions to this day. Their definition of engineering at that time highlighted the skills used, and emphasised working for the benefit of humanity: Engineering is… "the art of directing the great forces of power in nature for the use and convenience of man, as the means of production and of traffic in states, both for external and internal trade as applied in the construction of roads, bridges, aqueducts, canals, river navigation, and docks for international intercourse and exchange, and in the construction of ports, harbours, moles, breakwaters, and light-houses, … navigation by artificial power … construction and application of machinery, and in the drainage of cities and towns".

The focus of this definition was on civil engineering, and the activities listed were accordingly restricted. Specialisation within the ranks of non-military engineers began in the early 19th Century, following the development of steam power for factories and locomotion. This led to the formation in various countries of separate societies for mechanical, mining, electrical and chemical engineers. Further separation from these traditional societies occurred in the 20th Century whenever specialisation had increased to a level where it was appropriate to differentiate among the members and their skills. Specialisation within engineering has continued and indeed accelerated into the 21st Century.

An updated and more general definition of the overall field of engineering was made in 1958 by the US Accreditation Board for Engineering and Technology: Engineering is ... "the profession in which a knowledge of the mathematical and natural sciences, gained by study, experience, and practice, is applied with judgment to develop ways to utilize, economically, the materials and forces of nature for the benefit of mankind".

In this statement (quoted in Voland, 2004) the list of specific design skills has been removed and the theme of working for the benefit of mankind has been introduced. The idea of working economically has also been included. In fact, concepts of cost control and economic operation began to be used overtly in engineering work in the late 19th Century. These are important considerations in all modern engineering work, although when new, innovative techniques are being developed and used for the first time, the main problem may be to get the project completed successfully. A distinguished engineer has ruefully suggested that it is often better, financially, to be the second or third person to use a new, innovative process.

In the middle of the 20th Century there were still five main engineering branches: civil, electrical, mechanical, chemical and mining engineering, but, following the Second World War, development was rapid in all fields of engineering and many new branches of specialisation appeared, and matured. Subdivisions of the traditional fields, and the appearance of completely new engineering disciplines, have continued to this day. The newer engineering branches flourishing towards the end of the 20th Century included aeronautical, aerospace, agricultural, automotive, biomedical, coastal, computer systems, electronic, environmental, mechatronic, medical, optical, rehabilitation, and transport engineering, to name but a few. The pace of engineering innovation and development has if anything accelerated into the 21st Century, especially with the impetus from electronics, computer technology and information technology. Walesh (2012) has pointed out that 27 different types of undergraduate engineering programs were recognised in 2011 for the purpose of course accreditation in the United States by the organisation ABET (2011). With the relatively recent introduction of the Internet and with new developments in artificial intelligence, we can expect to see a rash of new innovative engineering applications in many fields in the near future.

1.3 THE NATURE AND SCOPE OF ENGINEERING WORK

Although engineering work is becoming increasingly more diverse and more specialised, some characteristics are common, and this allows us to identify the basic nature of engineering work. A common characteristic is the need for engineering work to be undertaken in minimum time, and using limited resources such as people, energy, materials and money. For this reason, modern engineering work is undertaken as individual projects, with clearly specified goals. Another characteristic is the use of scientific knowledge, technology and mathematics. Activities that are common to the different fields and branches of engineering include planning and design, modelling and analysis, implementation, research and development, organizing and managing, and the sale and marketing of engineering products and services. These activities are an indication of the nature and scope of engineering work in the early 21st Century. By discussing them, we can highlight the common aspects of engineering work.

Engineering projects

As we have seen, engineers have the responsibility of developing, maintaining, and extending the physical infrastructure, and thereby improving the conditions of life in the community. However, with limited resources on the one hand, and unlimited needs and demands of the community on the other hand, efficiency is a prime requirement, so that *engineering projects* need to be carried out with the minimum use of scarce resources, in limited time, and with clearly stated goals and definite starting and finishing states.

The goal of an engineering project may be to create a physical object, that is to say, a piece of hardware, that will be a component in a physical system. For example, the object could be a new clutch system for a car, an entirely new electric car, or a network of reticulation pipes for water supply in a new city suburb.

However, the goal might not be to create a physical object at all, but instead to create a list of steps that will define a factory process. It might be to determine a set of instructions, or a list of actions, that, when undertaken, will achieve a goal. For example, a protocol for the release of water from a group of dams in times of very heavy rainfall might be created with the purpose of alleviating flooding in a river valley. The term *plan* is commonly used to refer to such protocols and lists, and we will use this term in this sense. A computer program might also be regarded as a plan. The word *artefact* has been used in engineering texts as a general term that includes both physical objects and plans.

Some examples of engineering projects are: the design and construction of an overpass bridge (a physical object); the design, construction and operation of a chemical plant to produce fertilizer (a physical object and a plan); the adaptation and programming of a microprocessor to operate a programmed washing machine (a plan); and the widening and strengthening of a bridge to allow increased traffic in the future (both a plan and an object).

A considerable part of modern engineering work is concerned with the maintenance and upkeep of existing facilities. Although para-professional staff may well undertake maintenance work on a routine basis, a substantial plan is usually needed for the maintenance and repair of the major items of infrastructure. The term *asset management* is applied to such activities, and it is an important engineering activity.

Planning and design

Planning and design are activities in which the details of an engineering system or a process are worked out to the extent necessary for implementation to be undertaken. For example, before work can start on the construction of a building, a structural design must be carried out, and then the building processes have to be carefully planned. The mechanical, electrical, communications and other services also have to be planned and designed. The planning and design phases of an engineering project are crucially important, and usually lengthy and time-consuming.

The terms planning and design are not mutually exclusive and there is no consistency in the technical usage of these words. It could well be argued that a design is a type of plan. Certainly in structural engineering, the results of a design are presented as a set of working drawings, or “plans”. Nevertheless, it is useful to distinguish between planning and design, and the matter of terminology will be taken up again in Chapter 3.

Planning and design form crucial parts of any engineering project, irrespective of the engineering fields which may be involved. The quality of the planning and design work will have a decisive effect on the success of the project. Planning and design are also complex activities, so complex that they can usually only be carried out in an iterative manner, with much trial and error. A large part of this book is devoted to explaining how the basic engineering activities of planning and design are undertaken.

Implementation

The *implementation* phase of an engineering project usually commences when the planning and design phases are coming to a close. Implementation will be the most costly phase of projects that are aimed at creating physical objects, such as a building, a chemical plant or an automobile, where extensive use must be made of both physical and human resources. There are various specialised branches of engineering that relate to the implementation phase, such as industrial engineering, manufacturing engineering, construction engineering and supply chain engineering. Much detailed engineering knowledge and many specialised techniques have accumulated, and are dealt with in undergraduate courses and relevant textbooks.

The situation can be different when the aim of the project is to produce a plan. The implementation phase might then become less costly. For example, implementation of a simple plan for the release of water from a dam at times of heavy rainfall, to prevent flooding but to maintain sufficient water for consumption, might require little more than the production of a manual and short instruction sessions for relevant technicians and operators. However, a water-release plan is not always simple. In complex situations continuing engineering input and judgement may be required, and the implementation phase could then become costly. The development of the plan, with systems to measure and control the quantities of released water and water levels, and evaluations to produce feedback information, could also be a major part of the cost.

Engineering management

The planning, design and implementation of a typical engineering project involves intense human effort and the use of scarce resources. Careful *management* of the entire project is needed. Management involves organising and co-ordinating the efforts of the people and the controlled use of physical processes in all phases of the project, so that a successful outcome is achieved within a limited budget. In this sense, management is a planning process.

Co-ordination of people is a vital part of engineering management. Many specialists with highly focussed skills can be involved and so there is always the need for teamwork. It is common nowadays for special project managers to be appointed for the overall management of engineering work. If one looks at the billboard outside a major construction site, the name of the project manager is listed prominently, followed by the names of specialists, including architects, structural engineers, constructors, mechanical services providers, electrical services providers and so on.

As the scale and complexity of a project increase, so too does the management effort that is required. A small project such as the design and installation of a roundabout in an existing streetscape might require a team of 4 or 5 professionals, in addition to the construction workers. The team would include a project supervisor, design engineer, a draftsperson and administrative support. By the time the project size is measured in multiple millions of dollars, the team is likely to be interdisciplinary, with 10 to 12 people, including additional engineers with specialised knowledge, and an increase in management and support staff. Engineering project management is an important field separate to, but closely linked with, planning and design.

Management of an engineering project is a broad and important field that we cannot attempt to deal with in this book. However, management of the design and planning phase is itself an important task and will be discussed further in Chapters 3 and 5, while various aspects of team work will be outlined in Chapter 6.

Engineering analysis, modelling and testing

In the design phase of an engineering project, it is necessary to predict how physical objects, not yet created, will behave under normal and extreme environmental and operational conditions. Analysis, modelling and physical testing are used for this purpose. In the modelling phase, the aim is to create an *analytic model*, which is a simplified conceptual representation of the object or process that takes account of the essentials and ignores the inessentials. The behaviour of the model can then be analysed theoretically to predict behaviour under a range of realistic conditions. Engineering analysis can involve a variety of disciplines, such as solid mechanics, fluid mechanics, soil mechanics, electronics, chemical processes, and informatics.

For example, in the design of an aircraft frame, including the wings and fuselage and tail-plane, computer-based theoretical analyses will be undertaken, using the principles of solid mechanics, to evaluate the stresses and strains that will occur in critical regions of the proposed structure at various stages of operation, such as take-off and landing, and under extreme turbulent conditions. The results of such analyses can then be used to judge the adequacy of competing alternative designs in regard to strength under extreme loads, deformations under normal working conditions, and other problems that can develop over time, such as fatigue.

Biomedical engineering and bionic ears (Cochlear implants)

Biomedical engineering has assisted many millions of people, either disabled through injury or with serious birth defects, to lead a normal life with the aid of special prostheses. Biomedical engineering began in the mid 20th Century as an interdisciplinary field in which engineers, medical scientists and medical practitioners worked co-operatively. There are many types of prostheses in use today, including computer-controlled artificial limbs, computer-generated speech devices, replacement joints for hips, knees and shoulders, and pacemakers that control heart functions, to name just a few. Biomedical engineers also create devices that are used in the diagnosis and monitoring of sick people, and in various forms of therapy.

A cochlear implant is a biomedical device that was designed to improve the hearing of people with damaged inner ears or who have been born profoundly deaf. While hearing aids make sounds louder, cochlear implants take over the functions of the inner ear (the cochlea) and enable sound signals to be sent to the brain. The device consists of an external microphone and electronic circuit, usually fixed behind the ear, and small electrodes that are implanted in the cochlea and wired to the external circuit. The electrodes accept impulses from the microphone-circuit and then stimulate the cochlear nerve. The invention of this device is attributed to David Clark in 1978, when he was working at the University of Melbourne.

In a recent article in the *Melbourne Age* (Saturday, 28th May, 2016) he explained how he had been inspired by a walk on the beach: “I saw a large seashell which had the same spiral shape as the cochlea, or inner ear. I had been trying to work out in the lab how wires might reach the regions of the cochlea to transmit the speech frequencies. That day I picked up the shell and threaded a long blade of grass around the shell, and then I realised I was onto something.”

David Clark was a medical graduate. His PhD led to the cochlear implant. He set up the Department of Otolaryngology at the University of Melbourne. He is currently an Honorary Professor of Electrical Engineering and Distinguished Researcher at NICTA, which is Australia’s Information and Communications Technology Research Centre in Excellence.

Although over a million people have used the cochlear implant, the device has stirred controversy. Objections to its use have come from some members of the deaf community. For pre-lingual deaf people, sign language is the primary means of communication, and some have considered the cochlear implant to be a threat to this culture.

While computer-assisted modelling and analysis can nowadays be applied successfully to quite complex physical objects, some are too complex to be modelled accurately using present day computing facilities. The testing of a *physical model* is an alternative approach. Wind-tunnel testing of model airframes remains an important step in the design of aircraft, owing to the complexities of representing and analysing turbulent flow. In coastal engineering, when the performance of a geometrically complex harbour with breakwaters and jetties is to be investigated under extreme weather conditions, the analyses may have to be supplemented, or possibly even replaced, by experimentation. If it is an existing system, it might be possible to undertake experiments and measurements under field conditions to supplement the theoretical analysis. In the design of very complex, yet-to-be-created objects, theoretical analysis may not be feasible. Physical laboratory models could then be used. In such cases, the model is itself a physical entity and it can be smaller or larger than the proposed prototype.

Research and development

Traditionally, basic scientific research is carried out to obtain new knowledge about the world we live in, for its own sake. Engineering research is more goal-directed, and is undertaken to provide new, useful information and knowledge in a specific field. This information may be needed for the development of new processes or for the design of new physical systems. The range of fields for engineering *research and development* is as wide as the fields of engineering. Engineering research may require the testing of new materials and components and the development of new techniques for the analysis and modelling of systems.

In large engineering projects, it is often necessary to undertake directed research, with clearly specified goals, to obtain or confirm important design information that is not otherwise available.

The value of, and need for, engineering research and development is well recognised, and considerable effort goes into this. In fact, the most rapidly developing technologies tend to be those where most money is being invested in research and development, although which is cause and which is effect is not always clear.

Sales and marketing

Planning and design, implementation, management and research and development are in the mainstream of engineering work, but there are many other tasks that must be carried out to ensure the success of an engineering project. Clearly, *sales and marketing* are vitally important in the introduction of new consumer goods, such as a new model of an automobile, a new piece of software, or even a new aircraft. Good engineering knowledge is highly desirable, and sometimes indispensable, for the successful sales and marketing of engineering projects, especially projects which aim to produce products that will be manufactured or created in large numbers.

1.4 DEALING WITH COMPLEXITY

One of the characteristics of our modern engineered infrastructure is its complexity. To appreciate this, it is only necessary to examine a typical piece of the physical infrastructure, say a road system that allows traffic to flow into, out of, and around a modern city. Such a road system includes hundreds of kilometres of multi-lane access highways and even more kilometres of smaller urban and suburban roads and streets. The access highways will include one or more circumferential ring roads that allow traffic to bypass the city, as well as the radial roads that bring the traffic into and out of the central grid of local roads and streets. Grade separation with cross-overs and fly-overs will be used where the main access roads intersect the ring roads, to allow the traffic to flow smoothly. Many bridges, embankments and retaining walls will be needed to provide the grade separations and to traverse geographic obstacles such as rivers. In the smaller streets and roads within the city, intersections will be at grade but with traffic lights or roundabouts to control and improve the flow. A vitally important component of the road system manages and controls traffic flow. This is in itself a complex system, made up of traffic sensors at key locations to measure traffic volumes, a sophisticated computerised plan with programs and computing facilities to optimise the traffic flows, and control devices, including lights and warning signs, that communicate instructions and information to drivers.

The complexity of such a system makes it physically impossible for any one individual to keep all the details in mind, let alone to follow the detailed functioning of each component from second to second in time. Furthermore, it would be physically impossible for a single engineer working alone to plan and design additions to such a system and then supervise their construction in a reasonable amount of time. Nevertheless, some pieces of engineering infrastructure must be created economically and in a short time span. When installed, they must operate safely, efficiently, and near faultlessly.

To deal with such complexity, engineers use a *systems approach*, which, in effect, is a process of analytic decomposition, whereby each complex entity is progressively broken down into simpler components and then sub-components until a stage is reached where it is possible to analyse how each individual part operates and how these simple parts interact with each other. In this way, it is possible to evaluate how the overall system will perform in its various modes of operation. This decomposition approach will be discussed in some detail in Chapter 2, where some basic systems concepts will be introduced to formalise the approach.

1.5 ENGINEERING PROBLEM SOLVING

Analytic problems and open-ended problems

In the course of an engineering project, various types of problems arise that have to be dealt with in different ways. In particular it is important to distinguish between *analytic problems* and *open-ended problems.*

As the name suggests, an analytic problem is solved by analysis and logical deduction. These problems occur typically in the detailed stages of design. In structural engineering, for example, in the detailed design of an office building, the deflections that occur in the reinforced-concrete slab floors under normal working conditions must be checked to ensure that they are not excessive. The details of the slabs will have been chosen in an earlier step in the design process. As in other analytic problems, the deflection that occurs is a single real value, and an estimate of it will be obtained by modelling, analysis, and detailed computation. Methods of modelling and analysis in the different engineering disciplines are dealt with in detail in undergraduate and graduate course programs in universities.

Open-ended problems do not have a unique solution and they do not lend themselves to the analytic approach. It is possible to find a range of alternative, acceptable solutions to an open-ended problem. To provide a simple example, we will continue the discussion of the design of a floor system for an office building. At an earlier stage in the design process, before analytic problems concerning deflections and strengths are studied, a choice must be made of the type of floor system to be used. The problem here is to choose, on the basis of cost and performance, from a wide range of possible floor systems, such as flat plates, flat slabs, ribbed slabs in reinforced concrete or pre-stressed concrete, or concrete slabs supported on a grid of steel beams, or indeed a completely new innovative flooring system that will better suit the specific building project. An evaluation and costing of the main alternatives will produce a reasonable and acceptable solution, if not the best one.

Another example of an open-ended problem arises from the traffic congestion that occurs at peak travel times in many cities and urban regions. Various solutions are possible, some being obvious, such as improving the public transport system or the road system, but some are less obvious, such as staggering office hours and extending opening times for shopping complexes. A mix of solutions might also be appropriate. The point here is that alternative solutions are not always obvious, nor always easy to evaluate.

It is not always possible or desirable to look back to previous work to discover the alternative approaches that can be taken to solving an open-ended problem. When a large, new engineering project is commenced, and roughly similar projects have been successfully completed in the past, it can even be counter-productive to rely on information from past practice. If there are unique features in the new work, past solutions might be unhelpful and even misleading. Over-reliance on past practice restricts the innovative thinking that is needed to discover new and innovative solutions. In Chapter 3 we will take up the topic of dealing with open-ended problems. Techniques are described there that are useful for solving open-ended problems.

Convergent and divergent thinking

Analytic problems and open-ended problems are quite different in nature and different modes of thinking are needed to deal with them. These are sometimes referred to as *convergent thinking* and *divergent thinking*. In engineering work, engineers have to be adept in applying the convergent, problem-narrowing approach when they undertake modelling and analysis to solve an analytic problem, and also the divergent, problem-widening approach when confronted with open-ended problems.

Most engineers and engineering students have been exposed (sometimes over-exposed) to convergent ways of thinking. In undergraduate engineering degree programs, emphasis is placed on science and mathematics, and on analysis and logical argument. Engineering students therefore tend to find analytic problems easier to deal with than open-ended problems. The decomposition approach that engineers use to deal with complex systems and problems is based on a convergent way of thinking, as discussed in Chapter 2 of this book. Techniques that enhance divergent thinking, and that can be applied to open-ended engineering problems are considered in some detail in Chapter 3.

Ill-defined problems

Many engineering problems, both analytic and open-ended, are *ill-defined* when they first arise. This can be because new and emerging community problems and needs are discussed in vague terms by individuals and groups in the community. Sometimes, however, engineering problems are stated in overly specific terms, so that a solution to the problem is stated unintentionally in the formulation. For example, a new bridge in a specific location may initially be seen by the community as an obvious need to alleviate traffic problems. This might be correct, but alternatives, such as a tunnel, or other locations for the bridge, will deserve consideration and may, upon investigation, prove to be a superior solution. An overly specific problem statement is in fact an ill-defined statement. One of the first steps to be taken in engineering work is to identify the real problem that is to be solved, as distinct from community perceptions and the views and statements of individuals. An early step in an engineering project is to identify the real needs of the community that are to be satisfied. In this regard, engineers cannot purely be problem-solvers; they must also act as problem-framers (Donnelly and Boyle, 2006). They must apply their skills and experience to recast the initial problem statements to ensure that sensible, if not optimal, outcomes are achieved. Problem definition is discussed in some detail in Chapter 3.

1.6 ENGINEERING, SCIENCE AND MATHEMATICS

Engineers are not scientists, nor are they mathematicians. There are important features that distinguish the work of engineers from that of scientists and mathematicians. The engineer has an active role to play in using resources to satisfy community needs. Scientists are more concerned with understanding and modelling the physical and biological world, and are primarily engaged in acquiring knowledge, rather than using it. The active work of the engineer is also distinct from that of the mathematician, who is concerned with logical structures which may or may not be observable in the physical world.

Engineers routinely make use of scientific principles, scientific knowledge and mathematics in their work, so that an engineering project can be carried out efficiently, and in limited time. But they make use of *all* available relevant knowledge, including that provided by science, mathematics and current technology, and also from fields such as economics and psychology. Nevertheless, the available knowledge is not always sufficient, and may then be supplemented by goal-directed engineering research and development work. This may require the development of new methods of analysis and modelling, laboratory and field experiments, or even a search through the literature to learn from past experience.

While the prime engineering concern is with achieving community goals, and not with the discovery of new scientific or mathematical knowledge as such, good engineering work can lead to new scientific, mathematical and engineering knowledge and new technologies. The boundaries between science, mathematics and engineering are not rigid. From time to time, engineering problems will arise that require input from scientists, mathematicians and other professionals. A common program of work will then result, with engineers working together with sother professionals.

Cutting-edge engineering work can lead to the development of new mathematical procedures and scientific knowledge. In the mid-20th Century, for example, approximate methods of structural analysis were developed by engineers. They became known as "relaxation methods", and allowed approximate, but accurate, numerical results to be obtained for the analysis of complex structural systems, including airframes. The methods became widely used in fields where closed-form mathematical analysis was not possible.

1.7 ENGINEERING IN THE 21ST CENTURY

The National Academy of Engineering of the United States (2006) published the following list of the 20 greatest engineering achievements of the 20th Century:

- Electrification
- Aeroplane
- Electronics
- Agricultural Mechanization
- Telephone
- Highways
- Internet
- Household Appliances
- Nuclear Technologies
- Petroleum and Petrochemical Technology
- Automobile
- Water Supply and Distribution
- Radio and TV
- Computers
- High performance materials
- Spacecraft
- Imaging
- Health Technology
- Laser and Fibre Optics
- Air Conditioning and Refrigeration

Such a choice is of course subjective, and there are many other competing engineering developments from the previous century that have greatly improved the quality of life that much of our planet enjoys. Such a list presents a challenge to current and future engineers: to match the contributions of our predecessors. They also invite the question: what will be the important engineering developments in the coming decades? An article in *Technology Review* (2005), designed to give readers a sample of things to come, listed the following 10 emerging technologies:

- Airborne networks
- Quantum wires
- Silicon photonics
- Metabolomics
- Magnetic-resonance-force microscopy
- Universal memory
- Bacterial factories
- Enviromatics
- Cell phone viruses
- Bio-mechatronics

Although some of these might not be successful, they do suggest developments we can expect to see in the near future. Expansion at the frontiers of engineering will surely be rapid, and in ways that were science fiction only a few decades ago. Some current trends and problems, and some likely future developments are now discussed.

Continuing problems of sustainability and climate change

The world population was already around 8 billion people in 2016 and has continued to increase. Humans are now so numerous that they affect the earth and its ecosystems in many ways. Local effects in regions of high density are obvious and have been measurable for many years.

When Svante Arrhenius first suggested in 1896 that carbon dioxide in the atmosphere might cause global warming, he was thinking of natural sources of the gas. Continuing climate change due to natural (non-man-made) effects has of course been going on ever since a crust formed on the earth, and Arrhenius did not believe that man-made releases could have any effect, given the very small amounts involved at the time (Bengtsson, 1994).

But now, more than a hundred years later, the situation is different. The consensus of the scientific community is that humans have the potential to make an impact on the climate of the whole earth and may have already started to do so. The phenomenon has been referred to as *climate change* or *global warming*. Some scientists consider that man-induced climate changes are occurring, with significantly higher, and possibly lower, temperatures in different parts of the earth.

There is ample evidence of change, irrespective of causes. In agricultural and other climate-dependesnt industries, it has become necessary to adapt to climate change. Grape-growers and wine-makers around the world, from the Barossa Valley in South Australia to the Rhine Valley in Germany, have had to change their harvest times, production procedures, and even grape varieties, in response to increasing temperatures (Daniell *et al*. 2017). In the United Kingdom, wine-making has become (again, after many hundreds of years) a viable industry. With changes in weather patterns and ocean currents, important problems of sustainability arise.

In the past, engineers have managed to satisfy most of society's needs, for increased energy, faster communications, safer and more rapid transport and more reliable water sources. In the future, engineers will have to be creative and responsive in facing the challenges and opportunities of climate change, and the difficult problems of sustainability. The sustainable use of water, energy and other scarce resources will be of particular importance. Engineering societies around the world now include in their code of conduct the requirement that engineers work in a sustainable manner: this will be critical over the next few decades. Engineering aspects of sustainability are discussed in Chapter 9 of this book.

The physical infrastructure is getting smarter

An emerging trend in engineering is the implantation of small sensors in engineering systems so that performance can be monitored. Sensors in buildings and structures monitor stress and vibrations, and can foreshadow trouble long before it manifests itself. Some sensors are sophisticated electronic devices, but recent developments have shown that much can be gained from quite simple arrangements. For example, carbon fibres embedded in concrete during construction as a strengthening measure have been found to show variable resistance based on whether the material is under compression or tension (Hansen, 2005). This means that performance can be monitored without the need to have special devices embedded beforehand. Research is proceeding with the development of devices that power themselves using vibrational energy from the structure, thus avoiding external power supplies.

Artificial intelligence has been described in terms of thinking machines and computers. We are only starting to see the applications of artificial intelligence, and we must expect an enormous expansion of its use in the near future, in fields as diverse as routine engineering design, financial planning, data retrieval, medical diagnosis and the automated guidance and operation of vehicles on land and sea, and in the air. When used in conjunction with the Internet of Things, the automated control, guidance, operation and maintenance of even the most complex devices and systems will become routine.

Engineering systems are becoming progressively more complex

The engineering systems that make up the physical infrastructure are becoming more sophisticated and more complex. This can lead to far-reaching and unexpected side effects, for example when one or more components fail or behave in an unexpected manner. The more complex a system becomes, the more likely it is that unexpected and unwanted positive feedback effects will occur.

Examples of such problems in complex systems have occurred recently in electricity power generation and distribution systems. These systems involve massive networks that connect many different energy sources to the distribution network. Because of this, there is the potential for significant problems due to unexpected side effects. On August 14th, 2003, a power failure affected a large proportion of North America and Canada, including the cities of New York, Detroit, Cleveland and Toronto, and left up to 60 million people without power. According to reports, the high speed at which the blackout spread was unexpected, as was its extent. In a matter of seconds, circuit breakers, installed to protect electrical installations from sudden and potentially damaging power surges, tripped, taking out entire regions. In total, more than 20 power plants were temporarily shut down, including nine nuclear reactors in four American states. A task force report has suggested that the interconnectivity of the nation's power grid, which is the reason for its success most of the time, leaves it vulnerable to this kind of massive failure.

In September 2016, the state of South Australia suffered a severe, crippling and prolonged blackout. Roads without traffic lights resulted in traffic disruption and chaos in cities and towns. With elevators unable to move, people were trapped in high-rise buildings, and of course in the elevators. Without refrigeration, supermarkets and restaurants were unable to operate, with spoilage of food. In hospitals, intensive-care patients had to be cared for manually when the automated support systems ceased to function. Although the causes are still being debated in 2017, inadequate overall planning of the energy system undoubtedly allowed this disruption to occur.

The potential for such failures means that engineers must become increasingly involved in modelling very complex systems as a step in their designs. It is likely that much can be learned from natural systems, where the web of life shows that resilience is possible even in very complex systems.

Engineering applications are increasingly used in medicine

Biomedical engineering (or bioengineering) has become increasingly important and we can expect this trend to continue. Three of the items on the above list of emerging technologies are in this area. Metabolomics concerns diagnostic testing for diseases by analysing sugar and fat molecules that are the products of metabolism. Bacterial factories can use engineered bacteria to produce drugs to combat diseases such as malaria. *Bio-mechatronics* is the area of engineering that links robotics with the human nervous system. This has led to new artificial limbs that behave like the real thing and that can be controlled directly by the person's brain.

According to Citron and Nerem (2004) bioengineering has been recognised as an established branch of engineering since the late 1970s. They identify two key applications for engineering input: diagnostic imaging (ultrasound, computer assisted tomography, CAT and functional magnetic resonance imaging, fMRI) and implanted therapeutic devices (e.g., pacemakers, cardiac rhythm stimulators, prosthetic heart valves, vascular stents, implanted drug pumps and neurological stimulators). Together, these have improved the health of large sections of the population, using ideas and methods that were unimaginable a few decades ago.

The application of nanotechnology in medicine is a related area where engineers have made significant progress. In 2004, Dowling (2004) reported on current and future developments, including better artificial implants, sensors to monitor human health, and improved artificial cochleae and retinas. She warned then that some of these might not be realised for at least ten years – which is not really a long lead time. Duncan (2005) gave further examples such as nanodevices that would diagnose problems and then deliver appropriate drugs to promote tissue regeneration and repair. She stated: "these ideas may seem like science fiction but to dismiss them would be foolish". We can expect such applications to continue and intensify in the future.

Engineering devices are getting smaller

Engineers in many fields are developing and using technology on a minute scale. Computer engineers are currently constructing and planning computer circuits with wire sizes measured in nanometres (billionths of a metre). Mentioned in the list of emerging technologies, quantum wires are extremely low-resistance wires that use carbon nanotubes, where the diameter of single-walled tubes is in the order of nanometres. To put that in perspective, the DNA double helix has a diameter of 2 nm, while an average human hair has a diameter around 80,000 nm. These are developments in the area called "nanotechnology", which had its first tentative steps in 1959 (Gribbin and Gribbin, 1997).

One of the early promoters of nanotechnology was physicist and Nobel prize winner Richard Feynman. In 1959, around sixty years ago, he instigated two prizes:

one for the first person to build an electric motor that would fit inside a 0.4 mm cube; and a second for anyone who could reduce printed text by a factor of 25,000, which could be read by an electron microscope. He was surprised a year later when someone arrived at his office carrying a large box. It contained a microscope (required to see the miniscule motor) and at that stage he knew he was going to have to pay up (which he did).

It took 26 years for the second prize to be claimed by a graduate student who had written the first page of *A Tale of Two Cities* by Charles Dickens at the required scale. As a matter of interest, this scale would allow all 24 volumes of the Encyclopaedia Britannica to be written onto the head of a pin (Feynman, 1999).

More recently, the size of motors has been further reduced. In 2003, details were given of a motor that consisted of a gold rotor on a nanotube shaft that measured 500 nanometres across (Sanders, 2003). One consequence of such miniaturised devices is the enormous reduction in power required to run them. Here also we can expect significant future developments.

Energy is being harvested as well as mined

The Industrial Revolution of the 18th Century was driven by the advent of machines that replaced human- and animal-powered devices and which significantly increased output by using concentrated energy sources. Wood was initially the energy source, but was quickly replaced by coal, which has in part been replaced by petroleum products. Oil has been partially replaced by fissionable (radioactive) material. At the beginning of the 21^{st} Century, oil was the dominant source of energy for much of the world's transport, heating and industry. The advantages of oil include its relative abundance, its concentration of energy and its low cost in relation to other energy sources.

As alternative energy sources are progressively introduced, such as solar, tidal, wind and wave, problems of energy density arise. The energy in the wind or in the waves is much less concentrated and requires significantly more physical infrastructure to harvest it. Further improvements in energy storage techniques (batteries) will assist in the use of sustainable energy sources. As the technologies improve, we can expect to see the use of more distributed power systems, and renewable energy systems. We might also see the development of efficient and clean methods for obtaining energy from petroleum, gas and coal.

Existing infrastructure needs will continue, but disruptive technologies will change some infrastructure

In the future, the basic needs of society are not going to disappear. They will continue and the basic stock of engineering infrastructure will be maintained. However, with the development of disruptive technologies we can expect to see the less efficient technologies replaced by new ones. This can be observed today with the relentless replacement of electronic software and printed circuits by new and improved generations.

1.8 SUMMARY

Engineers have the task of creating, maintaining and extending the physical infrastructure that allows modern society to function. This infrastructure has developed over many generations, in response to the needs and demands of society, and has become more and more complex over time. To deal with this complexity, engineering work is now quite specialised, and the trend to greater specialisation will continue into the future.

To make effective use of scarce resources, engineering work is undertaken in individual projects with clearly stated objectives, and with specified starting times and completion times. Engineers work together on projects in teams, which may include specialist engineers and specialists from other professions.

Planning and design are the key activities that are used, together with management, to bring engineering projects through to successful implementation.

The engineering profession is presently changing, and is being shaped by trends that have been visible for some time. Engineers are working with technologies that are on a scale previously unimagined: both the smallest and the largest. The ultra-small scale of nanotechnology is taking engineers into areas such as biomedical and bio-mechatronic engineering. Large-scale projects demonstrate the need to understand, analyse and model massive complex systems and to deal with unexpected side effects that result from positive feedback, which can be very difficult to recognise, let alone deal with effectively.

In the next one hundred years, engineers will be faced with the challenges of satisfying the needs of an increasing world population that has increasingly high expectations of maintaining and extending the physical infrastructure with systems that are economical and sustainable, and hence, of ensuring that our planet continues to be safe, hospitable and inhabitable.

PROBLEMS

1.1 Find the employment section of a wide-circulation newspaper and look for engineering positions. For job descriptions that fall outside the traditional forms of civil, electrical, mechanical, chemical and mining engineering, classify them as specialisations under the relevant traditional field.

1.2 Computers are listed as one of the 20th Century's most important achievements. The famous statement: "*I think there is a world market for about five computers*" has been attributed, rightly or wrongly, to Thomas J. Watson, who was Chairman of IBM in 1943 (Maney, 2003). Discuss what problems the Chairman might have anticipated computers could and could not solve.

1.3 What are some of the infrastructure problems that artificial intelligence may be able to assist in solving?

1.4 Which disruptive technologies do you think will become important in the renewal of the existing engineering infrastructure?

1.5 Consider the problem of establishing a small scientific station based in Antarctica, to be permanently occupied by 20 people. The base is to have shipping access, and will receive supplies once a year in

December. Make a list of the items of engineering infrastructure, and the main components, which will be needed to run the base successfully.

1.6 Develop a list of five engineering achievements that you think may be on the list for the greatest in the 21st Century. Which branches of engineering will be responsible for these? Do these branches exist yet?

1.7 Make a list of the tasks involved in setting up a temporary raceway for competition cars on public roads in or near a city centre. Consider the time that each task would involve and develop a time-line. The planning should have as a goal the minimisation of disruption to everyday city traffic.

REFERENCES

ABET, 2011, Criteria for accrediting engineering programs, www.abet.org/Linked%20DocumentsUPDATE/Criteria%20and%20PP//E001%2010-11%20EAC%20Criteria%201-27-10pdf.

Bengtsson, L., 1994, Climate Change, Climate of the 21st Century. *Agricultural and Forest Management*, 72, pp. 3–29.

Carcopino, J., 1941, *Daily Life in Ancient Rome*. (Peregrine), 365pp.

Citron, P. and Nerem, R.M., 2004, Bioengineering: 25 Years of Progress – But Still Only a Beginning. *Technology in Society*, 26, pp. 415–431.

Daniell, J. E., Daniell, T. M., Daniell, K. A., Friedemann, W., Schaefer, A. M., Bijan, K., Kunz, M. and Burford, R., 2017, Wine: the changing risk of a highly vulnerable industry globally to natural disasters, *European Geosciences Union, General Assembly* 2017, Vienna, Paper No EGU2017-2315.

Donnelly, R. and Boyle, C., 2006, The Catch-22 of Engineering Sustainable Development. Forum, in *Journal of Environmental Engineering*, ASCE, 132(2), pp. 149–155.

Dowling, A.P., 2004, Development of Nanotechnologies. *Nanotoday*, December issue, pp. 30–35.

Duncan, R., 2005, Nanomedicine Gets Clinical. *Nanotoday*, August issue, pp. 16–17.

Feynman, R., 1999, *The Pleasure of Finding Things Out*. (Penguin), 270pp.

Florman, S.C., 1976, *The Existential Pleasure of Engineering*. (St. Martin's Press), 160pp.

Gallenkamp, C., 1960, *Maya. The Riddle and Rediscovery of a Lost Civilization.* (Frederick Muller Limited), 219pp.

Gordon, P. and Richardson, H.W., 1997, Are Compact Cities a Desirable Planning Goal? *Journal of the American Planning Association*, 63(1).

Gribbin, J. and Gribbin, M., 1997, *Richard Feynman. A Life in Science.* (Viking), 301pp.

Hansen, B., 2005, New Concrete Can Monitor Itself. Civil Engineering, The Magazine of the American Society of Civil Engineers, 75(11), p.35.

Kirby, R.S.; Withington, S.; Darling, A.B. and Kilgour, F.G., 1990, *Engineering in History*. (Dover Publications), 530pp.

Kramer, S.N., 1963, *The Sumerians. Their History, Culture and Character*. The University of Chicago Press, 355pp.
Lehner, M., 1997, *The Complete Pyramids*. (Thames and Johnson), 256pp.
Lienhard, J., 2000, *The Engines of Our Ingenuity*. (Oxford University Press), 262pp.
Maney, K., 2003, The Maverick and His Machine: Thomas Watson, Sr. and the Making of IBM. (John Wiley & Sons). ISBN 0471414638.
Mithen, S., 2003, After the Ice. A Global Human History 20,000-5000 BC. (Phoenix), 622pp.
National Academy of Engineering, 2006, Greatest Engineering Achievements of the 20th Century. http://www.greatachievements.org/ (Downloaded 12/4/06)
Richards, B., 1969, Urban Transportation and City Form. *Futures*, pp. 239–251.
Rivers, T.J., 2005, An Introduction to the Metaphysics of Technology. *Technology in Society*, 27, pp. 551–574.
Sanders, R., 2003, Physicists Build World's Smallest Motor Using Nanotubes and Etched Silicon. UC Berkeley News.
http://www.berkeley.edu/news/media/releases/2003/07/23_motor.shtml. Downloaded 26/3/06.
Slezak, Michael, 2016, South Australian windfarms revise safety settings after statewide blackout, https://www.theguardian.com/environment/2016/oct/19/south-australian-windfarms-revise-safety-settings-after-statewide-blackout?CMP=share_btn_link
Sprague de Camp, L., 1963, *The Ancient Engineers*. (Rigby Ltd.), 408pp.
Straub, H., 1952, *A History of Civil Engineering*. (translated by E. Rockwell). (C. Nicholls & Co.), 258pp.
Technology Review, 2005, 10 Emerging Technologies. *Technology Review*, 108(5), pp. 44–52.
The Age Newspaper, Melbourne, 28th May, 2016, Inventor of bionic ear wins prestigious award and inspires new field of endeavour. http://www.theage.com.au/national/education/voice/inventor-of-bionic-ear-wins-prestigious-award-and-inspires-new-field-of-endeavour-20131008-2v5bd.html#ixzz49pwEf3Ke. See also: @theage on Twitter,| theageAustralia on Facebook
Vocand, G., 2004, *Engineering by Design,* (2nd). New Jersey: Pearson Prectice Hall.)
Walesh, S.G., 2012, *Engineering your Future,* (3rd Ed., John Wiley & Sons), 470 pp.
White, R., 1978, *Transport*. In: Environment and Man, Volume 8, The Built Environment. Eds. J. Lenihan and W.W. Fletcher, (Blackie), pp. 82–119.

CHAPTER TWO

Engineering Systems Concepts

One of the characteristics of the engineering infrastructure is its complexity. To deal with complexity, engineers use a process of decomposition whereby a complex object is conceptually broken down into successively smaller and simpler component parts until it is possible to study and analyse the individual components and the way they interact with each other. The systems approach provides an effective framework within which the work of planning, design and management can be conducted. In this chapter we introduce some basic systems concepts and explain how they are used in the modelling and analysis of engineering systems.

2.1 INTRODUCTION

In this chapter we introduce a methodology to deal with the problem of complexity. The idea is simple: a complex entity is progressively broken into smaller and smaller component parts until each part is so simple that it is possible to study its behaviour and the way it interacts with other simple parts. The planning and design work in a complex engineering project can also be broken down like this, into a number of inter-related tasks and sub-tasks that can be handled separately by individuals or small groups of engineers and other professionals. In this way, a large project which requires many thousands of hours of engineering work can be completed by a team of engineers in a short period of time; however, careful management of the team and the work is needed if it is to be successful.

This approach is based on a convergent, or deductive, way of thinking, which is also used in engineering modelling and analysis and in many other fields of study, ranging from biology to corporate planning and from computer science to business management. To formalise the approach, we will introduce a number of basic systems concepts with examples.

2.2 SYSTEMS AND PROCESSES

The word *system* is used frequently, if rather vaguely, in everyday speech in phrases such as water supply system, communications system, political system and biological system. The word has already been used informally in this way in Chapter 1, where reference was made to water supply systems, communications systems, transport systems and city systems. In everyday usage, the word system means some complex “thing” that is organised in some way. We shall make frequent use of the word, but with the following specific meaning:

A system is a collection of inter-related and interacting components that work together in an organised manner to fulfill a specific purpose or function.

An example is the water-supply system for a city and its components, which were described in Chapter 1. Another example is a city office building, the main components being the structural (load-bearing) skeleton, the cladding (the outer protective layer, or skin), the usable floors and inner spaces, the mechanical services (including elevators and air conditioning), the electrical services, the information technology services, the water supply service, and the waste and sewage extraction services. The function of this system is to provide a comfortable, sheltered, serviced working space in a convenient location for office workers.

The word *process* is also used frequently but rather vaguely in everyday conversation. We speak of manufacturing processes, deterioration processes, and political, biological and chemical processes. In such usage, the word suggests some sequence of events that involves a progressive change in time. The corrosion that occurs in the small metal fasteners that join two sheets of cladding metal together is a process. The slow, progressive, complex changes that occur in a rainforest over time, either naturally or as the result of pollution and atmospheric warming, constitute a process. In sewage treatment works, a number of chemical and biological processes are employed to treat the wastewater. The word *process* also describes the complex sequence of events that occur in a human brain. In this book we shall use the word process with the following meaning:

A process is a sequence of inter-related activities that proceeds in time.

The seasonal movements and related cracking that occur in a house constructed on a reactive soil, which swells in the presence of moisture and shrinks in its absence, constitute a process. The construction of a bridge and the design of a building are processes. The planning and the design phases of an engineering project are also processes. A process is sometimes created by a system or a component of a system.

2.3 SYSTEMS HIERARCHY

An engineering system consists of *components* which in turn are made up of *sub-components*. When considered closely, each sub-component is also a system in its own right, and made up of *sub-sub-components*. For example, a water treatment plant, which is a component of a water supply system, consists of various components, such as mixing and holding tanks, control valves and measuring devices. The holding tank is a complex construction, a system in itself, with floor and walls and possibly a roof constructed from reinforced concrete, and with entry and exit pipes that connect it to other components of the water treatment plant. The term *subsystem* is used interchangeably with component.

It is a characteristic of a system that it will itself act as a component in a larger system, in conjunction with other systems. Thus, the water supply system is a component of the city system, whose other components include a transportation system, a communication system, and a sewerage system. The term *system hierarchy* describes the structure in which a higher-level system is composed of components that are themselves systems, made up of sub-components (or sub-systems) and so on.

The term *system of systems* is a term with a similar meaning to system hierarchy. It focusses on the upward integration of systems and components and has been used in fields such as transport and defence to indicate a group of systems that work together to produce an outcome, so that the group is superior to the sum of its parts. The concepts of system and sub-system will be used in Chapter 3 to assist in identifying and defining specific engineering problems by treating them as parts of wider engineering problems.

Just as a physical system can be treated as a set of inter-related components, so can a process be broken down into a number of sub-processes. Each of these sub-processes can in turn be broken down into sub-sub-processes. The design process is undertaken in small steps which are, in effect, sub-processes. Of course, the decomposition concept is not in any way new or novel. In another context, Jonathan Swift wrote:

A flea hath smaller fleas that on him prey,
and these have smaller fleas to bite 'em.
And so proceed ad infinitum.

The processes of engineering planning and design

Complex engineering infrastructure projects are carried out in separate but interrelated phases, of planning, design, implementation and maintenance. Careful management is needed of the entire project, and of its various phases, in order to bring the project to a successful conclusion. Each main phase is in itself a complex process and, using the decomposition approach, is broken down progressively into many simple, inter-related tasks that can be studied and evaluated individually. In this way, a large engineering project, requiring many thousands of hours of work, can be completed in a relatively short time period by groups of people working simultaneously on different tasks. Clearly, great care needs to be exercised in the initial breakdown of the work, and also in the planning, management and continuing coordination of the work of the separate groups, in order to ensure that an integrated outcome will actually be achieved.

As the scale and complexity of a project increase, so too does the need for careful planning and design. A systematic approach to project planning has evolved progressively over many years and is used in all of the major engineering disciplines.

Impetus was given in the 1960s to the development of a "systems" approach to engineering planning and design by the need to coordinate extremely large national projects in the United States, including the NASA space program.

The processes of engineering planning and design are explained in Chapter 3. Various techniques that are used in the project planning phase are discussed in Chapter 5.

2.4 FURTHER SYSTEMS CONCEPTS

Some additional concepts are useful in the modelling, analysis and design of engineering systems. They apply also to processes, but for reasons of brevity our discussion will concentrate on their application to systems. The concepts apply to real-world systems, to processes and to the theoretical and physical models that we use to study real-world systems.

System boundaries

In many engineering systems it is easy to distinguish one component from another, because there are clear and distinct physical *boundaries*. This is the case with the components of the water supply system previously described. However, the boundaries are not always clear-cut, and in some situations it may be necessary to introduce artificial boundaries in order to separate a system into components. For example, arbitrary boundaries are used in large, complex structural systems, such as dam walls or building skeletons or aircraft frames, when a theoretical analysis is carried out using the so-called finite element method. In analysing a massive dam wall to determine the combined effects of water pressure, gravity and earthquake motion, the wall is conceptually broken up, using a grid of systematically located boundary surfaces, into a very large number of small brick-like pieces, or “finite elements”. By first analysing the behaviour of a typical individual element, and then the way it interacts locally with adjacent elements, the behaviour of the complete wall can be evaluated. The breakdown of the system into component elements in this case is not based on physical boundaries, and, from a physical point of view, is arbitrary.

Taking another example, the fuselage, wings and tail structure are broadly distinguishable components of the frame of an aircraft and in the design of an airframe it is advantageous to deal separately with these components. Although the components are broadly distinguishable, there are no precise boundaries. In such circumstances, arbitrary boundary surfaces are chosen, so that separate analyses can be carried out for each component. Of course, the local interactions across the boundaries are vitally important, and depend in part on where the boundaries have been located.

The concepts of system, sub-system, hierarchy and boundary provide a useful starting point for engineers to describe, define, analyse and design complex engineering systems and processes. With these concepts the processes of analysis and design are broken down into tasks of manageable size. The decomposition approach, coupled with idealisations and simplifications that are introduced in the modelling and analysis, allows us to deal with large engineering systems that would otherwise be far too complex to handle.

Input and output

No physical system, whether engineered or natural, is self-sufficient or cut off from the rest of the world. On investigation, a system will always be found to interact in some way with other neighbouring systems. In other words, each system is a component in a wider system. It is therefore necessary to look not only at an individual component, or for that matter at an individual system, but also at the way the system (or component) interacts with its neighbours. A system interacts with other systems by means of *inputs* and *outputs*.

These usually (but not always) take the form of a flow of matter, energy or information through the boundaries of the system. The inflow, or inputs, may be thought of as the influence of the rest of the world on the system we are considering. The output, or outflow, becomes input to other systems and hence represents the effect of the system on the rest of the world. A system (and also a component) can be represented schematically as a box with lines indicating inputs and outputs, as in Figure 2.1. The inputs and outputs may vary in time, and are therefore written as time functions: $y_1(t)$, $y_2(t)$, $y_3(t)$, $z_1(t)$, and $z_2(t)$.

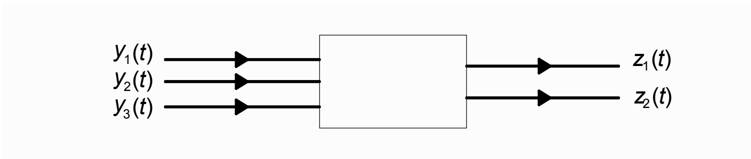


Figure 2.1 System with inputs and outputs.

A component in a system interacts with other components through its inputs and outputs. A system, consisting of components, each with inputs and outputs, can be represented schematically, as in Figure 2.2, by boxes to denote the components and by lines between the boxes to represent the inputs and outputs.

The input and output lines joining any two individual boxes represent the direct interactions between these two components. Additional indirect interactions may occur between two components via intermediate components. The inputs and outputs to the overall system are inputs and outputs to individual components, as shown in Figure 2.2.

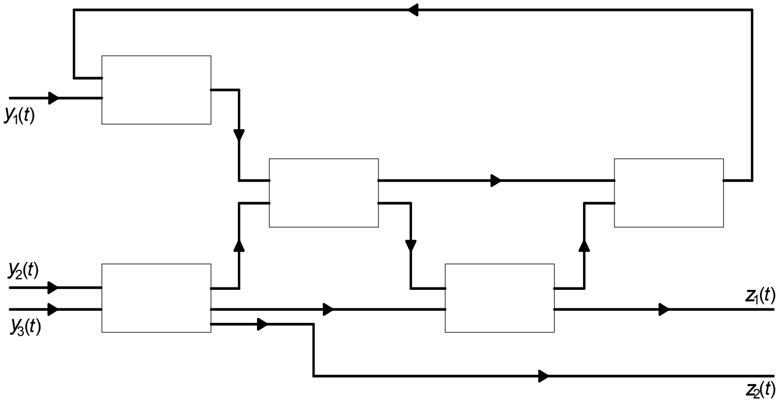


Figure 2.2 System and components.

Returning to the example of the water supply system, we have as system inputs the inflow of water into the reservoirs from the catchment areas. The quantity and quality of the water flowing from reservoir to reservoir and from component to component are input and output variables for the components.

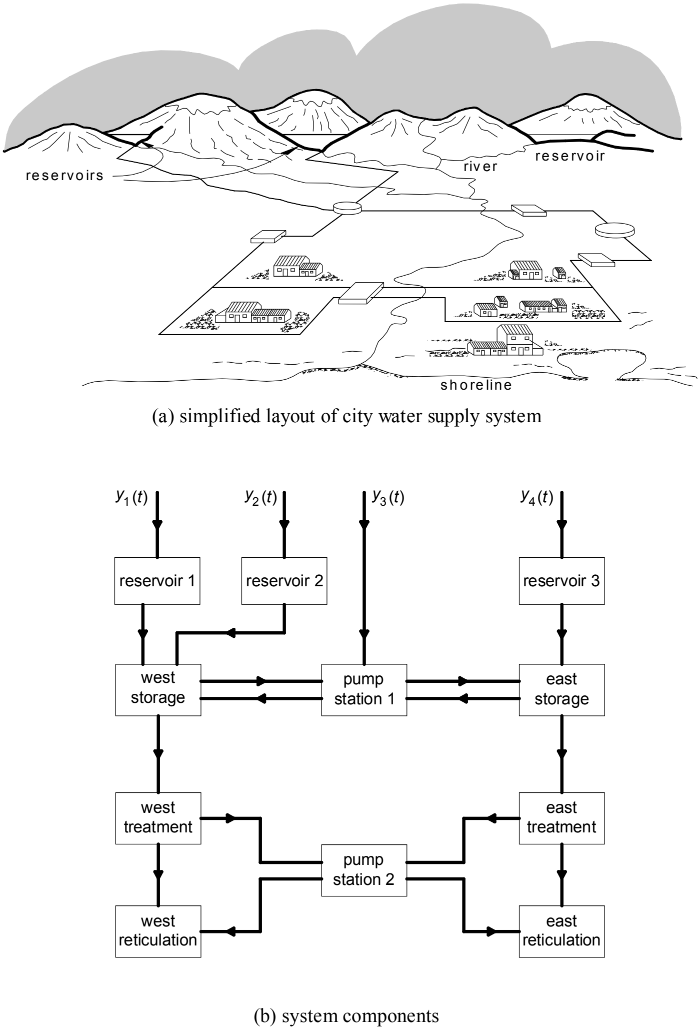
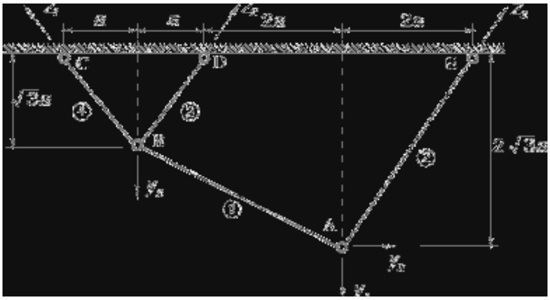
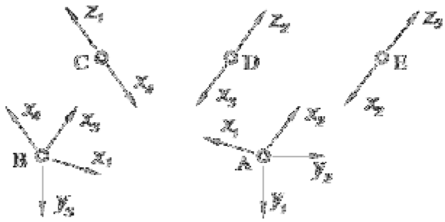


Figure 2.3 City water supply system.

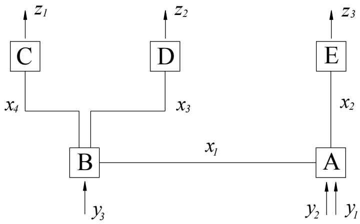
A schematic of the flow of water into, through and out of such a system is shown in Figure 2.3. This is simplified and does not show information on the quality and quantity of water in each component. Other arrangements of the components are of course possible, and there are other inputs and outputs, such as electrical energy and the chemicals used for water treatment, that have not been not included.



(a) tension-bar system



(b) forces acting at joints



(c) input-output schematic for joints

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1/2 & \sqrt{3}/2 & 0 & 0 \\ \sqrt{3}/2 & -1/2 & 0 & 0 \\ -1/2 & 0 & \sqrt{3}/2 & \sqrt{3}/2 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

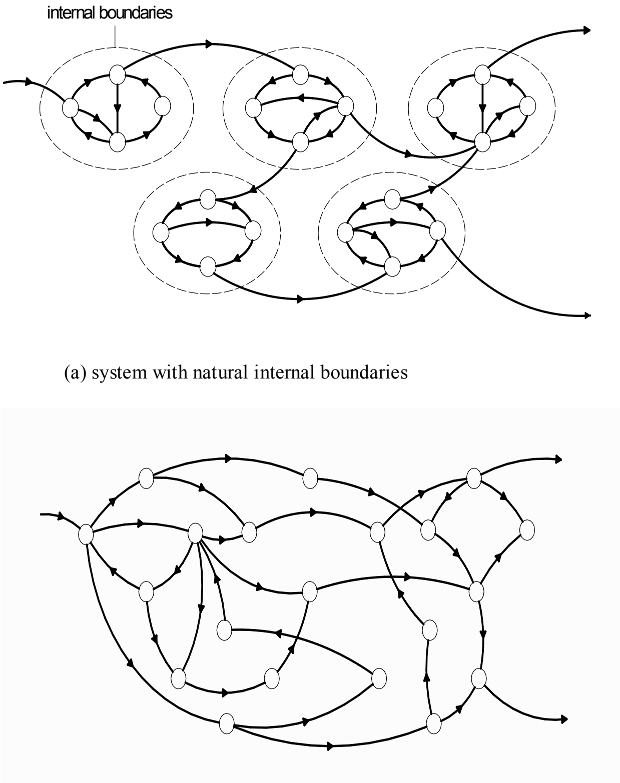
$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

(d) input-state, output-state relations

Figure 2.4 Structural tension-bar system.

An example of inputs and outputs that do not fall into the categories of matter, energy and information is provided by a structural load-bearing system which picks up applied loads and carries them into the supports, as shown in Figure 2.4. The pin-jointed frame is the system, while the tensile and compressive members and the joints are the components.

Most engineering students study such systems in introductory courses in the mechanics of solids. The inputs and outputs for these components are forces. The loads, applied at joints A and B, are the inputs to the system, while the reactions at joints C, D and E are the system outputs. The relations between output and input forces are presented in Figure 2.4(d). In a more complex structural system the inputs and outputs between neighbouring components might be stress fields that are distributed across the contact areas between components, together with the associated strain fields.



(a) system with natural internal boundaries

(b) system without natural internal boundaries

Figure 2.5 Component boundaries.

As we have seen, a real system might, but might not, display clear-cut natural internal boundaries among its components. The boundaries tend to be clear-cut if the sub-components separate themselves out from each other into clusters, with relatively few inputs from one cluster to another. Figure 2.5 shows schematics of two systems, one with natural internal boundaries, and the other without. Natural boundaries occur where there are fewest input-output connections among the sub-components. Even when it becomes necessary, for analysis or design, to break a complex system into arbitrary components, it is advantageous to choose the boundaries in such a way that the resulting inputs and outputs among the components are as few and as simple as possible.

Black-box systems and components

In order to achieve an overview of the functioning of a system component, it is not always necessary to study the internal functioning of the sub-components. In fact, by avoiding detail and by concentrating on the overall input-output relations among components, it is often possible to achieve a good overview of the behaviour of the system. When the detailed internal functioning within a component or system is ignored, and it is regarded simply as a device that transforms inputs into outputs, it can be thought of as a *black box*.

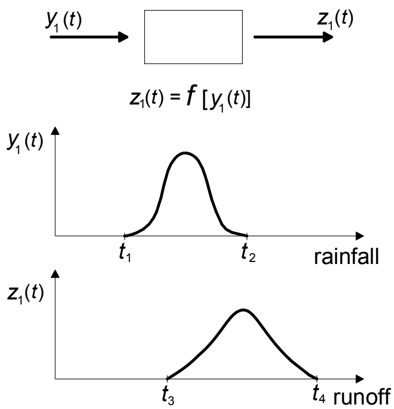


Figure 2.6 Black-box analysis: rainfall-runoff relations for a catchment area.

An example of a black-box representation of a component is the use of equations to calculate the run-off from a catchment area in terms of the rainfall history. In a very simple model, a time delay and a transformation are applied to the time history of the rainfall (the input) in order to represent the run-off (the output) as a second time history, as indicated in Figure 2.6.

A highly simplified and idealised treatment of a complete system can often be obtained by lumping closely linked components together and replacing them by black boxes with simple inputs and outputs. Such simplified treatments are often very valuable, for example at the preliminary stage of an analysis or of a design, when an accurate consideration of the functioning of the system is not needed. The idea of black-box components is used in electronic engineering for an arrangement of transistors, resistors and capacitors, assembled to fulfill a specific function. Such a device, designed to detect fish, is shown in Figure 2.7.

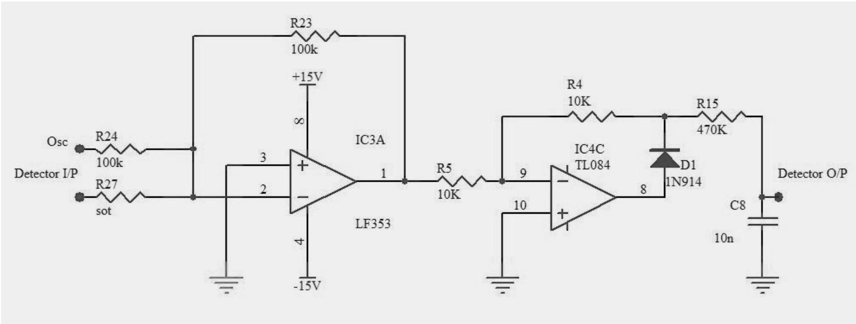


Figure 2.7 Electronic circuit for a fish detector. (Source: EngTest, University of Adelaide. "With Permission")

Control systems

Engineering systems operate with the purpose of achieving a specified goal. If the goal can be quantified, it may be possible to introduce a *control device* as a component to control the performance of the system and make it goal-seeking. The control device may be linked to a sensor that measures the condition of the system and sends the relevant information to the control device.

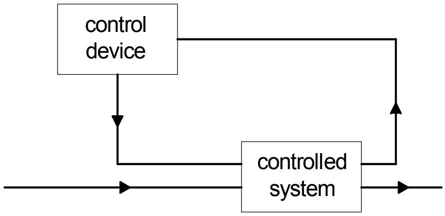


Figure 2.8 Closed-loop control system.

The actual condition (or state) can then be compared with a desired or target condition, and a control signal can then be sent to another component in the system to initiate changes that move the system progressively towards the target condition. This arrangement, shown in Figure 2.8, is a *feedback control system*, or *closed-loop control system.*

A simple example of a control system is an air-conditioned room (Figure 2.9). The desired air temperature is pre-set by the user. The thermometer senses the temperature in the room, and the control component switches the conditioning unit on when the difference between the measured and the target temperatures exceeds a specified margin.

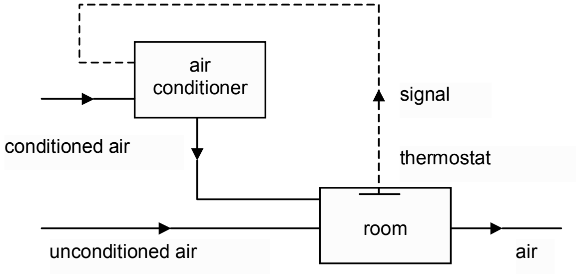


Figure 2.9 Air-conditioned room.

Another example is the traffic control lights at a road intersection, where sensors are used to detect and measure the volume of traffic flowing into and out of the intersection in all directions. A simple control device calculates the optimal green times for the flow of traffic in each direction, and a control signal is sent periodically to change the lights. While such localised traffic control systems are still used occasionally, it is far more common to use more sophisticated control systems to deal with the traffic in larger sectors of a city, including many intersections. The traffic density is measured continuously in key streets and intersections and the information is sent back to a central computer system which has been programmed to calculate a sequence of signal settings to control local flows and minimise overall flow time through the road system.

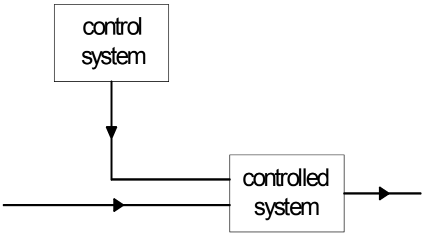


Figure 2.10 Open-loop control system

A simpler, but less effective arrangement for a control system is shown in Figure 2.10. The feedback control loop has been omitted and so the arrangement is called, somewhat illogically, an *open-loop control system*. As an example, we may consider a set of traffic lights at an intersection that operates without traffic flow counters. The pre-set timing of the lights is chosen to allow for, say, normal peak hour traffic flow. Obviously such a system cannot deal successfully with unexpected or unusual events, such as traffic accidents or heavy rain.

The state of a system

The concept of system *state* is used in analytic studies of the behaviour of dynamic systems, that is to say, systems that experience progressive changes in time. The state of such a system is represented by a set of key variables, called the *state variables*. The values of the state variables at any given instant in time completely define the condition of the system and take into account the effects of all previous states and all inputs up to that instant. The state variables are used to predict the state of the system at succeeding times (Ashby, 1965a, 1965b).

When a medical practitioner investigates the state of health of a patient by measuring blood pressure, heart rate, condition of the eyes, chemistry of the blood and urine (counts of red and white blood cells, levels of cholesterol, triglycerol, uric acid, and sugar), the values represent a primitive set of state variables which indicate, very approximately, the state of health of the patient. The patient is of course a highly complex biological system, and is much more complex than the simplified state model being used in the health check.

The idea of system state contrasts with the black-box representation previously described, because, in evaluating the state of the system, we look at its internal functioning and not just at the inputs and outputs. The treatment of system behaviour as a sequence of changing states will be discussed further when we look at methods of analysis.

From a formal mathematical viewpoint, the state of a system can be represented using a multi-dimensional space in which each state variable has an axis allocated to it. If there are n state variables, then the *state space* has n dimensions. The state of the system is represented by a point in state space and the behaviour of the system over time traces a line in state space. Although state space concepts are used mainly for dynamic systems, the idea of state can be seen in Figure 2.4, where the state of the tension-bar system is represented by the four forces in the bars, x_1, x_2, x_3 and x_4. The state variables in this case depend purely on the system inputs, which are the forces y_1, y_2 and y_3. An example of state space analysis for a linear dynamic system is presented in the appendix to this chapter.

Equilibrium and stability

Equilibrium and stability are vitally important concepts in engineering, and indeed in many other fields of study. These ideas are best explained using state space concepts. A system is in an *equilibrium state* if there is no tendency for it to move away from that state unless it is forced by external disturbances. The equilibrium state is represented by a point in state space. Depending on the complexity of the system, there may be more than one equilibrium state, or no equilibrium state at all. In a very complex non-linear system, there may even be small regions of state space in which the system will remain, moving around in a near-equilibrium condition unless it is forced away by external inputs. On the other hand, a system can be kept in a state of pseudo-equilibrium by controlling the inputs to maintain the state of the system at or close to the desired point in state space. An example of this is when a helicopter hovers: it is being held close to a point in three-dimensional space by the pilot.

The concept of system equilibrium is applicable in many fields other than engineering, such as economics and biology. An example of an engineering system in equilibrium is a building structure that safely supports the vertical loads on its floors and resists the horizontal wind loads that act on its external walls. The state variables of the system include the internal forces induced in the component members in response to the loads, and also the deformations and deflections that result from the loads.

We speak about the *stability* of a system in relation to its equilibrium states. A system is in a stable state of equilibrium if it returns to the same equilibrium state after it has been disturbed. If, following a small disturbance, the system moves progressively away from an equilibrium state, then that state of equilibrium is unstable. A system is in a state of *neutral equilibrium* if, following a slight disturbance, it does not return to the equilibrium state, nor move further away.

A building structure, consisting of slabs supported on beams and columns and carrying vertical floor loads, can be tested for stability by applying horizontal loads at the different floor levels and then removing them. The test may be carried out mathematically on a theoretical model, in the laboratory on a physical model, or may even be carried out on the real system. If the system returns to its original state with the same original internal forces upon the removal of the disturbing horizontal loads, it is in a stable equilibrium state. If, however, the building is potentially unstable, the small horizontal forces could induce large lateral displacements with buckling failure of the columns, followed by complete collapse of the building.

Of course, the response of a system to a disturbance depends on the magnitude of the disturbance. If the disturbance is large, a stable system may be damaged or even destroyed and will not then be capable of returning to its equilibrium state. A bomb blast or earthquake, if sufficiently intense, will bring an otherwise stable structural frame to the point of destruction. It is therefore useful to describe a stable system as one for which the response to a disturbance is *not out of proportion* to the magnitude of the disturbance. It is frequently necessary to investigate the equilibrium condition (or conditions) of a system and its stability in relation to its equilibrium states. Methods of *equilibrium analysis* and *stability analysis* are available, both for static and dynamic systems (Bazant and Cedolin, 1991).

System coefficients

The main system variables consist of the input and output variables and the state variables. The latter represent the progressive behaviour of the system in time, or, more precisely, the behaviour of the system model. However, various other quantities are needed to define the model. These are called the *system coefficients*. They are inherent properties of the system and its components. System coefficients, in contrast with the system variables, are usually fixed in value and represent the unchanging physical characteristics of the system. For example, in the frame in Figure 2.4 the lengths of the component members, the coordinates of the various joints, and the stiffness properties of the components are the system coefficients that are used to model the structural system. System coefficients typically include the properties of the materials comprising the system, and its geometry. Unlike the state variables and the output variables, they do not usually change in value in response to changes in the input variables.

Nevertheless, some system coefficients may change slowly in time, not in response to the applied inputs but because the system itself is gradually changing.

For example, some engineering materials change their properties in time. The modulus of elasticity of concrete (its stiffness) increases slowly with time, because of complex chemical changes that occur in the material. Progressive changes in system coefficients can occur if there is deterioration, as for example when there is corrosion of steel, abrasive removal of layers of material in a pipe, or scouring and erosion of the bed of a stream.

If one or more of the system coefficients vary significantly in time, then it may be advisable to restructure the model of the system, to treat the varying quantities not as coefficients but as system variables. This is the case when stream erosion is severe and a stream bed moves progressively over time, thus changing the geometry of the system. When this occurs, it becomes necessary to move from a fixed bed model to a movable bed model that can deal with the erosion.

Optimisation

An engineering system has to have a clearly stated set of goals and objectives if it is to be designed, created and operated. Identification of the goals and objectives is an important step in the processes of planning and design. The aim of the engineer is to achieve the best possible solution, or possibly the most economical solution, for any constraints that have to be satisfied. The term *optimisation* refers to the process of finding the best, or *optimal*, solution.

Optimisation techniques can sometimes be employed to find the best solution to a mathematically formulated problem, irrespective of its physical nature. In optimising a system or a process, it is necessary not only to identify the goals and objectives but also to quantify these goals. The process of optimisation is otherwise useless. Specific optimisation techniques are discussed in Chapter 13 of this book.

The error of sub-optimisation

It is important to understand that optimisation of the overall system is *not* usually achieved by independently optimising each of its individual components. Indeed, such an approach can lead to serious malfunction of a complex system. The optimisation of even one of the components in isolation may be detrimental to the overall system. The term, *error of sub-optimisation*, is used for this important idea and refers to the process of optimising one or more system components to the detriment of overall system performance.

Examples of sub-optimisation should make the matter clear. In the design of an aircraft, the aim is to achieve optimal overall performance. If the engines are designed to achieve maximum performance without due regard to their fitting into the airframe, to their size and weight, and to the way they work in conjunction with the airframe, then the result will be less than optimal as far as the aircraft is concerned. Indeed if the engines were designed disregarding the other components, it is unlikely that the aircraft would fly.

A type of sub-optimisation error can occur in the structural design of buildings if a minimum weight solution is sought, on the assumption that this will lead to a minimum use of materials and hence to minimum cost. In fact, minimum weight rarely leads to minimum cost. Structural systems designed for minimum use of materials are usually extremely difficult to construct, and therefore very costly. The additional costs of construction far outweigh the savings in materials in a typical building structure. On the other hand, minimum weight is far more important in the design of a space vehicle because of the advantages in regard to payload and fuel use.

The design of an engineering system is always a compromise, whereby each component has to be tuned, not to perform optimally in itself, but to contribute to the optimal performance of the overall system. Sub-optimisation errors are prone to occur in large projects when insufficient contact and communication occur among the different groups of engineers working on the different components, which eventually have to be brought together. An important task in project planning and management is to avoid, or at least minimise, errors of sub-optimisation.

2.5 MODELLING AND ANALYSIS

In order both to create new pieces of infrastructure and to maintain and improve existing ones, we must be able to predict how real engineering systems and processes will behave in the future, under both normal and unusual operating conditions. Often we have to make such predictions before the systems and processes have even been created. We rely on *modelling* and *analysis* to do this. These terms are closely related in meaning, and are sometimes used interchangeably. Nevertheless, it will be useful for us to distinguish between the terms.

Mathematical models

The process of creating a simplified, idealised representation of a real-world system is referred to here as *modelling*, and the simplified representation is a *model*. The purpose of the model is to mimic the behaviour of the real system to an acceptable level of accuracy. The model is the starting point for *analysis*, which is the term that describes the activity of using the model in order to predict how the real engineering system and its components will behave when subjected to a specific set of operating conditions and environmental conditions.

In current engineering work, a model is most often in the form of a set of mathematical equations or computer algorithms, which we call a *mathematical model*. The analysis is then theoretical, and involves computations that are often complex, and usually undertaken by computer. Before powerful computing facilities became available, much greater reliance had to be placed by engineers on *physical models*, which will be discussed shortly.

Use of mathematical models in theoretical analysis

The improvements in computing hardware and software that have occurred in recent decades have allowed mathematical modelling and analysis to be undertaken for a wide range of engineering problems.

The processes of mathematical modelling and analysis of a physical system are indicated in the schematic in Figure 2.11, which shows the steps used to predict how an existing system or a proposed new system will respond to a specific set of operating conditions and inputs. There are three processes in Figure 2.11 and they are represented by the arrowed lines denoted as A, B, and C.

Process A represents the creation of a conceptual model that consists of a set of equations or algorithms. This is a simplified representation of the real system, but it shows how the real system will behave under a variety of different conditions and inputs. Process B is the theoretical analysis of the system model for a specific set of operating conditions and inputs. Approximations and idealisations always have to be made in the modelling process, and sometimes also in the calculations that comprise Process B. It is therefore necessary to consider carefully the effects on the predicted behaviour of all the simplifications and approximations that have been introduced in Processes A and B. This is the role of Process C. In Process C, the behaviour of the real-world system is predicted using the results of the analysis of the theoretical model, and information about the approximations made in Processes A and B. Only if the modelling Process A and the analysis Process B are very accurate, will the model behaviour be a very good prediction of the behaviour of the real system.

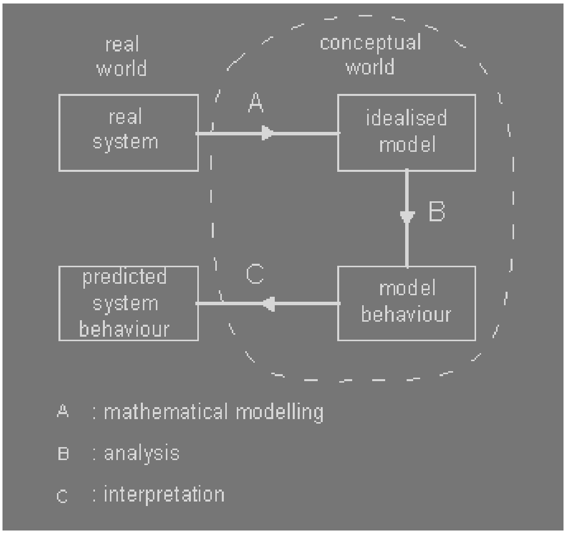


Figure 2.11 Modelling and analysis to predict system behaviour.

To complete the picture, the operation of the real system, under the same operating conditions and inputs as were assumed in the analysis, is shown as Process D in Figure 2.12. There are always differences between the behaviour of the real system and the predictions of the modelling and analysis, but the more accurate the modelling, analysis and interpretation, the closer will be the correspondence between real behaviour and predicted behaviour. Analysis is an essential tool in the design of engineering systems.

Physical models

As already mentioned, there are some systems that are too complex to lend themselves to theoretical analysis or computer simulation. They require physical modelling. The physical model may be the same size as the physical system to be studied, but it will often be necessary to scale-up or scale-down the model so that it can be created and tested in the laboratory. For example, complex coastal engineering processes are often studied using scaled-down physical models, which are constructed in large laboratory tanks.

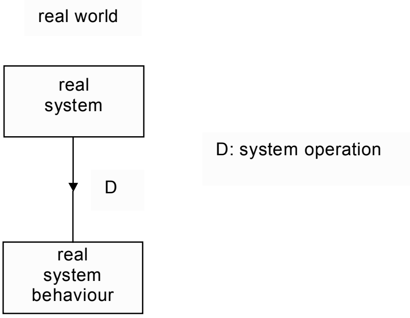


Figure 2.12 Real system behaviour.

The experimental results from tests on a physical model are used to predict how the real system will behave under a given set of conditions. In such cases the analysis phase requires an interpretation of the test data, with allowance for various factors that depend on the modelling, such as size effect.

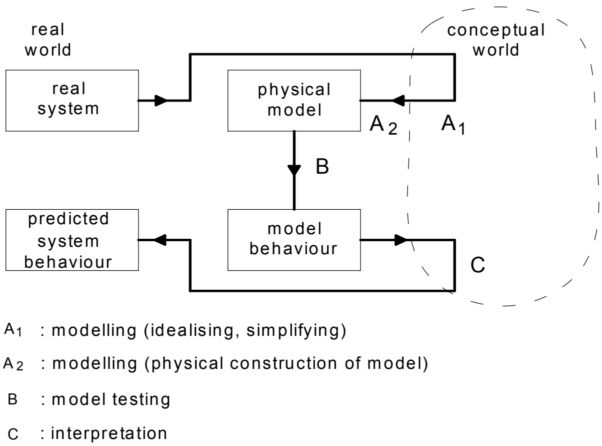


Figure 2.13 Use of physical modelling to predict system behaviour.

The steps in the creation and use of a physical model are shown as A_1, A_2, B and C in Figure 2.13. Steps A_1 and A_2 together represent the creation of the model, which begins with some conceptualisation (Step A_1) and then the physical work of constructing the model (Step A_2). In the construction of a physical model, simplification and idealisation are just as necessary as in the case of mathematical modelling. This occurs in Step A_1.

Physical models that have a similar appearance to the original system are sometimes referred to as *iconic models*. It is also possible to create *dissimilar models* that do not resemble the original physical system at all. For example, electrical circuits can be used to represent groundwater flow. Such dissimilar models can be created when the same set of underlying mathematical relations applies to two different physical situations.

If a physical system is going to be used in large quantities, it may be possible to test prototypes. A full scale prototype axle shaft for a new car would for example be tested in the laboratory for fatigue, while complete cars would be used in safety tests. On the other hand, a chemical engineer may create a scaled-down laboratory model of a proposed chemical engineering process. Structural engineers often construct small-scale physical models of tall buildings to observe the effects of complex gust loadings in a wind tunnel.

A scaled-down, physical, iconic model of a coastal breakwater in a laboratory is shown in Figure 2.14. An adjustable wave-making machine, seen in the foreground, simulates extreme weather conditions. Using this physical model, various breakwater geometries can be tested in the laboratory and their performances compared.

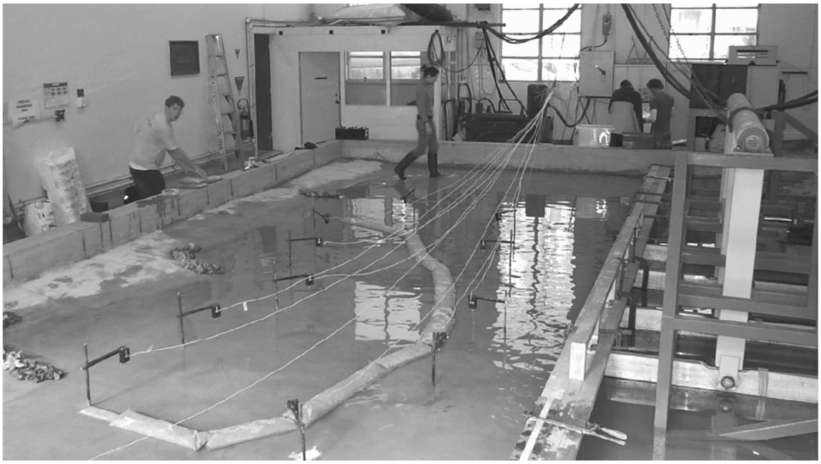


Figure 2.14 Physical model of a coastal breakwater.

Dealing with too much information

In order to create a model (either physical or mathematical) of an engineering system, the problem of complexity has to be dealt with. The concepts of system and hierarchy can be useful in this regard. However, there is another related problem. Simply put, there is usually too much information available about a real system, and most of the information will be either irrelevant or of marginal value when a model (either mathematical or physical) is to be created. It is necessary to separate out the aspects of the system that have an important influence on the behaviour that is to be studied, and those aspects which are of marginal or zero importance. An important step in modelling thus involves discarding irrelevant information. Initially it may be difficult to decide which information to keep and which to discard. Engineering experience and judgement are relevant, and the decision becomes easier when the behaviour of the system is better understood. The amount of information retained (if carefully chosen) can also determine how accurate the model will be.

If we wish to model a component of a bridge, say a slab which carries wheel loads and transfers them to a supporting girder, the factors of prime structural importance will be the size and shape of the slab and the stiffness and strength properties of the materials. The colour of the components and the surface details of the road on top of the slab will not be of relevance and will be discarded, even for the most accurate analysis. However, this information could be important if an analysis of temperature effects were to be undertaken. Separating out the important from the unimportant is equally necessary in mathematical modelling and in constructing a physical model.

Occam's Razor

William of Ockham (or Occam) was an English philosopher who lived in the 14th Century. He was an empiricist who wrote on science and logic, and argued that logic should be separated from theology and metaphysics. His statement in Latin, and in English translation, has become known as Occam's Razor:

Essentia non sunt multiplicanda praeter necessitatem.

(Do not multiply hypotheses beyond necessity.)

In the context of engineering modelling and analysis, Occam's Razor is an argument for simplicity. According to this principle, of two equally accurate theories or explanations the simpler is always to be preferred. It follows that we should keep our models as simple as possible, and introduce no more assumptions than are absolutely necessary. A modern equivalent might be the well-known motto used in engineering design: KISS (Keep It Simple, Stupid).

The trade-off between complexity and accuracy

In engineering modelling it is not possible or even desirable to take full account of all of the complexities of the real system. Some trade-off between accuracy and simplicity is always made in the modelling process. It follows that there can be no single "correct" model of a real system. On the contrary, there will be an unlimited number of possible models that vary in complexity and accuracy, depending on the degree of idealisation employed in the modelling process. There will rarely be a single "best" model.

Although a simple model will necessarily have limited accuracy, it has the advantage that quick (but approximate) answers can be obtained, and, possibly, a good overview of the functioning of the system. It is often of benefit to designers and analysts to have several alternative models available, including a simple, rough model and an accurate, detailed model. Designers will use simple models in the early conceptual stages of the design process and the accurate, more complex, models when they are working out the final design details. A very simple calculation, using a simplified model, can also be used to make order-of-magnitude checks on a very complex set of computer calculations.

Modelling and Formula 1 cars

A key aim in the design of Formula 1 cars is to minimise the drag they experience when they are being driven at high speed. Early difficulties in mathematical modelling led to the use of physical modelling to predict drag. This involves much more than building the car to a convenient scale and then blowing air past it. The way the system behaves must be understood and the tests designed accordingly. For example, It is relatively easy to build a 1/10th scale model of a car, but there are questions to be answered, such as: what air speed to use? If it is a 1/10th scale model, then perhaps the air speed should also be 1/10 of that expected, say 30 km/h to model the full scale 300 km/h? This is not correct. In fact, drag is an inverse function of scale, so that if the physical dimensions are reduced by a factor of 10, then the air speed must be increased by a factor of 10! The correct answer would be 3000 km/h. At this speed there are other problems (such as breaking the sound barrier) that spoil the modelling. It is necessary to test at as large a scale as possible, and this means that a huge wind tunnel would be needed to test a full size car.

Problems with testing large-scale physical models led to the development of sophisticated computer models that can deal accurately with many of the processes involved in drag. These now provide a viable alternative to testing physical models. One computer model, called FLUENT, has been developed by Sauber Petronas and Fluent Inc. It computes and displays the flow lines around a formula one car. For details see: http://www.fluent.com/about/news/pr/pr81.htm.

An analytical model can only be created to predict the behaviour of a system if the physical principles that govern behaviour are well understood. The model may then be a set of equations that comprise a mathematical model. With the advent and widespread use of powerful computer facilities, it is common to use mathematical models, with lengthy computations, to study the behaviour of quite complex engineering systems. Today, mathematical models are normally cheaper to prepare and use than physical models.

Reliability and numerical stability

To be acceptable and useful, a model has to satisfy a number of important requirements. First and foremost, it must capture the essential behaviour of the real system, otherwise it could lead to erroneous and dangerous conclusions. A process of *validation* is used to check the reliability of the model and to determine its likely accuracy, and the order of magnitude of prediction errors. There are various ways of validating a model, but generally this requires comparisons between model predictions and the observed behaviour of a real physical system. If the modelling is being undertaken for the design of a new, not yet existing system, then verification can be undertaken by comparing model predictions with the behaviour of similar real systems that have been previously constructed and tested. Physical models, tested in the laboratory, can also provide data for evaluating the adequacy and accuracy of a conceptual model. In some situations it is possible to calibrate a new model against an existing model that has itself been validated.

The potential accuracy of any analysis depends on the underlying model. However, misuse of a model can lead to inaccuracy, with dangerous results. A careful interpretation of the results of every analysis is therefore needed to identify possible differences between the real system and the model. This is an essential step in any analysis and was referred to in the discussion of Figures 2.11 and 2.12.

Validation becomes crucially important when we model complex systems, particularly non-linear systems, because some effects, such as positive feedback, are non-intuitive and are difficult to recognise. Unfortunately, the more complex the system, the more difficult it is to validate the model.

Yet another requirement of a conceptual mathematical model is *numerical stability*. That is to say, computational problems should not arise when the model is used to analyse a range of different systems, or a single system with varying coefficients and varying inputs and operating conditions. Numerical instability is a common bugbear of mathematical modelling, and is particularly likely to appear when non-linear system models are being used.

Numerical instability can lead to non-convergence of certain operations, so that the calculations cannot proceed to their end. Even when the computations proceed to a conclusion, numerical instability can lead to serious errors in the results.

2.6 ENGINEERING ANALYSIS

Many different types of analysis can be carried out for a given physical system or problem, depending not only on the specific aspects of behaviour to be investigated, but also on the type and accuracy of the model being used. We shall look briefly at some of the forms of engineering analysis that are commonly undertaken.

If the real system displays significant variability in its response to repeated, apparently identical inputs, then it may be necessary to undertake a *probabilistic analysis*. Alternatively, a *statistical analysis* could be undertaken with the coefficients and inputs treated as random variables. Another possibility is to use a deterministic model, but to carry out a *sensitivity analysis* by making a large number of computational runs while systematically varying the values of the system inputs and coefficients. Sensitivity analyses can also be used when the coefficients or inputs are not known precisely, but only statistically. A large number of runs might also be undertaken with randomly generated values of the coefficients and inputs, in order to look at the variability to be expected.

On the other hand, variability might be ignored in favour of a simple, deterministic *input-output analysis* for the overall system. This could be appropriate when approximate estimates of behaviour are needed. An input-output analysis may be thought of as a causal relationship between input (cause) and output (effect). Alternatively, it may be useful to consider that the system executes a process that transforms the input into output.

In a more basic analysis, the behaviour of the individual components will be examined, as well as the way they interact with each other. Scientific and engineering principles will be drawn on from the areas of knowledge relevant to the particular physical system.

It is a characteristic of real systems that their behaviour changes over time. If the changes occur rapidly, dynamic effects become important and so a *dynamic system analysis* should be undertaken, using a dynamic system model. *State space analysis* is used to study dynamic system behaviour. An advantage of state space analysis is that the equations of state are in a very convenient form for computation. The n equations of state for an n-dimensional system show the rate of change of each state variable with respect to time, as a function of the current values of all the state variables. This is very easy to program for a computer. An example of state-space analysis, with a set of state equations, is given in the appendix to this chapter.

If the changes do not take place rapidly over time, it may be possible to ignore the time dependent effects and use a *static* (or *steady-state*) *analysis*. For example, in modelling a bridge to study the effects of highway traffic, consideration will be given to braking, impact and collision and this will require dynamic analysis; on the other hand, dynamic loads might be approximated by static loads in modelling a small building constructed in a seismically inactive area.

Computer methods and computer simulation

A computer is normally used to carry out the detailed calculations required for the analysis of a complex engineering system. This increases enormously the quantity of information that can be generated. In fact, the form of the analysis and the form of the underlying model are sometimes adapted to suit the available computer software. This is because a large number of standardised computer software packages are commercially available in most engineering fields that can be applied to problems of analysis and design.

The software packages are written to deal with a general class of problem. Thus it is often the case in engineering work that an appropriate software package is first chosen, and then the modelling and analysis decisions are undertaken in a way that is compatible with the software.

Computer simulation of the behaviour of a system is made by calculating successive values of the state variables (or input-output variables) over an extended period of time for a prescribed set of operating conditions and system coefficients. The aim of computer simulation is to give a detailed picture of the way the real system will behave should it be subjected to the operating conditions chosen for the analysis. System simulations can be repeated many times with systematically varied values of the coefficients and inputs in order to observe the sensitivity of the predicted behaviour to variations in the system properties. A sequence of such calculations has previously been referred to as a "sensitivity analysis."

The term *Monte-Carlo analysis* is used to describe repeated analyses undertaken with randomly chosen variations in system characteristics and inputs. Such analyses can be used to determine the effects of random variations on system performance. This in turn provides a means of studying stochastic system effects. The values of the coefficients and inputs are determined by random sampling from frequency distributions which represent variable real-world quantities.

Modelling complex physical systems and processes

The modelling work that has been undertaken in an effort to manage and maintain the River Murray mouth in South Australia provides a good example of how a range of models of varying complexity can be used. The aerial photograph below shows the mouth as it was in April 2000, choked by a large influx of sand and with a small number of channels allowing water to flow into the lagoon adjacent to the land.

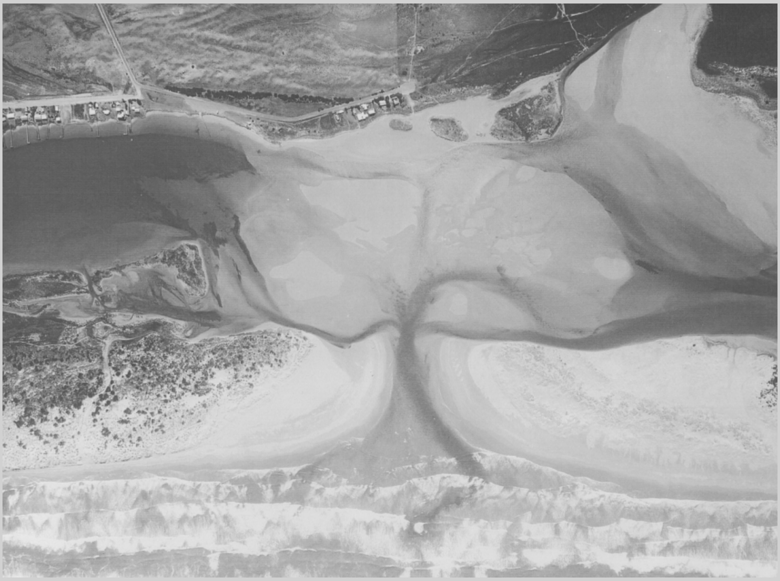


Figure 2.15 Areal view of chocked mouth of the River Murray (2000).

Reduced river flows due to increased irrigation and upstream usage had led to large deposits of sand inside the mouth area, brought by the tides that flood in twice daily. In an optimisation study of river discharges, two numerical models were developed: a simple, one-dimensional model that represented the state of the mouth by a single number (the cross-sectional area), and a complex three-dimensional model of the mouth, its channels and sandy islands over an area of several square kilometres. While the detailed three-dimensional model could be used to evaluate discharge options and to give predictions of which channels would be choked with sediment and which would run freely, it took approximately one day of computer time to simulate one day of operation. Simulations of anything over a month were thus impossible due to time constraints. While the one-dimensional model could not predict what was happening to individual channels, it did allow a 100-year simulation of the mouth size to be undertaken in a matter of minutes and gave a useful prediction of the long-term fate of the mouth of the river.

The results of the large numbers of analyses can then be used to create frequency distributions to represent system behaviour, thus avoiding a probabilistic analysis of the problem.

In the case of complex non-linear systems, behaviour can be almost impossible to anticipate intuitively. Modelling, analysis and validation then become critically important. The only feasible approach is to use models that have been developed specifically for computer analysis. State space concepts and state space analysis are very useful in such circumstances.

Global Warming: modelling and analysis

The earth, together with its oceans and atmosphere, forms what is now known to be a highly complex system that is acted on by the sun and moon and other celestial bodies. This understanding is relatively new and it is interesting to see how it has developed (and continues to develop) with time.

In the early 19th century Joseph Fourier determined that without the earth's atmosphere, normal surface temperatures would be below freezing. In 1859 John Tyndall tested a number of the atmosphere's components and found that while oxygen and nitrogen were ineffective at trapping heat, methane and carbon dioxide were effective. The first two greenhouse gases had been identified. In 1896 Svante Arrhenius suggested that if for some reason carbon dioxide levels were to rise, for example as a result of increased volcanic activity, this could lead to a general warming which in turn would increase atmospheric moisture, and this would create a feedback loop and further warming. These processes, including the feedback between temperature and atmospheric moisture, might be shown as in Figure 2.16, which is of course a highly idealised model of the atmosphere because it ignores many other important processes.

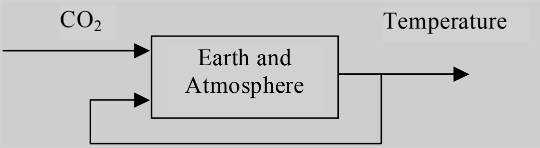


Figure 2.16 Feedback loop in global warming.

Other factors that are now believed to be important include cloud cover, ice and snow cover, the oceans and their ability to store heat and absorb carbon dioxide, the ocean currents, the earth's ecosystems that act as both sources and sinks for carbon dioxide, man-made features such as aeroplanes and their exhaust contrails. The list goes on. Many of these have significant feedback effects and it is for this reason that there are so many arguments about global climate change. In all this uncertainty, one thing is certain: it would not be possible to model the climate without a systems approach to describe and quantify the actions, reactions and interactions between all the processes at work.

Complex, counter-intuitive behaviour in systems was demonstrated by Forrester (1971), one of the early pioneers in the use of numerical modelling and computer simulation of complex system behaviour. Forrester used the idea of positive and negative feedback loops to show how apparently simple changes can lead to unexpected and drastic outcomes. An example of a positive feedback loop is the effect of birth rate on population. Over time, even a small positive birth rate leads to an exponentially increasing population because of the compounding effect. In contrast, the death rate has a negative feedback effect on population. When several such opposing feedback effects occur simultaneously, the results can be very difficult to anticipate.

An example of modelling a system with such feedback effects was given by Dörner (1996), who investigated the problem of controlling excessively bad odours in a small pond. Some of the feedback effects he considered were related to temperature, oxygen, organic and inorganic substances present in the pond, and the anaerobic microorganism population.

2.7 SUMMARY

System concepts are valuable tools in the work of engineering planning and design. In this chapter we have introduced the ideas of system and process, sub-components and hierarchy, input and output, state, equilibrium and stability, and system goals and optimisation. A system model is a simplified and idealised representation of a real physical system or process. It may be a physical laboratory model or a conceptual (mathematical) model. Conceptual models provide the starting point for mathematical analyses that predict how a real system behaves under a given set of operating conditions. The process of engineering modelling has been briefly described and some of the main types of analysis have been discussed.

System concepts are also useful in dealing with the complexity that is inherent in engineering systems and in problem solving. A decomposition process is used, whereby a system is broken down progressively into smaller components and sub-components that are so simple that we can study how they behave individually and how they interact with each other. Complexity is also dealt with by discarding irrelevant information.

In Chapter 3 we will see how these concepts are used in planning and design.

PROBLEMS

2.1 Consider the airport for a city that caters for both local and international flights:

(a) What are the main components (i.e. sub-systems) of the airport system?
(b) Choose several components and identify the sub-sub-systems.
(c) What are the main inputs and outputs for the airport system?
(d) In the planning and design of this airport, which processes should be analysed quantitatively? Which pieces of hardware should be analysed quantitatively?
(e) Which variables could be used to represent the state of the airport system at any time instant?

2.2 Answer the questions listed in Problem 2.1 in relation to a ten-storey city office building.

2.3 Identify and describe the functioning of some of the control systems to be found in the house or apartment where you live, and in the car you drive and the bicycle you ride.

2.4 What would you choose as the state variables when analysing the dynamic behaviour of a diving board above a swimming pool? Which are the system coefficients?

2.5 Figure 2.7 in this chapter represents the layout of a city water supply system, but no outputs are shown. What are the outputs? What happens to these outputs? How does this system link up with other components of the city system?

2.6 Identify possible sub-optimisation errors in the design of a road network you are familiar with, and in the building where you live.

2.7 What simplifications and idealisations have been made in the derivation of the equations of state in the Appendix to this chapter?

REFERENCES

Ashby, W. R. 1965(a). *An Introduction to Cybernetics*. London: Methuen.
Ashby, W. R. 1965(b). *Design for a Brain*. London: Methuen.
Bazant, Z. P. & L. Cedolin 1991. *Stability of Structures*. Oxford: Oxford University Press.
Chapra, S. C. & R. P. Canale 1998. *Numerical Methods of Engineering*, 3rd Ed. New York: WCB McGraw-Hill.
Coates, J. F. 2000. Innovation in the future of engineering design. *Technological Forecasting and Social Change*, 64, 121–132.
Dörner, D. 1996. *The Logic of Failure*. New York: Basic Books.
Elgerd, O. 1967. *Control Systems Theory*. Tokyo: McGraw-Hill Kogakusha.
Forrester, J. W., 1971. *World Dynamics*, Cambridge (Mass): Wright-Allen Press.
McLoughlin, J. B. 1969. *Urban and Regional Planning, A Systems approach*. London: Faber & Faber.

APPENDIX 2A: EXAMPLE OF STATE-SPACE ANALYSIS

To show how a set of equations of state can be developed, we consider the dynamic analysis of the two-storey frame shown in Figure 2.17. The building is subjected to horizontal loads due to wind or earthquake.

To idealise and simplify our analysis, we assume that the floors are stiff in comparison to the columns and move laterally under the influence of the horizontal forces, as indicated in Figure 2.17(a), and do not bend. To further simplify our model, we assume that the wind forces $y_1(t)$ and $y_2(t)$, act at the first and second floor levels. They are the input variables, and vary over time. As the floors move horizontally, the supporting columns deflect accordingly, as in Figure 2.17(a).

We now choose appropriate state variables to represent the condition of the system. The horizontal displacements at the floor levels, $x_1(t)$ and $x_2(t)$, are clearly appropriate choices, but these two variables alone are not sufficient to specify the state of the system: initial values of these two variables at some starting time t_0, $x_1(t_0)$ and $x_2(t_0)$, do not contain enough information to allow all future states to be calculated when an input history, $y_1(t)$ and $y_2(t)$, is stipulated.

We need to include two further state variables and we choose the velocities of the floors as the additional variables. Denoting the velocities as x_3 and x_4, we have:

$$x_3 = \frac{dx_1}{dt} = \dot{x}_1 \tag{2.1}$$

$$x_4 = \frac{dx_2}{dt} = \dot{x}_2 \tag{2.2}$$

If initial values of these four state variables are known at time t_0, a dynamic analysis will allow us to determine all subsequent states of the system for a known history of the input variables. The equations of state for this simplified and idealised system are derived from simple dynamic principles.

The top floor is represented as a point mass which is accelerating under the resultant of the two forces acting on it. One of these forces is the applied load $y_2(t)$; the other is the restraining force exerted on the floor by the tops of both second storey columns. The stiffness of the two columns together is k_2 Newtons per mm. This means that a force of k_2 Newtons, applied to the columns at the top floor level, would produce a movement of 1 mm, relative to the bottoms of the columns at the first floor. Using Newton's law of motion, we equate the total force acting on the top floor to the mass times the acceleration, to obtain:

$$y_2(t) - k_2\left[x_2(t) - x_1(t)\right] = m_2\ddot{x}_2(t) \tag{2.3}$$

From Equation (2.3) and the definition of the fourth state variable (Equation 2.2), we obtain the following state equation:

$$\dot{x}_4(t) = \frac{1}{m_2}\left[y_2(t) - k_2\left(x_2(t) - x_1(t)\right)\right] \tag{2.4}$$

Similar considerations apply to the mass at the first floor. In this case, however, there are three forces acting on the mass: the horizontal load $y_2(t)$; the force from both second floor columns acting on the mass; and the force from both first floor columns acting on the mass. We thus obtain another state equation that represents the dynamics of the first floor mass. The four state equations, including Equation 2.4, are thus:

$$\dot{x}_1(t) = x_3(t) \tag{2.5}$$

$$\dot{x}_2(t) = x_4(t) \tag{2.6}$$

$$\dot{x}_3(t) = \frac{1}{m_1}\left[y_1(t) - k_1 x_1(t) + k_2\left(x_2(t) - x_1(t)\right)\right] \tag{2.7}$$

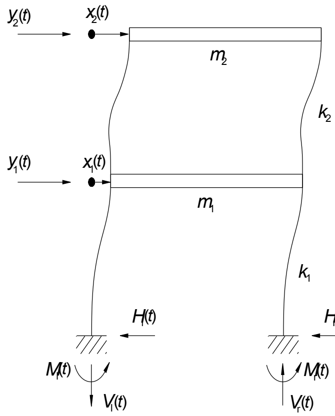
$$\dot{x}_4(t) = \frac{1}{m_2}\left[y_2(t) - k_2\left(x_2(t) - x_1(t)\right)\right] \tag{2.4}$$

These equations of state are in what is called the “normal form,” whereby the time rate of change (the first derivative with respect to time) of each state variable is expressed as a function of the current values of the state variables and the inputs. The equations of state allow us to calculate the incremental change in each state variable over the next small time interval from the current values of the state variables.

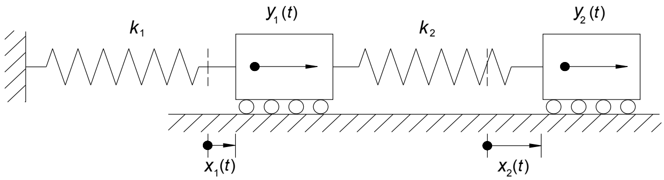
The normal form of Equations (2.4) to (2.7) is very convenient for undertaking numerical calculations. Various numerical and analytical techniques are available to solve the equations and hence obtain the values of the state variables at a succession of time instants, from the starting state and the inputs. We are not concerned here with the mathematical techniques that are used to solve the equations of state, and we shall not pursue the example further. Table 2.1 below lists the state variables, the input and output variables and the coefficients for this simplified model.

Table 2.1 System variables and coefficients for dynamic analysis

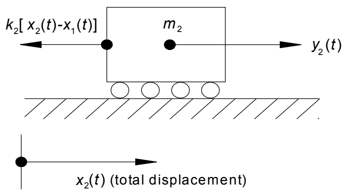
Input variables	
$y_1(t)$ (Newtons):	**horizontal force applied at first floor level**
$y_2(t)$ (Newtons):	**horizontal force applied at second floor level**
State variables	
$x_1(t)$ (mm):	**displacement at first floor level**
$x_2(t)$ (mm):	**displacement at second floor level**
$x_3(t)$ (mm/sec):	**velocity at first floor level**
$x_4(t)$ (mm/sec):	**velocity at second floor level**
Output variables	
$z_1(t)$ (mm):	**displacement at first floor level ($=x_1(t)$)**
$z_2(t)$ (mm):	**displacement at second floor level ($=x_2(t)$)**
(In this particular analysis, these two output variables happen to be the same as two of the state variables.)	
System coefficients	
m_1 (kg):	**mass of first floor**
m_2 (kg):	**mass of second floor**
k_1 (N/mm):	**stiffness of lower columns**
k_2 (N/mm):	**stiffness of upper columns**



(a) simplified dynamic structural system



(b) equivalent dynamic system



$$y_2(t) - k_2[x_2(t)-x_1(t)] = m_2\ddot{x}_2(t)$$

$$\dot{x}_4(t)=\ddot{x}_2(t)$$

(c) forces acting on top floor

Figure 2.17 State space analysis of a two-storey frame.

CHAPTER THREE

Engineering Planning and Design

Planning, design and management are the core activities used to drive an engineering project from its inception through to a successful conclusion. These are essentially problem-solving processes. In this chapter we first look at a simple, common-sense method for solving open-ended problems that provides us with a methodology for undertaking engineering planning and design. The main steps in the processes of planning and design are discussed in some detail. Some important related considerations are also mentioned, including safety and risk, the role of codes and regulations, and legal and social issues. Alternative, non-traditional approaches to engineering work are also discussed, including the agile approach and sprints and scrums. Wicked and intractable problems are discussed briefly. In an appendix, a planning example is discussed, which deals with the future water needs of a small coastal city.

3.1 TERMINOLOGY

As we have seen in previous chapters, the terms planning and design are not mutually exclusive, and there is overlap in the way both words are used. Broadly speaking, *planning* is undertaken to ensure that a proposed action will be successful. It involves deciding on the goal that is to be achieved by the action, and then identifying the steps needed to achieve the goal. These steps, when clearly formulated, constitute a *plan*. Planning is a crucial step in any engineering project. *Design*, as distinct from planning, is undertaken to produce the information needed to create a new physical system. Design work is also undertaken when an existing system is to be modified and improved. *Management* refers to the effective use of available resources to achieve a designated outcome. Management is the means by which the results of engineering planning and design are implemented and brought to a successful conclusion.

In this book, we will refer to the *design of physical objects*, and to the *planning of operations and processes*, in short: to the *design of hardware* and the *planning of software*. Accordingly, we design a bridge, but plan the process of its construction. Design work is undertaken with the purpose of determining precise details of devices, machines, and other physical objects, so that they can be manufactured or constructed. Planning work is undertaken to determine the details of the procedures and processes which form part of any engineering project.

These descriptions of planning, design and management can be sharpened if we apply the systems terminology of Chapter 2. An engineering project is thus undertaken to change the current state of some part of the physical infrastructure to a preferred new target state, either by introducing new systems and processes or by modifying existing ones. The planning for an engineering project begins with a -study of the current state of the infrastructure and proceeds to the identification of the desired future target state. It also involves working out the steps needed to change the present state into the new state. Design is the process of determining the details of any required new systems (or the details of modifications to be made to existing ones) that are needed to achieve the desired changes to the infrastructure. The purpose of management is to bring about the changes efficiently, in a designated time frame, and with an effective use of available resources.

By way of example, consider a project that aims to improve the quality and quantity of water to be delivered to a township over the next 25 years. In the initial planning work, it is first necessary to determine details of the present water supply system, and in particular the quantity and quality of the water currently used together with details of the delivery systems. It is also necessary to evaluate the expected demographic changes in the city over the planning period, including increases in physical size and population, changes in business and industry activities, so that future water requirements can be estimated. Possible new sources of water may be required, such as groundwater and domestic rainwater tanks, and possible sites for new reservoirs and dams. Combinations of the various alternatives will also be considered. Additional options in the case of inadequate water sources may include recycling of water, the use of pricing policies to curb demand, and the construction of desalination plants if there is ample salt water close by. The progressive changes to the water supply system during the planning period, to allow for the increasing demands for water, will be an extremely important planning outcome.

Design work for the project would be needed for any new components, or modifications to existing components. For example, if a new dam is to be constructed, design work is needed to determine the size, shape and other details of the dam wall, and how the wall is to be keyed into the valley sides and base at the chosen site. But before the design work can begin, it will be necessary to identify and survey possible sites and to obtain data on rainfall and runoff patterns, as well as relevant geological information. Careful planning is needed to determine the sequence of steps for preparing the site and constructing the dam wall. The resulting construction plan for the dam will include activities such as providing access roads to the site, bringing in heavy equipment, excavating soil and rock, shaping and preparing the foundations, grouting unsound regions of bed rock, and progressive construction of the main wall. Design work is also needed for the downstream components of the delivery system, such as pipes, pumping stations, holding tanks, and pressure tanks.

Comparable planning and design work would be needed for any other options that are chosen. The planning and design activities themselves need careful management, as do all the steps in the implementation phase of the project. Progress needs to be monitored regularly to ensure that the design and planning work proceeds on time and within budget.

The term *project planning* is applied to the overall planning of the project, while the planning undertaken at sub-project level is often referred to as *activity planning*. The planning required to carry out a particular step in an engineering project is thus activity planning. The planning of an activity results in a sub-plan, or component, of the project plan.

Using an hierarchical view of planning, Griffis and Farr (2000) go further and use the term *program planning* for planning at a broader level where, for example, various related engineering projects are identified and sequenced, taking account of available resources. Program planning thus takes place at the regional, state or national level. A program would include a number of projects that are to be carried out together or in sequence. A flood mitigation program in a fertile river valley might include the construction of several dams in the upper reaches, silt dredging near the river mouth and the building of levy banks in intermediate regions along the river.

Although our focus here will be on project planning and activity planning, we need to mention several other types of engineering planning. *Strategic planning* is undertaken by businesses and other organisations to produce long-term plans that are aimed at their continued successful operation and survival. Another form of long-term planning, of interest particularly in situations where the future is uncertain, is *scenario planning* (Meier et al. 2016). As an example of scenario planning, NASA and ESA (the European Space Agency) have developed a joint experimental program to test whether the impact of a relatively small missile could nudge an asteroid out of a collision path with the earth (Cosmos *Community Input,* April 2017). According to *Cosmos*, a different approach is being studied by the Makeyev Rocket Design Bureau in Russia, which aims to design and construct a missile that could destroy asteroids if they come dangerously close to earth.

The design activities in an engineering project can be broken down into tasks and sub-tasks. This breakdown is needed to deal with the complexity inherent in a large infrastructure item such as a dam, a road, a bridge, a building or an airport. In the design of a unit that is to be manufactured in large numbers, such as a washing machine, a mobile telephone, or an automobile, the design is undertaken using the hierarchical approach. It is also used in the planning of processes, such as the manufacture and delivery of ready mixed concrete to building sites in a city.

Planning and design are complex activities; so complex, usually, that they can only be carried out iteratively, with considerable trial and error. It will be appreciated that the activities of planning, design and management for a large project will be closely inter-related and will probably overlap. Basically, they are all problem-solving activities.

3.2 DEALING WITH ILL-DEFINED AND OPEN-ENDED PROBLEMS

In Chapter 1 we saw how engineering problems can be complex, open-ended and ill-defined. The systems approach was introduced in Chapter 2 to deal with complexity in engineering systems and in engineering problems, but it does not provide a means to solve open-ended problems. We now consider how best to deal with open-ended and ill-defined problems.

An ill-defined problem is one that is vaguely or ambiguously formulated. There are good reasons why engineering problems are initially poorly defined. The need for an engineering project can gradually become evident over time as inadequacies in the infrastructure are experienced. Community dissatisfaction leads to demands for improvements. In such situations the needs of the community may be expressed forcefully, but not necessarily precisely and unambiguously. This is how some large engineering projects begin. The first step is to identify and clearly formulate the real problem.

Even when an engineering problem has been clearly formulated and stated, it will be open-ended, in the sense that there will not be a single "correct" solution. Indeed, a characteristic of engineering problems is that there are alternative, potentially acceptable solutions. The problem is not to find "the" solution but rather to find the "best" solution. For example, road traffic congestion in a part of a city might be dealt with by increasing the capacity of the existing roads. An alternative could be to introduce an efficient and cheap public transport system. Some social engineering might also be considered, by modifying the starting and finishing times for employees of industries and businesses in the problem region.

At the commencement of an engineering project, the initial state will not be known precisely and the target state is still undefined. In some situations it may be very difficult to identify the required target state and, in a worst case scenario, it will not be possible to know whether the correct engineering decisions have been made until the project has been completed. For example, the planning and design for a new and improved model of an automobile requires a great deal of expensive engineering work, and then manufacture, but the introduction of the new model is, in the end, a risky commercial venture. Its success can only be tested in the market place after the engineering work has been concluded. The open-ended nature of engineering work raises a number of important questions. In particular:

- how do we solve an open-ended problem if there is not going to be a unique, correct solution?
- how can we know if we have found a good solution to an open-ended problem?
- how do we go about identifying alternative solutions to an open-ended problem?

To find answers to such questions we need to look at problem-solving strategies.

3.3 METHODOLOGY FOR SOLVING OPEN-ENDED PROBLEMS

People use a range of problem-solving strategies in their everyday life. One popular procedure is to choose a method (or even a solution) that was successfully used in the past. This approach might work well in relatively simple situations, provided the problems, past and present, are identical, or very similar in nature. Difficulties will arise if there are substantial differences between the present and past problems.

In engineering work, it can be useful to draw on past practice for ideas for solving current problems. However, it is dangerous to rely exclusively on this approach. Adaptation is always needed, because differences inevitably exist between past and present engineering problems, and great care must be exercised to ensure that these differences are allowed for. Furthermore, past practice is not relevant when a problem has new features. Often there is no relevant, reliable past practice to rely on and engineers must then create innovative, new approaches.

Problem-solving strategy

When faced with a problem that is both ill-defined and open-ended, common sense suggests that the first step should be to clarify the problem and re-state it in clear and unambiguous terms, in so far as this is possible. If a range of alternative approaches can be found, after the problem has been clearly formulated, then the initial instinct may well be to choose the most "obvious" one and develop this into a solution. The obvious or instinctively most acceptable solution is likely to be one that has been used previously and that we are therefore familiar with.

On reflection, however, it becomes clear that intuition alone will *not* lead necessarily to the best, or even to a good, solution. Rather than concentrating initially on a particular solution, it is better to take just the opposite approach and look for as many different solutions as possible, that are feasible and promising. To be able to do this, we need to formulate the problem in general (not over-specific) terms so that unusual but promising solutions are not excluded.

Our search for unusual but promising ideas will not be successful if we use convergent thinking and rely on an analytic approach. We need to use divergent, lateral thinking. Creativity is extremely important because we are seeking unusual, non-routine solutions. In Chapter 4 we shall look at the question of creativity and how creative approaches to problem solving can be developed.

If we are successful in creating a range of different and promising solutions to our problem, our next task is to evaluate them, compare and rank them and hence identify the best approach. These simple thoughts lead to a problem-solving strategy which can be applied to poorly formulated, open-ended problems. The steps are listed in Table 3.1.

Table 3.1 Strategy for solving complex, open-ended problems

Step	Action
1	Formulate the problem clearly but in general terms
2	Develop a wide range of promising approaches for solving the problem
3	Evaluate and compare these approaches and hence identify the best one
4	Work out the details of the solution, based on the best approach
5	Implement the solution

Methodology for complex engineering problems

The strategy of Table 3.1 will be appropriate for relatively simple, open-ended problems where it is easy to evaluate and compare the alternative options. This may be the case in simple activity planning tasks. If, however, the problem has many, relatively complex, alternative solutions (and this is typically the case in engineering work) then an enormous amount of effort would have to be expended in the third step. A full set of designs and plans would have to be created for each approach. This would be extremely time consuming, costly and inefficient, because it would mean solving the detailed problem many times, once for each option.

For example, if we consider a harbour city and the problem of choosing the most suitable water crossing for the road traffic, there may be a dozen main options, including bridges of various types at different sites and under-harbour tunnels with alternative routes. Each bridge and tunnel option will be unique. If the problem solving strategy of Table 3.1 were applied unmodified, it would be necessary to carry out detailed designs for each option, in order to make proper and fair comparisons. The design and planning costs in engineering work are always substantial and in this case they would be many times the cost of a single design, and prohibitive. The bill for the planning and design work alone would take up a large portion of the budget.

Planning and design for the Great Pyramid of Giza

Studies of the great pyramid of Khufu (or Cheops) have shown how complex and ingenious the planning, design and construction for this mammoth engineering undertaking must have been. The pyramid, located on the outskirts of present day Cairo in Egypt, was built more than four thousand years ago. It has a 230 m square base and covers an area of 5.3 hectares or 13 acres. Over 140 m high, it is the greatest monumental construction of antiquity. Apart from its special chambers, the pyramid originally consisted of three main components: a massive inner core of stone blocks, a thin surface cladding of white limestone, and an intermediate fill layer between the cladding and core. The surface cladding must have given the construction a spectacular appearance. It has been vandalised over the millennia. The main inner core contains over 2 million large stone blocks of limestone and granite, mostly weighing between 2 and 6 tonnes. These were sourced from various quarries, then sized and shaped before being transported by Nile barges to the site, where they were moved on rollers and raised to the working level (up to 140 m high) before being placed precisely in position. One theory suggests that a temporary inclined helical ramp was built around the partially completed pyramid and used to move the blocks up to the working level. It is estimated that the project was completed in a little more than twenty years by around 20,000 men working in small teams.

Pyramid *design* was complex and ingenious. In earlier pyramids the stone blocks were given a slight inward inclination to improve stability and reduce settlement. The great pyramid was constructed on bedrock which had been cleared, levelled and shaped to give stability and prevent settlement. The blocks were placed in horizontal layers, generally with larger blocks at lower levels and with smaller blocks in the higher layers. Corbelled ceilings with granite beams were used for special rooms, and stress-relieving chambers were introduced above the burial chamber.

The *planning* for the construction must have been of the highest order, comparable with that needed for a modern monumental engineering project. On average it was necessary to transport, deliver, raise and place about 300 blocks per day. With say a twenty-man team to look after each block, an average on-site workforce of above 6000 would have been needed. However, it seems that the main building work took place during the yearly flooding of the Nile, when agricultural work could not be undertaken. The building site must have then been a hive of activity. Even looking after the needs of the workforce would have been a highly complex exercise in planning and logistics.

The planning, design and management skills evident in the construction of the Khufu pyramid had been acquired by experience and trial and error over generations of pyramid building. The name of Imhotep, a vizier, medical doctor and scribe, is associated with pyramid design and construction, in particular the step pyramid of Saqqara. He is perhaps the first engineer to be known by name. The vizier responsible for the great pyramid is thought to have been Hemiunu.

Sources: DTV Atlas (1974); Craig B. Smith (2004)

A modification to the strategy in Table 3.1 is needed. The costs in time and effort in the planning and design phases can be reduced enormously if, instead of fully evaluating every option, we use an incremental approach, and start with simple and rough comparisons of the options. On the basis of quite limited information it will usually be possible to identify and eliminate the less competitive options. This first step is not costly because we are using simple, approximate, order-of-magnitude calculations. A second, rather more detailed, evaluation of the remaining options then allows a further cull to be made. Proceeding step-by-step in this way, with increasingly detailed calculations, we obtain a short list of the most promising approaches. For this strategy to work, we must have a range of models and analysis methods that vary from simple and rough, up to accurate (and presumably complex). We have already discussed modelling and analysis in Chapter 2, where we saw that a range of approaches of varying complexity, and therefore accuracy, can be developed in most engineering fields.

Proceeding step-by-step in this way, we eventually identify a single best option. If not, we will at least have a short list of the most promising options and we can then proceed to a final round of comparisons in which we use the most accurate analyses and calculations.

With this methodology we substantially reduce the total amount of effort. Furthermore, the more detailed calculations are made for the more promising options. The steps in this modified methodology are listed in Table 3.2.

Table 3.2 Methodology for solving complex, open-ended problems

Step	Action
1	Formulate the problem clearly and in general terms
2	Develop a wide range of promising approaches
3	Choose criteria for ranking the alternative approaches
4	Cull the least promising approaches using simplified evaluations
5	Cull progressively, with more detailed evaluations, until a short list remains
6	Choose the best approach from the short list, using detailed evaluations
7	Develop the best approach into a detailed solution
8	Implement the solution

Iteration

It might appear from the discussion so far that we only have to follow the linear sequence of steps in Table 3.2. This is not so. As the design and planning work proceeds, it will often become clear that a modification, improvement or correction is needed in one or more of the steps already completed. In Step 4, for example, when a new and unusual, but promising, approach is being evaluated, it might be found that the original problem statement in Step 1 is unnecessarily restrictive. Rather than staying rigorously with the original problem statement it is best to go back and restate the problem in a more encompassing way, as our aim is to obtain the best solution. Likewise, when alternatives are being compared and ranked in Steps 5 and 6, it might become clear that an improved problem statement will allow improvements to be made to some of the approaches.

Even in the final phase, when the short-listed alternatives are being ranked and ordered, it might be found that some approaches have been unnecessarily penalised or even eliminated because of arbitrary, unnecessarily restrictive criteria. A reformulation of Step 3 would then be justified if it can lead to a better solution to the problem. It might even be possible to modify some approaches in Step 6 to achieve more favourable evaluations and hence a better solution to the problem. Iteration is an essential ingredient in this methodology.

In some situations it may be advantageous to undertake several steps simultaneously. If it is difficult to choose good evaluation criteria in Step 3, before the alternative approaches have been examined, it might be preferable to undertake Steps 3 and 4 simultaneously. A possible complication then to be avoided is unintentional bias in the choice of criteria that favours an intuitively appealing and favoured approach.

We will now see how this methodology is applied to the processes of planning, and then design.

3.4 ENGINEERING PLANNING

A methodology for undertaking engineering planning is shown in Figure 3.1. It follows closely the problem-solving methodology listed in Table 3.2 but typical feedback loops have been added to emphasise the need for iteration.

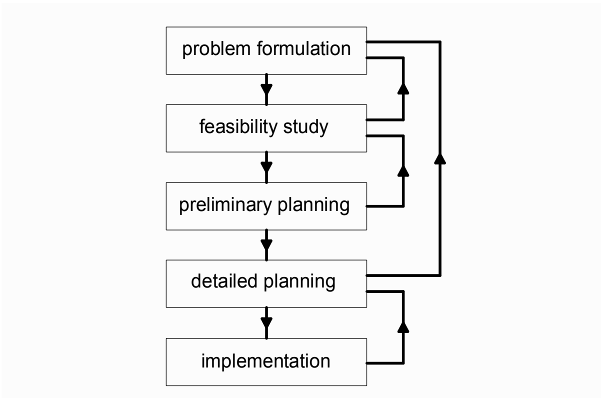


Figure 3.1 The process of engineering planning

There has also been a regrouping of the steps from Table 3.2 into five main phases, in order to conform with common engineering usage. Some new terminology has been introduced. In particular, the step of choosing the evaluation criteria (Step 3 in Table 3.2) has been combined with Step 1 to become the *problem formulation* phase. The *feasibility study* incorporates Steps 2 and 4, while *preliminary planning* consists of Steps 5 and 6. Step 7 has been renamed *detailed planning*.

The overall aim of the initial problem formulation phase is to clarify and, if possible, quantify the problem. This is not as simple as it may at first appear because engineering problems are usually many faceted. For example, *constraints* always apply to an engineering project, and limit in various ways the approaches that can be taken. Engineering projects are always constrained by limits on overall cost, on available time, and on the quantity of materials and human resources that are available. Account needs to be taken of these constraints in the initial problem formulation phase.

The *feasibility study* is the second phase of the process in Figure 3.1. This term emphasises that there might not be a feasible solution. The desired outcome from this step is of course a short list of acceptable approaches, which will demonstrate feasibility. As already mentioned, the intention at this stage is not to focus on a specific approach, but to find as wide a range of different and contrasting feasible alternatives as possible.

The preliminary comparison of alternatives commences during the feasibility study, after the range of options has been identified. This allows us to test out the evaluation criteria and to modify them as necessary. The feasibility study continues until we have a short list of good, feasible options. If no feasible approach can be found, then we must abort, postpone or modify the project. Reasons for not finding a feasible solution might be related to difficulties in satisfying the constraints that are imposed on the solution. For example, costs might be unexpectedly and unacceptably high for all alternatives. Sometimes the technology will not be sufficiently advanced to allow an economical and safe solution to be formulated. If no feasible solution is found, an option is to repeat the feasibility study, but with relaxed constraints and less optimistic goals.

In the third phase of the process, *preliminary planning*, further rounds of comparisons and evaluations are carried out, in order to identify the best option. The comparisons now require more detailed analysis and evaluation than previously. If one of the alternatives stands out clearly from the others, then the preliminary planning phase can be quickly and easily completed. On the other hand, if the alternatives are closely matched, this phase can become protracted because of the need for increasingly detailed comparisons.

The purpose of the *detailed planning* phase is to take the best option, when it has been identified, and develop it to work out all of the details needed to undertake the implementation. If much analysis and evaluation work has already gone into identifying the best option, then there will be correspondingly less detailed planning work needed. The end result of this phase is a detailed plan for implementation. However, this might still not be the final plan. Problems can arise during implementation that make it necessary to revisit some of the earlier steps. Indeed, the plan will not be finalised until the project is completed. There is an often quoted statement: it is not the plan that is important, it is the planning.

The last phase, *implementation*, may require the construction or manufacture of new components of infrastructure or the implementation of new processes. At first sight, implementation may seem to be a separate part of the project, to be undertaken when the planning and design work has been completed. This is not so. Whatever form the implementation takes, it is almost always necessary to return to and revise some of the earlier planning work.

In later sections of this chapter we will discuss in some detail how to undertake each phase of the planning process. Before doing this, however, we will discuss the phases of the design process.

3.5 THE DESIGN PROCESS

The design process shown below in Figure 3.2 parallels that in Figure 3.1 for planning, although there are some differences in terminology. In particular, the term *concept design* has been used to describe the second phase of the process, when alternative approaches are sought. The emphasis here is on finding alternative concepts, or design options, rather than on investigating feasibility. Despite this difference in wording, the aim of the second phase is the same in both design and planning: to produce a short list of the best feasible options for solving the problem.

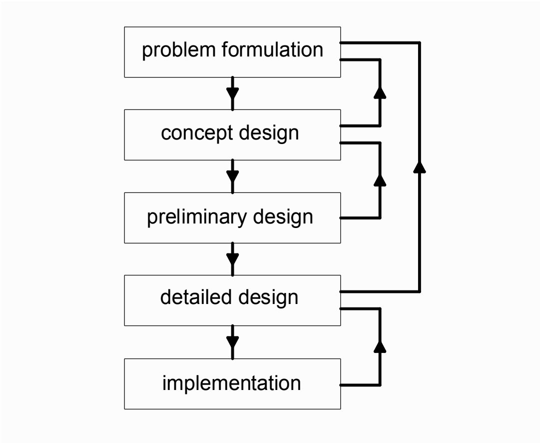


Figure 3.2 The design process

In the structural design of a bridge, the concept design phase is undertaken with the purpose of finding suitable alternative structural forms that provide safe "load paths" for the design loads. The load paths carry the applied loads, such as self-weight, vehicular loads and wind loads, through the different structural components and into the supporting foundations. For a large-span bridge, the alternative structural forms might include an arch, a suspension bridge, a cable-stayed bridge and a multiple span system of beams with intermediate piers.

Initially, the aim is to produce a wide range of alternative, promising concepts, but then to reduce these down to a short list, using the evaluation criteria developed in the problem formulation phase, using simplified calculations and approximate models and analyses.

The most general, far-reaching and financially most important design decisions are made in the initial stages of the concept design, when new, original and unusual approaches are sought. As the design proceeds towards the detailed phase, the design decisions become more and more specific, and they have less impact on overall costs.

In Figure 3.2, the terms *preliminary design* and *detailed design* correspond to their counterparts in Figure 3.1. The purposes of these phases are similar in both planning and design and, as we shall see, they are carried out in the same way. The purpose of the preliminary design phase is to study the short-listed concepts in sufficient detail to allow the best concept to be identified. If it is difficult to choose from several very good concepts, then successive rounds of calculations and comparisons of increasing complexity are needed. These studies may become, in effect, detailed designs. It is not unusual for the preliminary design phase to blend into the detailed design phase, with detailed designs being carried out for several of the very best options.

In the design of a dam for a water supply system that was planned for the Hastings district in New South Wales, Australia, 24 sites were identified in an initial study. During preliminary design this was narrowed to two, before the best was chosen for the final design (Thompson, 2005).

In the detailed design phase, the aim is to take the best approach, now identified, and generate all the information needed to allow full implementation. If one concept has been clearly identified at an early stage as the best, then a fair amount of additional, detailed work will be required in this phase. On the other hand, the effort needed in the detailed design is correspondingly reduced when the chosen concept has already been investigated in some detail.

As previously noted, reconsideration of previous phases of the design may become necessary when implementation is underway. Poor engineering design will almost always be the result if inadequate attention is paid to how the design is to be constructed or manufactured or otherwise implemented.

Although Figure 3.2 adequately represents the engineering design process, some designers prefer to describe the process slightly differently. The term *preliminary design* is sometimes preferred to concept design. It is also not unusual to treat problem formulation as a part of concept design. The design process could then be described as a three-phase sequence consisting of concept (or preliminary) design, detailed design, and implementation.

Irrespective of the way the design process is described, the activities of problem formulation, concept design and preliminary design (in the sense explained above) are the basic steps. We have chosen here to describe design in terms of the five phases of Figure 3.2. There are two reasons for this: firstly, the importance of problem formulation in design is highlighted; secondly, this formulation emphasises the close similarities between the processes of planning and design.

What is not shown in Figure 3.2 is how the need for design work is often first recognised during the project planning process. This is when the need for a new physical system might first be identified. Initially the need is not formulated very clearly or precisely, so that problem formulation remains an important step to be undertaken at the commencement of the design process. The following statement by Richard Seymour, and quoted by Liston (2003), emphasises the crucial role of problem formulation in engineering design:

> *Something which is often forgotten or misunderstood is that the vast majority of the work involved in design is finding out what the problem really is– and it's rarely what you think it is. More often than not the client is asking the wrong question. Once you've found out what the problem really is, then things virtually design themselves because you've so comprehensively understood the problem that the solution is self-evident.*

3.6 COMMUNITY INPUT

Large engineering projects are designed to improve the physical infrastructure, and the community as a whole are beneficiaries. Nevertheless, a minority of people can be seriously disadvantaged by even the most successful projects. For example, when upgrades are made to access roads into a city, land acquisition can disrupt the lives of people who live and work in the affected areas. In such circumstances, financial compensation can become an important part of the process.

People have the right to know about proposed new developments that will impact on their lives, and they also have the right to express opinions and concerns about such developments, and indeed to take part in the decision making processes. In many countries, freedom of information laws ensure that information is available to the community on proposed new engineering projects. When large-scale engineering work is to be undertaken, the sectors of the community that are likely to be affected need to be advised and brought into the process, not only through written information but also via public meetings and consultations. The views of individuals and groups in the community need to be known during the initial planning phases of the project and also as work progresses.

Differing views are normal in any group of people. Alternative and opposing views can arise in regard to an engineering project, and will usually be accomodated in the community consultation process. A compromise view is often reached which is acceptable to, if not welcomed by, all participants. Unfortunately, this is not always the case. Occasionally, opposing or competing community views of a proposed engineering project can become entrenched, that seem not to be resolvable. The term *intractable* has been applied to such situations, and the associated problems have been referred to as *wicked*. These are rare, but do occur, and are discussed in some detail in Section 3.14 below.

3.7 PROBLEM FORMULATION PHASE IN PLANNING AND DESIGN

Table 3.3 below contains a checklist of actions that can assist in the problem-formulation phase of engineering planning and design. The terminology used comes from Chapters 1 and 2. Some of the actions are information-gathering activities while others are aimed at formulating and, to the extent possible at this stage, quantifying the problem. Consultation with the community and its representatives will often be required. Although intended specifically for planning and design, these actions can be useful in problem solving generally. We now discuss them in turn.

Identifying the problem as part of a wider problem

As we have already seen, an engineering project can begin as a response to perceived needs in the community, and the objectives may be initially poorly defined. The problem as stated may be too vague; alternatively, the problem may be stated in over-precise terms which imply an “obvious” solution. In the latter case, the implied solution will rarely be the only possible one, and may not be the best one. It is therefore important to begin the problem formulation phase by trying to identify the problem as part of a wider or larger problem.

Table 3.3 Checklist of actions for the problem formulation phase

Action
Identify the problem as part of a larger or wider problem
Identify the relevant engineering system as part of a wider or larger system
Identify the components of any relevant system
Find interest boundaries for the problem
Determine the real underlying needs that are to be addressed
Gather relevant background information
Search for possible side-effects
Identify constraints
Specify objectives and identify possible conflicts among objectives
Specify design criteria, performance requirements and operating condition, as appropriate for each new system
Devise measures of effectiveness

By way of example, we return to the problem of providing traffic access between two parts of a city separated by a waterway. Suppose that the cross traffic is already catered for by several bridges which cannot handle the increased traffic volumes. The "obvious" problem might be seen as providing an additional bridge to cater for present and future traffic. Is this a statement of the problem, or is it statement of a solution to a problem? A broader problem statement would be: how to improve the cross-harbour traffic flow. An advantage of this statement is that it widens the range of possible solutions. It is now possible to consider a tunnel, a range of small peripheral bridges, various forms of water transport, or even air traffic as alternatives. This example was used by Noel Svensson in his book on engineering design in 1974.

In fact, the city of Sydney has been faced with a traffic problem of this nature for many years. In the 1950s there was a single bridge to carry traffic across the harbour between the northern and southern suburbs. A second bridge was constructed on the periphery of the harbour at Gladesville, but traffic volumes continued to grow and traffic congestion became worse. There were community calls for another large harbour bridge. The next step was in fact to construct a tunnel to address the cross-harbour traffic problems in the medium term. The tunnel, integrated into the arterial roads of the city, was initially successful. However, as the city continues to grow so too do the problems of traffic congestion. New solutions are needed.

It is useful to further widen our example. Is the real problem simply one of regularly increasing the traffic capacity on all the main routes in response to growing volumes? A broader problem statement might be to reduce or eliminate the current and future traffic congestion. This statement allows more varied approaches to be considered. It allows, for example, a reduction in peak traffic by staggering starting and finishing times for various industries, and the use of flexitime in offices. It also encompasses initiatives such as imposing surcharges on traffic that uses critical facilities at peak times, and encouraging the increased use of public transport. A form of congestion tax has been introduced in cities such as Singapore and London to limit the number of vehicles entering the inner city area.

Broadening the problem statement does not disallow the initial “obvious” solution. On the contrary, it opens up the problem to a range of alternative, competing solutions and so tests the adequacy of the “obvious” solution.

Identifying each engineering system as a component of a wider system

To assist in identifying the problem as part of a wider problem, it can be useful to identify the relevant engineering system as a component of a larger or wider system. In the above example the broader problem statements have focused on the transport system, of which a bridge is just one possible component.

It can be useful to take the problem-broadening technique further. In the above example, is the problem purely one of achieving an efficient city transport system? The costs involved in providing roads, tunnels and bridges to improve the traffic network of a large and sprawling city run into many billions of dollars. When such expensive traffic options are considered, other radically different possibilities deserve consideration. It might be possible to achieve a more workable city by reducing the need for cross-harbour traffic. Here we are extending our discussion beyond traffic engineering and into the realm of town planning. We can go further: is it the best use of resources to improve road transport in an already congested city? Even if the road transport system were to be improved, the result is likely to be more city growth, followed by further congestion and yet more calls for improvements to the again-inadequate transport system. An alternative and better use of the resources might be to introduce decentralisation schemes and encourage the relocation of industry and population away from the city, into regional growth areas. We have now moved on from town planning questions to a consideration of regional planning and various associated political issues.

Identifying interest boundaries for the problem

How far should the process of problem broadening be taken? Clearly we could extend the bridge argument further, beyond regional planning and into areas of national or even international planning. At some stage the problem broadening argument breaks down. How do we recognise this stage? The questions cease to be relevant when they are so broad that new issues will barely influence the problem statement. This occurs when the expanded system is so large that the original problem (in this case, the need for a bridge, or some alternative) ceases to be relevant. When this happens, we have clearly gone too far.

The limit, where the problem has been widened to the extent that the original problem is of marginal relevance, is referred to as the *interest boundary* for the problem. The relevant solution options will be found within this boundary. In Figure 3.3 the interest boundary is shown for the bridge example previously discussed. In this case the very large costs indicate that the interest boundary should certainly include the city system and perhaps extend to the surrounding region or even state.

The construction of a large bridge in a small country can become an international issue. An example was the K-B Bridge in Palau, a small island nation in the Western Pacific, somewhat less than a thousand kilometres west of the Philippines. The bridge was constructed in the 1970s with aid money from the

United States when Palau was its protectorate. At the time, the bridge was the largest prestressed concrete box-girder cantilever arch construction in the world, linking the two main islands of Korror and Babeldaob. It served a crucial purpose for the whole population of Palau because essential services, including the airport and power generation equipment, were located on one island while most of the population lived on the other. The bridge collapsed suddenly and apparently without warning in July 1996, shortly after it had undergone extensive refurbishment. Life in Palau was severely disrupted, even though a ferry connection was established between the islands. The construction of a new bridge was clearly beyond the resources of Palau and had to wait until it could be undertaken with international aid. The interest boundary in this case extended well beyond national boundaries.

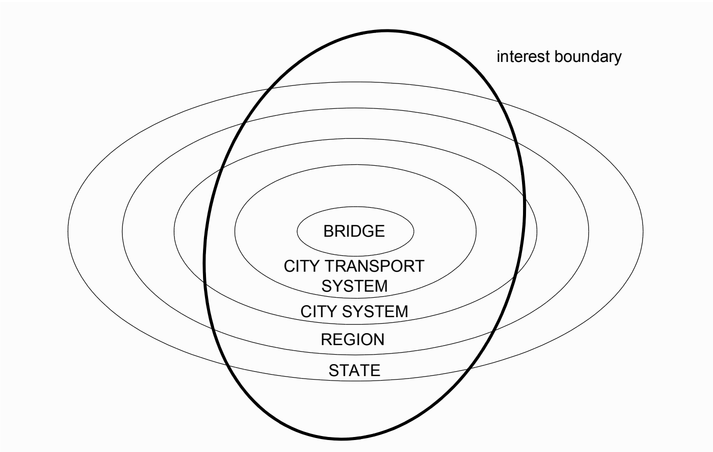


Figure 3.3 Problem-widening: interest boundary for a bridge.

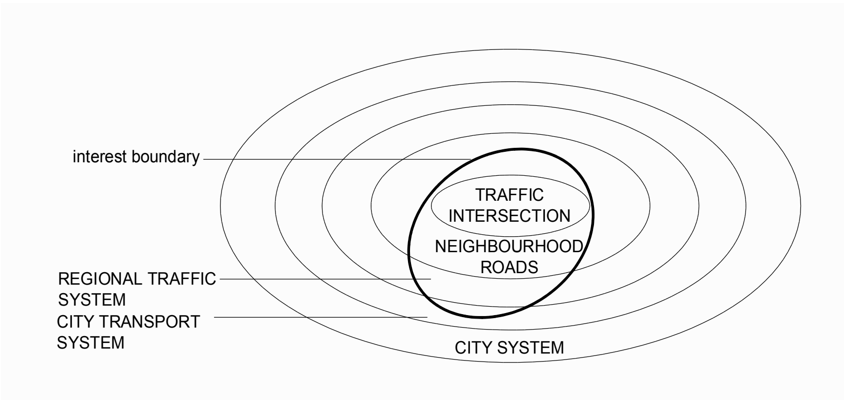


Figure 3.4 Problem-widening: interest boundary for a traffic intersection.

As another example of an interest boundary, we noe choose a relatively small one, but again in the field of traffic engineering. We consider the redesign of a hazardous traffic intersection in a suburban area.

The traffic intersection may initially be seen as a part of the local traffic system, consisting of the intersection and the immediately adjoining streets. This can in turn be regarded as part of the traffic system in one region of the city. The interest boundary is indicated in Figure 3.4. In this case the costs of the proposed project are a small proportion of the annual road transport budget. Conditions at the traffic intersection are of only marginal relevance in relation to the efficient flow of traffic through other regions of the city.

In summary, the problem-broadening process should be halted when a further widening of the boundary has only a marginal impact on the original problem statement. The interest boundary can be identified in this way.

Determining the real underlying needs as distinct from the stated needs

The real needs that initiate an engineering project are not always those that are initially perceived. The process of problem widening and the identification of the engineering system as a component in a wider system usually lead to a better understanding of the real needs. In the bridge example, the need is not for a bridge *per se*, but for a reduction in traffic congestion, or, possibly, a more efficiently functioning city. When the needs are properly identified the problem can be formulated in a more general and much more useful way.

In situations where the size and scope and influence of a project are limited, it may be much easier to identify the relevant underlying needs. As an example, consider the construction of a swimming pool as part of a house design for a private client. Before the design work is undertaken the real needs of the client have to be clarified. The pool may be needed for swimming training, for diving, for exercise, or for social entertaining and relaxation. It might be required mainly for prestige reasons, or for a combination of reasons. In each case a different design could be appropriate. The problem statement should be formulated through sensitive discussions with the client.

The attempt to identify underlying needs sometimes moves attention away from purely technical problems towards social issues and political problems, especially when the proposed work is large in magnitude and costly. This is almost inevitable if large sums of public money are to be spent. As the problem-widening and system-widening processes are undertaken, the questions acquire an increasingly political and social flavour and require community input.

Gathering relevant background information

When work commences on a new project there is usually a lack of relevant background information. This has to be gathered as early as possible, and before the problem formulation phase is completed. The required information may be scientific and technological, but it may also be non-technical and sociological, legal or political in nature. The data-gathering exercise may require consultation with the community, the use of libraries, the internet, textbooks, and databases. Sometimes it will be necessary to gather field data and conduct laboratory tests, as for example when the properties of a foundation material have to be determined.

The type of information needed in engineering work is unlimited. When international one-day cricket matches were initially planned in Australia, a lack of suitable turf wickets was an important problem and the design of replaceable, transportable turf wickets was investigated. In this case the necessary background information included the way the game of cricket is played as well as the laws that govern the game. Detailed horticultural information was needed on the appropriate types of grass for preparing turf wickets, how to grow the grass, and the depth of soil needed, not only to allow the grass to grow but also to provide the right "bounce". Other information was needed on how to achieve the best type of drainage and compaction, and the required wearing properties of the turf. The structural engineering questions, relating to creating a rigid, transportable structural tray to support the pitch, seem relatively uncomplicated. In unusual projects, engineers will rarely have adequate knowledge or training in the relevant specialist fields. It is necessary to gather information rapidly and may require working in association with consultants and other professionals with specialised knowledge.

As a further example, consider the relocation of arterial roads in a suburban area. There is an obvious need for technical information concerning the soil properties along the proposed routes as well as the present and expected future traffic volumes. To obtain this information programs of soil testing and traffic counting might be undertaken. Other relevant background information of a quite different nature relates to the social structure of the community through which the roads are to pass, for example the location of the roads in relation to natural community boundaries, including the catchment areas for local schools. The political implications of such a project, measured in terms of votes gained and lost, may also turn out to be important background information, influencing any political decisions on whether or not the project proceeds.

Searching for side effects

Each engineering project is designed to bring about some change in the world in which we live. Although the intention is to improve the infrastructure and thereby satisfy some community and individual needs, it is inevitable that there will be side effects. These may be obvious and desirable, but also hidden and undesirable. They may only become evident after a considerable period of time, when the project has been completed and is in operation. In the case of a large-scale project there will usually be a wide range of side effects, some beneficial and some detrimental.

The identification of side effects should begin as early as possible during problem formulation but should continue into the evaluation stages because some side effects will depend on the chosen solution. It is clearly important that the relevant side effects be identified and allowed for when the costs and effectiveness of each alternative approach is evaluated.

Some disturbance of the environment may be inevitable when a large infrastructure project is undertaken, and environmental side effects can become a focus of community attention. If a road or freeway is constructed, either in an urban or rural development, some modification of the surroundings is involved, such as the demolition of existing older buildings. It may involve cutting into a hillside or filling in part of a valley floor in the countryside. Even when a large-span bridge is constructed across a valley to lessen the physical effects on the environment of the roadway, there is a visual impact. Of course the impact may be pleasing to many, but a serious problem to others.

A side effect of improved car security

In a report in the Australian Newspaper of 5th June, 2006, a rise in the number of "car jackings" was attributed by police to improved "immobilisation technology". The new technology had made the theft of unattended late model vehicles almost impossible. According to the newspaper report, thieves had therefore taken to "holding up drivers across Sydney's wealthy suburbs".

In 2011, work commenced on a high road bridge to cross the valley of the Mosel river in Germany. The main purpose of the bridge was to improve links between the city of Frankfurt and Dutch and Belgian harbours. However, as reported by Asimov (2010), the project became controversial, and serious opposition developed, when adverse side effects on the wine industry and tourism in the region were identified.

Technical side effects may be advantageous or disadvantageous for a project. The provision of lift wells and associated walls in a multi-storey building design improve the resistance of the building to lateral loads. Non-technical side effects can become significant, such as increased or decreased job opportunities, which may be temporary or long lasting. The influence on the tourist industry of the construction of a monumental, iconic building, or the development of water sports and recreation facilities following the construction of a dam, are examples of non-technical side effects which may have an effect on the outcome of a project. They need to be identified and spelt out in the problem formulation phase.

Identifying constraints

It is important in the problem formulation phase to identify the constraints that apply to the problem. Constraints restrict the possible solutions to a problem, and arise in various ways. They may be technical, legal, economic, social, environmental or political in nature. Monetary cost is a constraint that applies in one way or another to every engineering project. Constraints arise if certain side effects are unacceptable or undesirable, such as excessive atmospheric pollution. Technological limitations create other constraints.

Legal constraints apply, for example, to building construction. They limit the maximum footprint size of a building on a city building site and also the maximum building height. Other examples of legal constraints are the emission control requirements for automobile engines and the minimum requirements for fire protection and toilet facilities in buildings that house people.

Legally enforceable industry standards are often the means by which constraints arise in engineering design. Limits for noise control and thermal insulation thus apply in the design and construction of apartment buildings. Design constraints may be introduced for the different components that make up a system so that they fit together and work in harmony with each other. Constraints can sometimes be quantified as physical limits or as minimum performance requirements which must be achieved.

Identifying Constraints: Power supply for Waterfall Gully

As already noted in the main text, much of South Australia's power is carried on the distinctive steel and concrete Stobie poles. However, there are situations where these cannot be used. In the late 1970s the increasing demands for electricity in Adelaide's eastern suburbs led the state electricity body, the Electricity Trust of South Australia (ETSA), to investigate the provision of additional capacity to suburbs adjoining the Adelaide Hills. This required 66kV power lines to traverse difficult terrain over a number of hills.

Various constraints were identified in relation to the location and nature of the proposed work, such as:

Cost and visual appeal: although underground power lines would have been preferable from a visual point of view, the cost of this solution had recently increased significantly and essentially ruled out underground lines as an option.
Legal: planning approval was required from the Department of Environment and Planning and this led to further constraints on what could be installed.
Environmental: the Department of Environment and Planning imposed constraints in terms of visual amenity and the physical environment, including limiting the scope of roads that might be used to access the area.
Social: although the residents of nearby councils were the beneficiaries of the new power supply, they exerted significant social and political pressure on the planners in an effort to maintain their views and the natural characteristics of the area.
Technical: without the ability to drive roads and tracks through the area a solution had to be developed using available technology. The solution was built around the use of a helicopter. This had technical constraints associated with it, because there was a limit to the load that could be lifted, in this case a spare carrying capacity of less than 140kg.

The constraints that were imposed left few viable options. The one that was chosen was to build and partially assemble three steel structures. The components were to be brought to the site by helicopter and bolted together while secured by guy ropes.

Difficulties encountered in assembling the pylons, coupled with other issues, led to the deaths of four workmen who were on one of the structures when it suddenly fell to the ground during an operation designed at securing it in alignment. Three died instantly and the fourth was pronounced dead on arrival at the Royal Adelaide Hospital.

Source: Grabosky (1989)

The unavailability of certain resources can lead to other constraints. For example, telephone poles were commonly made of wooden tree stems in most parts of Australia in the mid-20th century. However, the lack of trees in country South Australia severely constrained the use of timber and led to the early and widespread use of a steel-concrete composite pole, called a Stobie pole after its designer. Stobie poles have been a distinguishing feature of the South Australian landscape for many decades.

Another important constraint is time. Completion deadlines are crucial constraints in the construction of sporting complexes and facilities to be used every four years for the Olympic Games. The timelines for construction are determined by the set dates of the games and have to be met, in one way or another.

Physical constraints apply to some engineering problems. In the early stages of the design of the Gateway Bridge in Brisbane, Australia, a minimum clear height above water level at mid span was required in order to allow shipping to pass under the bridge. On the other hand, a maximum overall height limit was imposed to give air clearance for flight paths to the nearby airport. The initial constraints on minimum and maximum heights clashed for this bridge and no solution was possible until the constraints were examined more closely and relaxed.

In the case of large engineering projects with overt government support, it is not unusual to find that political considerations lead to explicit or implicit constraints on the engineering work. Local sourcing of materials and the creation of employment are frequently negotiated before any engineering work is undertaken and become constraints on the engineering work. Such constraints have to be identified early in the problem clarification phase of the project.

Defining objectives and identifying conflicts among objectives

It is important to have a clear and unambiguous statement of the goals and objectives of any engineering project. This is, in effect, a statement of the problem, and of what is to be achieved. While it is important to clarify the initial problem statement, it can be advantageous not to finalise it until at least some of the background information has been collected. In particular the process of identifying relevant systems as components of larger systems and identifying the problem as part of a larger problem can assist in identifying the objectives.

Furthermore, engineering projects usually do not have just one, but rather a number of different objectives, and some objectives are very likely to conflict with others. The twin requirements of maximum performance and minimum cost are always going to be in conflict. It is always necessary to identify potential conflicts among the objectives.

For example, a dam can usually be used both to store water for supplying to a township or region and to assist in reducing down-stream flooding. Some dams are also designed for electricity generation. For effective storage, the dam should be kept nearly full, whereas for flood prevention it needs to be nearly empty and able to accept runoff after heavy rain. The operating policy for the dam has to recognise and allow for the implied conflicts.

Potential conflicts among objectives must be explicitly recognised. The question of how to deal with conflicting objectives will be discussed in some detail in Chapter 9.

Specifying design criteria, performance requirements and operating conditions

The goals and objectives of a project are expanded, clarified and quantified through the use of design criteria. Performance requirements and operating conditions are used jointly to specify how the system that is to be created must perform. These are used both in the detailed phases of planning and design, and in the preliminary phases to evaluate and rank alternative approaches and concepts.

In the creation of a physical system or device it is necessary to determine both the conditions under which it will operate and the level of performance that will be required. For example, if a pipeline is to be used to move natural gas from a production field to a consumption site, the performance requirements for the system will include the maximum and average quantities of gas that will be transported in a time period, as well as the minimum life of the operation. Such information is also needed for the planning and design of other components of the overall system, such as storage tanks and pumps.

The operating conditions of interest in the pipeline might include the variation in ambient temperatures throughout the year, the magnitudes of possible seismic action, and the chemical properties of the gas in relation to its effect on the pipe material.

Clear and unambiguous statements of goals, objectives, design criteria and performance requirements are particularly important should the completed project or system not reach expectations, with resulting legal dispute.

Safety and reliability are further performance requirements which require very careful consideration. They are discussed briefly in Section 3.13 following, and in more detail in Chapter 11.

Devising measures of effectiveness

The goals and objectives of a project are expanded, clarified and quantified through the measures of effectiveness, which are used in the feasibility study phase of project planning and in the preliminary and the detailed design phases. When a project has more than one objective, one or more measures of effectiveness are needed for each specific objective, and ideally with an overall measure that takes account of the different objectives.

Total cost is often appropriate as the overall measure of effectiveness. This is the case if the various objectives can be stated in terms of equivalent cost. Overall cost is also appropriate if the various objectives can be reformulated as minimum performance requirements, or as constraints that have to be satisfied. Not all situations lend themselves to monetary evaluation. For example, the measures of effectiveness might need to take account of matters such as aesthetics, environmental effects, risk of injury and loss of life.

Even when cost is used as the measure of effectiveness, conditions can become complex if carefully analysed. In the replacement of the superstructure of a bridge, total initial cost might at first appear to be entirely appropriate, so that it would be a simple matter to choose, say, from construction based on steel trusses, steel plate girders, in-situ reinforced concrete girders and precast, prestressed concrete girders. Purely in terms of cost of construction the plate web girder solution may be a clear winner. However, when maintenance costs are taken into account for the expected lifetime of the bridge, which may be between 50 and 100 years, the reinforced concrete girder solution may take precedence. But other factors might also influence the decision, such as the need for minimum disruption to traffic during construction. The problem is now to establish a time-cost trade-off. This might lead to the choice of a prefabricated truss system. Yet again, if appearance is important, as well as cost and time needed for construction, an extremely difficult trade-off has to be made between monetary value and appearance. The precast, prestressed concrete solution may now be competitive.

The manner in which the measures of effectiveness are formulated can be enormously influential in determining the direction that a project will take. In some circumstances an engineering project might be undertaken without the explicit use of a measure of effectiveness. For example, a design engineer might choose a design approach subjectively without a study of alternatives and a measure of effectiveness. If the project is small, the argument could be that an exhaustive problem formulation study would only add to the cost of the project. There are inherent dangers in such an approach, and they should be evident from the discussions to date. One possibility is that the wrong problem is solved. Errors of judgement easily occur if there is no measure of effectiveness. Even an attempt to define a measure of effectiveness can provide a fresh starting point, because it gives new insight into the project.

Iterating

The various activities listed in Table 3.3 all occur within the problem formulation phase; nevertheless, they may need to be undertaken iteratively. The type of background information that is required becomes progressively clearer after some attempt has been made to identify side effects and constraints and possible approaches. Likewise, some of the constraints and side effects are more easily recognised after some background information has been gathered. It will also be necessary to return to the problem formulation phase as we proceed through the later phases of planning and design.

3.8 FEASIBILITY STUDY AND CONCEPT DESIGN

The purpose of the feasibility study in planning is to show that the project can be carried out successfully. When a short list of promising alternative approaches has been identified, any one could form the basis of a solution. Likewise, the aim of the concept design phase is to establish a short list of promising concepts, each of which may result in a successful engineering design.

It is emphasised that it is best to begin with a wide a range of alternatives. We have seen how each option can be investigated superficially in order to cull non-feasible and non-competitive candidates. A further study, with a somewhat more detailed analysis of each option, leads to a further culling. The culling continues, using additional information, until there is a short list of feasible, promising options. A list of steps that can be useful in the feasibility study and in concept design is shown in Table 3.4. The steps are discussed below.

Table 3.4 Checklist of actions for feasibility study

Action
Check available resources
Investigate and quantify the constraints
Develop as many promising concepts as possible
Compare alternative concepts using the measures of effectiveness
Progressively eliminate non-competitive approaches
Modify the problem formulation as necessary
Modify the measures of effectiveness as necessary
Scrap or defer the project if no feasible approaches can be found
Identify the most promising concept

Checking available resources

Resources are used up in the course of an engineering project and an initial check is needed to ensure that sufficient resources will in fact be available. Resource availability (or non-availability) can determine the viability of particular options, so that this check should be undertaken with all alternative concepts and approaches in mind.

It is necessary to consider human, financial and technical resources, as well as any special materials and machinery that will be needed during implementation. Engineering expertise and scientific knowledge are resources, as are relevant trade and craft skills. Specialised design and analysis skills may be required. Time is a resource because engineering work always has to be completed within a limited time frame.

Investigating and quantifying the constraints

The technical and other constraints that have been identified in the problem formulation phase have to be investigated and, if possible, quantified. Some constraints are legal-technical, and are quantified in codes and standards and in legislation. In the design of a large city building, constraints on overall height, minimum services for water, sewerage, fire protection, thermal and acoustic insulation and even vertical transport, are imposed through the relevant ordinances and building acts.

Developing as many promising concepts as possible

This is the key step in the entire processes of engineering planning, design and problem solving. The success of any design or planning work depends on the range, quality and appropriateness of the approaches and options that are generated. As already emphasised, diversity in the alternatives is important. Innovative and creative new approaches, as well as traditional and proven ones, should be included in the initial list of alternatives. It is by no means clear whether an innovative new approach or a well-tried standard approach will eventually prove to be best. Both types need to be investigated.

The technique of problem-widening, described previously, should prevent an overly specific problem statement that would otherwise stifle new and unusual approaches. Creativity is central to this activity and is discussed further in Chapter 4, together with techniques that may prove helpful in the search for new and different options.

Comparing alternative concepts and progressively eliminating non-competitive approaches

The preliminary comparison of the options, made using simple calculations and preliminary information, will allow the poorer options to be culled. In the design of a highway bridge to cross a river, rough order-of-magnitude design calculations are sufficient to provide very approximate sizes for the main structural components of alternative systems. This allows preliminary costs to be estimated and for checks that the constraints and operating conditions are met. Information on costing is provided in Chapter 8.

In further rounds of comparisons, the increased accuracy of the analysis and design calculations lead to more refined comparisons and further culls.

Modifying the problem formulation as necessary

The need for an iterative approach throughout the problem solving process has been emphasised. Modification of the original problem statement may be needed during the feasibility study. As work proceeds and new and unusual approaches are considered and investigated, the understanding of the problem is inevitably improved, so that a better formulation of the problem may be possible. New and unexpected approaches can challenge the validity of the original problem statement, particularly in regard to the measures of effectiveness that are used to compare and evaluate the alternatives.

Scrapping or deferring the project if no feasible approaches can be found

If none of the options are feasible, perhaps because they do not satisfy constraints relating to time, cost or resources, then the project cannot proceed. It might be best to cancel the project, or to defer it until technical knowledge has improved to the level needed, or until additional resources become available. Another possibility is to look for new, innovative approaches that will make the project feasible. Yet another possibility is to reformulate the problem with different, more modest goals and less severe constraints.

3.9 PRELIMINARY PLANNING AND DESIGN

The purpose in this phase is to bring the search for the best option to a positive conclusion and to identify the approach that will lead to the best solution of the problem. Each short-listed option is investigated in turn and in sufficient detail to allow comparisons and rankings to be made, using the measures of effectiveness.

Even at this stage, very accurate comparisons are avoided if at all possible because of the cost implications. On the other hand, if alternatives are eliminated on insufficient grounds, the most appropriate alternative might also be incorrectly eliminated. The step-by-step approach therefore continues until all but one of the alternatives are eliminated.

Although attention is focused on the original options that emerged from the feasibility study, the search for new and better alternatives should not be discontinued. Modifications to existing approaches should be made in the later stages of the process if improvements are achieved. As work proceeds, there can be a build up in expertise which can lead, even at this stage, to further improvements and changes to the original problem statement and to the measures of effectiveness, as well as to new or modified design or planning concepts.

3.10 DETAILED PLANNING AND DESIGN

The aim now is to work out the details of the solution that are needed for full implementation. For example, at the end of the preliminary design of a reinforced concrete bridge the form of construction and the approximate overall dimensions of the component members will have been chosen. It is now necessary to fix the details, including final member sizes, the amount, type and location of the reinforcement in each member, non-structural fitments, concrete strength, concrete finishes, special road surfaces, handrails, bearing pads to support any main girders, and storm water pipes, so that full construction plans and specifications can be prepared.

Accurate calculations are needed in the detailed phase. But even here, iteration may become necessary, for example to develop alternative trial details for some components. If adjustments are made to the design details, with the aim of improving performance or decreasing cost, a new analysis might be needed to check whether the improvements have in fact been achieved. Such iterations continue until an effective and economical design or plan has been achieved, that meets all the requirements.

It is in the detailed phase of planning and design that optimisation techniques may be employed. If the behaviour of a component lends itself to theoretical modelling, it should be possible to improve the design by mathematically optimising the parameters that define or characterise the component. The process of optimisation is discussed in some detail in Chapter 13, together with various mathematical optimisation techniques.

At all stages of the detailed design, checks are made that the design constraints are not violated. In some situations overly severe constraints may add disproportionately to the cost or detract from the effectiveness of a solution. Even during the detailed planning and design phase it may be advisable to modify decisions made previously in the problem formulation phase.

In the detailed design of the components of a system, the primary focus should be on the overall operation and cost of the parent system. If the design of a component has a disproportionate effect on overall cost and effectiveness, then modifications may be possible for this component and for the interacting components, so that an improved overall design is achieved. If components are to be manufactured in quantity it may be desirable, depending on the nature and expected cost of the component and the number to be produced, to construct prototypes and test and modify them as an adjunct, or alternative, to the theoretical analyses.

The important final step in the detailed design and planning phase is full documentation, with a permanent record of relevant calculations and analyses and any other investigations that have been used to produce the final plan or design.

3.11 IMPLEMENTATION

Implementation of a plan, a design or a solution to an engineering problem can take many forms, depending on the context of the work. Special engineering fields with their own undergraduate programs and text books are devoted to the different types of implementation, such as construction engineering and manufacturing engineering.

Poor engineering design occurs when inadequate consideration is given to how the design is to be implemented. Constructability is an important criterion that is too often forgotten in the structural design of buildings. An undue focus on optimum design can lead to an elegant design with a minimum use of materials, but exorbitant construction costs.

We have specifically mentioned (if only briefly) the implementation phase here as a reminder of its importance in the overall scheme of planning and design.

3.12 THE SOLUTION-FIRST STRATEGY

The sequence of steps shown in Figures 3.1 and 3.2 follows from the methodology for solving open-ended problems. The necessity of iteration, and the advantages of sometimes undertaking several steps simultaneously, have been mentioned.

In some situations a rearrangement of the sequence shown in Figure 3.1 or 3.2 may be advantageous. It can be argued, for example, that choosing the evaluation criteria *before* the range of alternative approaches has been identified can prove to be an overly analytic approach. An alternative is to choose the evaluation criteria after some or all of the options have been identified. This may sometimes be a better alternative, although it may also introduce an unintentional bias in the criteria that favours some approaches over others. A better alternative might be to undertake these activities simultaneously. Various rearrangements of the sequence are possible and desirable in special circumstances.

To emphasise the fact that alternative sequences may be appropriate in some circumstances, we now discuss briefly a solution-first strategy. This stands in sharp contrast to the sequence in Figures 3.1 and 3.2. Numerous examples can be found in the history of engineering where an important project has *not* commenced with the identification of a problem but, on the contrary, has started with a potential solution. The task is then to search for an appropriate problem. A good technical idea may arise from some technical or scientific development, or by bringing new, potentially useful knowledge from another field of engineering.

Post-it Notes® (the adhesive that wouldn't stick)

The development of the Post-it note pads is a good example of the application of the solution-first strategy. In 1968 a 3M scientist, Dr Spence Silver, discovered a new type of adhesive, one that was quite different from anything currently available but whose properties defied conventional use. Dr Silver tried for five years to generate some interest in the new product, but because it was apparently inferior to existing adhesives he was unsuccessful.

Eventually a company researcher, Art Fry, took notice of the adhesive and its properties and started using it as a bookmark. The advantage here was that it was sticky enough not to fall out, but not so sticky that it left a residue on the page. By using the adhesive in this way it was soon noticed by others around the 3M Company and the idea of a lightly adhesive note pad soon developed.

The Post-it® note was introduced commercially in 1980 and named outstanding new product by 3M in 1981.

Source: 3M (2006)

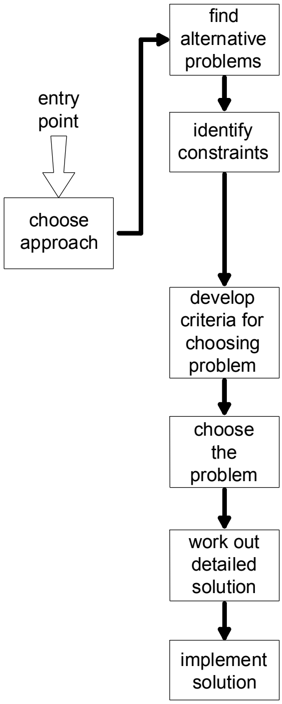


Figure 3.5 Solution-first approach to problem solving.

Ideas may arise concerning applications of new materials with unusual and useful properties that have been developed in non-engineering fields. Such scenarios lead to the search for possible applications of new ideas.

The solution-first approach shown in Figure 3.5 applies to such situations. A specific solution is the starting point to the procedure, which is a rearrangement of Figure 3.2, the steps being similar to those already discussed. This again emphasises the important point that engineering problem solving is an iterative activity. The entry point to the process is not of prime importance.

In a similar vein, Rogers (1983) has argued that answers often precede questions. He has suggested that while organisations face many problems, they typically possess only limited knowledge of a few innovations that can offer solutions. The chance of identifying an innovation to cope with a specific problem is therefore small. However, by commencing with an innovative solution there is a good chance that it can be matched to some problem facing the organisation. According to Rogers, a strategy for organisations is thus to scan for innovations and to try to line up a promising innovation with relevant problems.

3.13 OTHER ASPECTS OF PLANNING AND DESIGN

Risk, safety and failure

Once the goals and objectives of an engineering project have been established, the expectation is that they will in fact be achieved. However, engineering work inevitably involves some uncertainty and risk, and the possibility always exists that a system or a component will not perform as expected, or indeed that the entire project will not fulfil all of its objectives. When any engineering work is undertaken there is a risk of failure that, although small, is real and unavoidable. However, the levels of risk are controllable to some extent, and can be reduced to very small values that are acceptable to the community.

It is important to note that the term “failure” can mean different things to different people, and within the context of engineering work, it can be used with different meanings in different contexts. By way of example, consider the design and construction of levee banks as part of a flood mitigation scheme in a river valley. For the general public, and especially for people living in the valley, the word failure would be used in the event that the floodwaters overtop the levee banks and flow onto the adjacent land. In this sense, the word failure refers to a temporary condition, from which the system can recover. Indeed, the designers would have considered the risk of the flood event and the levee banks would have been constructed to resist the physical forces that occur during flooding, so that they will remain viable after the flood waters have receded.

However, in the design of the levee banks the engineers face several problems. Firstly, the maximum rainfall event that will occur over the design life of the project is unknown, and no upper bound can be placed on it. A second problem facing the designers is that the financial and physical resources available for the project will be limited. It is not therefore feasible to create a system that will contain all flood levels, irrespective of the intensity of the rainfall. Given these limitations, the task of the design engineer is not to create a system that will never “fail” in the sense that the floodwaters will overtop the banks, but rather to create a system that will successfully accommodate large rainfall events “most of the time”. The realistic design aim is therefore for the levee banks to prevent flooding, except for the extremely large rainfall events that will occur about once in every thirty years (or perhaps fifty years or even a hundred years, depending on the resources available and the resulting damage from overtopping). In one sense the system fails whenever the banks are overtopped, as we have seen, but from the design engineer's point of view the levee banks will be a success provided they reduce the frequency of flooding to an acceptably low frequency, of about once in every thirty (or so) years. For the designer, failure occurs if the levees are overtopped more frequently than once in thirty years on the average. In this example, the term "failure" has two quite different meanings, both of which are legitimate.

From the designer's point of view, a practical and useful engineering approach is to say that failure occurs if the performance of the system does not meet the performance levels that were set for it. These performance levels are carefully chosen during the design process when the minimum performance requirements, the design criteria and the effectiveness measures are identified. Nevertheless, in the previous example the system must be designed so that it can recover from flooding. We could also say that the system suffers a temporary failure during each flooding. In this text, and especially in Chapter 11 (which deals with risk and reliability), both meanings of the word failure will be used, but the meaning will be made clear from the context.

In regard to an engineering project, we can say that failure occurs if the aims and objectives of the project are not substantially achieved. Another quite different kind of engineering failure can occur if unintended, unpredicted and unacceptable side effects arise from the engineering work. Such failures can arise from errors made in the initial investigation stages of the project and from an incomplete and inadequate understanding of the problem.

Irrespective of the way we choose to define failure, there can be many different possible forms, or "modes", of failure to consider. Furthermore, the consequences of the different failure modes can vary enormously. The consequences may be very minor, such as when a pump in a water treatment plant breaks down, or they may be catastrophic, as in the collapse of a dam with loss of property, injury and even loss of life. Minor engineering failures are not uncommon, but the catastrophic ones, such as the collapse of a building, a bridge or a dam, are fortunately very rare.

Turning to an example from the field of structural engineering, we note that many different modes of failure are possible in critical cross-sections of a reinforced-concrete girder in a building frame. These include failures in flexure, shear, torsion, tension and compression, and various combinations. There are other possible modes of failure of the entire member to consider, such as excessive deflection. The many different failure modes have very different consequences, varying from severe (a torsion failure followed by a partial building collapse) through to minor operational disruption (excessive deflection under a temporary overload).

In any project, the designer must identify the possible modes of failure, evaluate the likely consequences of each mode, and design the system so that there is an acceptably low risk of failure appropriate for each mode. The level of risk that is acceptable will depend on the mode of failure and on the consequences of the failure.

There are many possible causes of engineering failure, ranging from human error and poor material quality, through to catastrophic natural events, such as severe earthquake or extreme flood. Human error can occur by miscalculations made in the design phase of the project, or by errors made during the implementation phase, or by mismanagement or accident during the operational phase.

The cause of a catastrophic failure may even be a very unlikely event that has been foreseen and allowed for in the planning and design phases. In the original design of the Tasman Bridge in Hobart, Australia, careful consideration was given to the possibility of a ship colliding with the main piers. Although this was considered to be a remote possibility, steps were taken in the design to limit the effect of such a collision (New et al, 1967). Nevertheless, the unlikely event occurred in January 1975, when a ship did collide with a pier of the bridge. The collapse of the main bridge deck had catastrophic consequences, with tragic injury, the loss of twelve lives and severe disruption to life in Hobart (Laurie, 2007).

It is not uncommon for the actual use of resources during the implementation phase of a project to be greater than was allowed for in the planning phase. Also, engineering projects are not always completed within the planned time schedule. Such over-runs on cost and time may be due to errors made in the planning, the design or the implementation phases of the project. They may be caused by supply chain problems, or even by extreme weather conditions. The over-runs may be small or large, but, strictly speaking, these are failures in the sense that we are using the word.

In open-cut mining projects a great deal of planning effort goes into minimising and managing risk, but it is accepted that accidents are unavoidable. Rescue and recovery operations are therefore included in the planning, design and on-going management of such projects. In petroleum and crude-oil handling facilities, all procedures are carefully planned and monitored with the express purpose of preventing fire and explosion. Accident, fire and explosion cannot be prevented completely, and have to be allowed for in the planning and design of such projects. Management procedures are carefully worked out to handle accidents when they occur, in order to minimise the consequences.

Even with the introduction of ingenious fail-safe concepts, the probability of failure in real engineering systems cannot be reduced to zero. The incremental cost of reducing risk increases sharply as the risk level decreases. In other words, it is extremely expensive to achieve a marginal increase in safety when the safety level is already high. Eventually a stage is reached in any design situation where the cost of improved safety is impossibly high. Engineering design must therefore deal with calculated risk and, in rare cases, the prospect of malfunction and failure.

Given that an element of risk is inevitable in engineering work, one of the difficult decisions is what is an acceptable risk level to be applied in planning and design. If the risk is too high and the safety requirements too low, the rates of occurrence of failure will be unacceptable to the community. If the risk levels are set too low, the cost of achieving the safety levels becomes unrealistic. A consequence of extremely low risk levels and high project costs is that fewer important projects can be undertaken. Methods for risk management and for dealing with failure always have to be analysed in detail. They are discussed further in Chapter 11.

While engineers understand that absolute safety is not an achievable or a feasible objective, this is not always appreciated in the community, or in legal circles. The risk levels adopted in engineering work have an indirect effect on the entire community, in regard both to the safety levels achieved and to the cost of the engineering infrastructure. It is for this reason that national codes and standards provide guidance and set minimum safety and performance requirements in many fields of engineering.

3.14 INTRACTABLE PROBLEMS AND SYSTEMS

The procedures for planning and design that we have discussed so far in this chapter make use of systems concepts that were introduced in Chapter 2. In effect, they provide a common-sense approach to the solution of complex problems which are ill-defined and open-ended.

Are all engineering problems amenable to this common-sense (and traditional) systems approach? Do all engineering systems lend themselves to decomposition and analysis?

It has been observed that some real world problems are too ill-structured to be dealt with using the standard systems approach. These have been described as *wicked problems*, as distinct from the *docile problems* that are amenable to standard approaches (Rittel and Weber, 1993). Wicked problems usually have a strong human dimension, for example when a number of individuals and groups of people are closely involved, both in the problem itself and in defining and finding an acceptable solution.

It has also been suggested that some real world systems do not lend themselves to standard planning and design techniques. A distinction has been drawn between *hard systems*, for which the procedures described in this book are applicable, and *soft systems*, which are intractable. Soft systems have been described as highly complex, lacking any clear structure, and possibly containing uncoordinated sub-systems that pursue their own independent goals. Again, such problem systems usually have a human dimension. A *soft systems methodology* has been developed by Checkland (1984) and his co-workers at the University of Lancaster for dealing with complex, intractable problems that are related to soft systems. It has also been suggested that the self-contained, limited, engineering project may not be the best way to deal with soft systems, and that alternative approaches may be more successful.

Wicked problems

Wicked problems have been discussed since the 1970s in regard to social planning and management problems that involve human activity, and more recently in regard to engineering planning, design and management. As already discussed in Section 3.6, there is an important human element in most modern engineering work, and community consultation and community input normally lead to acceptable decisions and successful outcomes. On the rare occasions when this is not the case, problems can become wicked, and this is usually because conflicting views and opinions in the community. According to Rittel and Weber (1993), wicked problems cannot be adequately formulated because additional relevant aspects are continually brought up for consideration. They suggest that a solution only occurs when a decision is made that the current trial solution is “good enough". It follows that there is no correct or best solution to a wicked problem, only that alternative solutions can be identified, compared and ranked. They also consider that each wicked problem will be unique, and that each wicked problem will lead to another problem.

Perhaps inevitably, sub-classifications of wicked problems have been proposed: *super-wicked problems* have been defined as wicked problems which have serious time constraints imposed on them, and which are partly caused by the same people that are dealing with the problem. The term "*mess*" was introduced by Ackhoff (1974) to describe a set of inter-related problems, or a system of problems. Further information on wicked problems and social messes can be found in the book by Ritchey (2011).

Various strategies have been proposed for dealing with wicked and intractable problems. When a problem is made intractable by the involvement of many paticipants with opposing views, an *authorative approach* may be possible. A small group of carefully chosen people, including experts, is given the responsibility of coming up with a solution. This approach is time-efficient because it sidesteps the need to deal in detail with the competing views and beliefs of the interested parties; however, it relies on the perspectives and experience of the group of people chosen to deal with the problem. An entirely different approach is to invite solutions from all persons and parties involved with the problem, and then to evaluate them and identify the best one. This is an *adversarial approach*. It can lead to tensions in the community and mutual distrust. A *compromise approach* has also been proposed, whereby an attempt is made to include all people who are likely to be affected, and to achieve collaboration using meetings and discussions to clarify the issues among competing interests. This approach will be very time consuming, but can lead to a solution that will be satisfactory to the majority of people, optimal for a few, and unsatisfactory for a minority. Compensation is always an added option for dealing with those significantly disadvantaged by a project.

In important engineering projects it is certainly true that very large numbers of people become involved, in one way or another. When different interested groups hold entrenched views, the project can take on a political dimension so that the engineering problems display the characteristics of wicked problems. Community consultations and negotiations then become very important, but may require additional political input.

Some commercial engineering problems can be exceedingly costly and can only be evaluated as successes or failures after they have been completed. Examples are to be found in the design, manufacture and marketing of new innovative appliances, and even of new models of automobiles and aeroplanes. In this respect the planning and design problems have the characteristics of wicked problems, although the normal procedures of planning and design and open-ended problem solving are applicable.

In previous generations, an authoritative approach was commonly taken to large engineering projects undertaken for the public benefit. A small group of politicians, engineers and other professionals were responsible for the decisions that affected the lives of many people.

Today, large, important and controversial engineering projects receive continuing public scrutiny, beginning at the initial stages of problem formulation.

Public opinion and comment are openly sought. This does not mean that the engineering methodology is not used. Rather, it means that the relevant phases of the planning and design processes can receive useful input from the public, and that the community will be well-informed and can become involved in the decision making processes. Nevertheless, on rare occasions agreement might not be reached, perhaps due to opposing, entrenched opinions, and political input may then become necessary. There is a growing literature on wicked and super-wicked problems on the internet which makes for interesting and useful reading.

Soft systems methodology

Checkland (1984) and his research group at Lancaster University in the United Kingdom, and other researchers, drew a distinction between hard systems and soft systems. Hard systems are relatively well-defined and lend themselves to traditional evaluation and analysis using the procedures presented in this book. In contrast, soft systems are ill defined and possibly indefinable because individuals and groups of people with differing views become involved and do not agree on what constitutes the system and the purpose of the system, nor on the problem that is to be solved. The idea of soft systems was originally developed in regard to problems in management and business, but is applicable to some engineering problems, particularly those which involve opposing views with political and sociological elements.

The soft system methodology proposed by Checkland and further developed by other writers is usually presented as a seven-step process. It is a sociological process rather than an engineering process, and is aimed at achieving accord among people and groups with radically differing views. Further details of the soft systems methodology can be obtained from a variety of internet sources.

When an engineering problem is intertwined with sociological and political issues, the standard systems concepts and planning and design methodologies can become unworkable in the face of divergent and opposing philosophical and political views in the community. In such situations sociological and political input is needed to allow a basic starting point to be found for developing an engineering solution.

Agile software development

If engineering problems and systems become intractable, it is usually because there are social, political and business issues to be resolved. But this is not always the case. Developers of large, complex, computer software packages have found that traditional planning and design approaches are inappropriate because of the extreme complexity of the systems being developed, and the length of time needed to create the complete, final plan or program, and then bring it into operation. Multi-purpose, multi-faceted software packages are quite different to, say, a piece of engineering hardware, such as a bridge, which must be brought into use as a complete entity. In contrast, various parts of a large software package can be developed, delivered and put into use long before the complete package has been conceived, let alone created. Some business software packages fall into this category. In their final form, these packages gather enormous amounts of data that fluctuate second by second; they then process, evaluate and analyse the data and so provide rapidly changing advice concerning the financial markets.

An alternative, *agile approach* to software development has become popular, which allows solutions to evolve progressively and adaptively, with different teams working simultaneously on coding diverse parts of the package. The emphasis is on the early production of working software and the regular delivery of additional components. Close collaboration among the teams of developers is obviously necessary. However, the approach also encourages close contact between the developers and the end users of the package. Furthermore, it allows changes to be made to the overall requirements and function of the package, even at late stages of development. This is an adaptive approach in which, initially, there is no accurate statement of the requirements of the end product. On the contrary, the requirements can change and develop over time and as the detailed work progresses.

This is a radical departure from the traditional methodology in which implementation begins as the detailed planning and design come to an end. Some proponents of the agile approach have emphasised that they are not promoting a different methodology, but rather the principles of a different approach.

The origins of agile thinking in software development can be traced back to the latter decades of the 20th Century, but the approach gained real momentum at the beginning of the new century, when varying views on agile principles (and even including a manifesto) were presented on the internet and published (Highsmith et al 2001). Generally speaking, the agile approach incorporates the following ideas:

- customer satisfaction is to be achieved by the early and frequent delivery of useful software;
- working software is the main measure of progress;
- adaptability is paramount, allowing requirements to change, even in the late stages of a project;
- close co-operation is needed, not only among the teams developing different parts of the software package, but also between users and developers; and
- teamwork and good communication are essential among the self-organising development teams that progressively produce the architecture and overall requirements.

In summary, the agile approach emphasises an adaptive, iterative and evolutionary approach to software development, rather than the traditional sequential approach of plan-design-implement.

Not surprisingly, many procedures have been used in implementing the agile approach. We will consider just one, called *scrum*, which is used to manage the work of a software development team. The aim is quick delivery of software, with an ability to respond rapidly to changes in requirements, to emerging technologies, and to changing user requirements (Schwaber and Beedle, 2002). The word "scrum" comes from the game of rugby football. In the scrum approach, a small team works closely together to produce new software components on a continuing, regular basis. Various roles are allocated to the team members. There is a product owner, who represents the interests of the customer or end user, but is also responsible for communication. The scrum master is the facilitator. A small team, usually of three to nine people, undertakes the development of software increments.

Work is undertaken in short *sprints* which typically last just several weeks, the aim being to produce useable software by the end of each sprint. A planning meeting is held at the start of the sprint to clarify the aims and the detailed work to be undertaken. A short stand-up meeting is held daily during the sprint and this is called a scrum. Each team member reports on progress made since the previous scrum, on the proposed work for the coming day, and, importantly, on any impediments that have been encountered.

Further detailed information on the scrum method, and on other agile methods, are to be found on the internet. Of course, problems have inevitably been experienced in the introduction of the agile approach, some of which have been documented. One recurring difficulty is accurate budgeting for an agile based project.

Agile management methods in engineering

It has been suggested that agile methods may be useful outside of software development and in engineering management and business management. The agile approach has been proposed for use in the automated development of engineering products such as computers, motor vehicles and medical devices. A suggested advantage of the agile approach in management is that it is much easier to deal with change and with unexpected occurrences when work is organised in short sprints. It is too early to judge how successful agile methods will prove to be in traditional engineering management applications. New engineering applications of the agile approach will be followed with interest.

A potential problem in applying the agile approach, and one not to be underestimated, is how to achieve a smooth and successful changeover from one mode of operation to another. Even when significant advantages are to be found in employing an agile approach, a too-rapid change in the management approach would cause disruption in any organisation. Careful planning for the changeover would be essential. A useful approach might be a step-by-step “hybrid” alternative, in which some suitable parts of the project are chosen and managed using agile principles. Components of both approaches could in fact be cherry-picked to obtain advantages from both approaches

3.15 SUMMARY

Large engineering projects require extensive planning and design work that is mainly carried out prior to the implementation phase. A major part of this chapter is devoted to describing and explaining procedures that are suitable for undertaking engineering planning and design. These procedures are based on a common-sense methodology of open-ended problem solving.

The processes of planning and design have many features in common, and in practice the terms are not mutually exclusive. However, we have used the term *design* when a physical object (or hardware item) is to be created, and the term planning when a non-physical object such as an operation or a process (or software item) is to be created. Despite this distinction, the processes are very similar, and we have emphasised the similarities in Sections 3.4 and 3.5 of this chapter, where the procedures for planning and design arew discussed.

The traditional plan-design-implement approach to engineering work is well suited to large and small engineering projects which deliver new items of infrastructure, or improvements to the existing infrastructure. Even when an engineering project is widely approved and leads to improvements in the physical infrastructure that will be enjoyed by the community as a whole, there are usually individuals and organisations that will be seriously disadvantaged. Financial compensation is a means of achieving equity in such situations.

Individuals and organisations in the community can become stakeholders when they take a keen interest in an engineering project, and can become active participants in the decision making processes. Public consultation is a normal part of modern engineering planning and design, and can be beneficial to all. The outcomes from public participation can be very positive, provided time and resources are allocated appropriately.

In some circumstances, however, problems can arise in the consultation process, especially if there are differing and entrenched views in the community. Sociological and political issues can then become intertwined in the engineering work. The consultation process may then end in controversy, without easy resolution. The terms *wicked* and *intractable* have been applied to such situations. When they arise, the plan-design-build approach may require political and/or social intervention. Wicked and intractable problems are relatively rare, but techniques for dealing with them are needed, and are discussed in Section 3.14.

There are some engineering problems for which the traditional approaches, described in this chapter, do not work satisfactorily, even when there are no social and political issues. Computer software development is a prime example. The *agile* approach has been developed specifically for the management of software development work, and various procedures, including *scrum* and *sprint*, have been developed and are now used to apply agile principles. There is potential for the use of these alternative approaches in traditional engineering planning and design, but substantial applications are still to come.

PROBLEMS

3.1 What are the main engineering problems to be faced in establishing a permanent human settlement on the moon?

3.2 How would you undertake a feasibility study for the problem of creating a lunar base to house between 10 and 25 persons in a reasonably comfortable environment for living and for undertaking scientific work? List the steps you would use in undertaking this feasibility study. Describe each step in a short paragraph and mention what you see as the key considerations. For example, if one of the steps is to obtain background information, what kind of information would you want to obtain?

3.3 What are the main engineering problems to be solved in establishing a small permanent Antarctic base for, say, twenty scientists? Describe briefly the steps to be taken in the concept design for such a base.

3.4 Consider the energy supply and distribution system of the city where you live. How would you undertake a planning study to provide adequate energy over the next thirty years? What background information would you want to gather in undertaking this study? List several alternative sources of energy that might be used to increase supply to meet demand. What other energy options would you consider in your study?

3.5 In constructing a building it is usual to begin from the foundations and work progressively upwards, floor by floor. Is this the only way to undertake building construction? Can you think of any situations where construction follows another sequence? Under what circumstances might it be advantageous to develop an alternative construction sequence?

3.6 The Government of South Australia is considering developing a nuclear waste storage site in a remote area of the state. Go through the steps of problem formulation for this project using Table 3.3 as a guide. Include in this process a list of the background data that you would require in order to adequately formulate the problem,

REFERENCES

Ackoff, R. L. 1974. Redesigning the future. John Wiley and Sons. New York.

American Society of Civil Engineers 1986. *Urban Planning Guide*. New York: ASCE.

Asimov E. 2010. *In Germany, a Highway Threatens the Mosel Wine Region. New York Times.* 26 March, 2010.

Chadwick, G. 1978. *A Systems View of Planning*, 2nd Ed, Oxford: Pergamon Press.

Checkland, P. B. 1981. Systems thinking, systems practice. John Wiley and Sons. Chichester, UK.

Department of Water, Land and Biodiversity Conservation, South Australia, 2005. *Water Proofing Adelaide, a Thirst for Change.*

DTV, 1974. *Atlas zur Baukunst*, *Band 1*, München, (Deutscher Taschenbuch Verlag).

Dym, Clive L. & Patrick Little 2000. *Engineering Design, A Project-Based Approach*. New York: John Wiley & Sons.

Fowler, B. & Rasmus, J. 2005. Seaside Solution. Civil Engineering, Magazine of ASCE, 75 (12), 44–49.

Grabosky, P.N. (1989) Electricity Trust of South Australia: Fatal Accident at Waterfall Gully. Wayward Governance: Illegality and its Control in the Public Sector. Australian Studies in Law, Crime and Justice Series, 161–171. Web reference: www.aic.gov.au/publications/lcj/wayward/ch10.html (Downloaded 24th January, 2006).

Griffis, F. H. & J. F. Farr 2000. *Construction Planning for Engineers*, McGraw-Hill International Editions, Engineering Series.

Hall, P. 1980. *Great Planning Disasters*. London: Weidenfeld and Nicolson.

Highsmith, et al., 2001. http://agilemanifesto.org/

Hyman, Barr 1998. *Fundamentals of engineering design*. New Jersey: Prentice-Hall.
Laurie, Victoria, 2007, Time Capsule: Tasman bridge disaster claims 12 lives. The Weekend Australian Magazine, 6–7 January, 2007, p. 6.
Lee, Colin 1973. *Models in Planning*. Oxford: Pergamon Press.
Maier H.R., Guillaume J.H.A., van Delden H., Riddell G.A., Haasnoot M. and Kwakkel J.H. (2016) An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together?, *Environmental Modelling and Software*, **81**, 154-164, DOI: 10.1016/j.envsoft.2016.03.014.
Neilson, A. M., 1991, Sydney Harbour Tunnel – Why? *Proceedings of the Seventh Australian Tunnelling Conference*. Institution of Engineers, Australia. Also in Tunnelling and Underground Space Technology, *Elsevier Science*, 6(2), pp. 211–4.
New, D.H., J.R. Lowe & J. Read, 1967, The superstructure of the Tasman Bridge, Hobart. *The Structural Enginee*r, 45, pp. 81–90.
Office for Water Security, (undated), South Australian Government, *Water for Good*, www.waterforgood.sa.gov.au.
Ritchey, T. 2011. Wicked problems- social messes. Springer Verlag
Rittel, H. and Weber, M. 1993. Dilemmas in general theory of planning, Policy Science, 4, 155-69.
Rogers, E. M. 1983, *Diffusion of Innovations*, 3rd Ed., The Free press, Macmillan Publishing.
Saaty, T. L. 1990. *The analytic hierarchy process: Planning, priority setting, resource allocation*, 2nd Ed., Pittsburgh: RWS Publications.
Spinner, M.P. 1997. *Project Management Principles and Practices*. New Jersey: Prentice Hall.
Smith, Craig B. 2004. *How the Great Pyramid was Built*. New York: Smithsonian Books, (Imprint of HarperCollins).
Schwaber, K. and Beedle, M. 2002. Agile software development with Scrum. Prentice Hall. ISBN 0-13-067634-9.
SSFM Engineers, Inc, 1996. *Preliminary Assessment of Korror-Babelthuap Bridge Failure for US Army Corps of Engineers*. Honolulu, Hawaii, 21 October.
Svensson, N. L. 1974. *Introduction to Engineering Design*. Randwick (NSW): University of New South Wales Press.
Voland, G. 2004. *Engineering by Design*. New Jersey: Pearson Prentice-Hall.
Yee, A. A. 1979. Record span box girder bridge connects Pacific Islands. *Concrete International*, June, 1(6).
3M (2006) www.3m.com/about3m/pioneers/fry.jhtml(23 January, 2006).

APPENDIX 3A: PLANNING FOR A CITY WATER SUPPLY SYSTEM

As an example of how the steps listed above in Tables 3.3 and 3.4 are used in the early phases of a project, we now discuss briefly the planning of the future water supply for a small coastal city in a very dry climate for, say, a 30 year period. The city of Adelaide in South Australia is the focus for the discussion. The aim here is not to provide valid, long-term conclusions in regard to the needs of this city. The emphasis is on the process, not the results. We will restrict our attention to the problem formulation step and the feasibility study.

Problem formulation

The steps in problem formulation, listed in Table 3.3, are applied (in hindsight) to the planning of the water supply system for the problem as it was in the year 2010 .

Identifying the problem as part of a wider problem

Given that there is already a serious lack of water throughout the region, the problem of planning for the future water supply in Adelaide has to be considered as part of the wider problem of providing water for the entire State of South Australia. Although about 70 per cent of all economic activity in South Australia occurs in Adelaide, an attempt to look at the Adelaide water supply problem without reference to the water supply problems of neighbouring areas, including industrial cities such as Port Augusta and Port Pirie, the South Australian wine industry and other rural industries, would be a serious mistake. The fluctuating demand for water in remote regions of the state, generated by extractive industries and mining operations, is also relevant and needs to be taken into account.

The geography of southern Australia and its lack of river systems makes it necessary to identify the problem as part of a wider problem: the management and use of water throughout the Murray-Darling river basin, which extends through three states in the eastern region of the Australian continent. This basin supplies the majority of the water used in South Australia, but there is significant upstream use of Murray water before the river enters South Australia, and this restricts the amount of water available.

The apparent change in climate that has occurred in recent years in the region (but also in many other parts of the world) suggests that there might be global aspects to this "local" problem.

The water supply system for the city also has to be looked at as one among many system components that make up the physical infrastructure of the city. Other components of particular relevance are the sewerage and storm water systems, which have to take up most of the input water, but are also potential suppliers of additional (processed, recycled) water. The parks and gardens and other public areas place significant demand on water. The transport system and the energy generation and supply systems are other components of the city system, as are industrial regions and the suburbs where people live.

Relevant background information

Quantitative data on the current state of the water supply system and its desired future states in thirty years time and in intermediate times need to be gathered. The population of Adelaide was a little over a million in 2010, when the water supply system provided about 200,000 ML per year. In addition, a little under 100,000 ML was used in neighbouring rural areas including the Adelaide Hills. On the demand side, about 45 per cent of city water was used by suburban households, about 28 per cent went to primary production, but only 10 per cent went to the commercial and industry sectors. Community use (including parks and gardens) made up the remainder, about 17 per cent.

Looking specifically at households, we find that about 40 per cent of domestic water went into gardens and other outdoor uses, such as swimming pools. Bathrooms (baths and showers) took up 20 per cent, with toilets at about 11 per cent, while laundries and kitchens use around 16 and 11 per cent, respectively.

On the supply side we find that although the water for the city comes from a variety of sources, in 2010 the two main ones were, firstly, the reservoirs and catchment areas in the Adelaide Hills, and, secondly, water pumped from the River Murray. Other minor sources of water for Adelaide include groundwater, rainwater tanks and storm water. In a “normal” year the two main sources provided roughly 60 and 35 per cent respectively of the water for the city and surrounding rural areas, but in “dry” years, such as 2007 and 2008, the component from the Adelaide Hills reduced to between 10 and 20 per cent. The additional water was obtained by extra pumping from the Murray, and severe restrictions were placed on the outdoor use of water.

These figures give a broad picture of the supply and demand situation in and around the city in 2010. Similar figures for the rest of the state, and indeed for south-eastern Australia provide additional important background information. For example, of the water available from the River Murray, about ten per cent went to towns and cities, while the rest was used in irrigation. Throughout the state there were about 50 small desalination plants in 2010, usually located in small and isolated regions such as Penneshaw on Kangaroo Island.

Background information is needed on the desired “target” state of the system in thirty years time, and at intermediate times. Such information regarding future demand necessarily depends on expected demographic changes to the city and state, and cannot be as reliable as the data on the current state. It could therefore be advantageous to look at several alternative scenarios, including optimistic, pessimistic and most likely, in order to obtain a range of figures for the target state.

The future population of Adelaide has been estimated to be about 2 million by the year 2027, and 2.5 million by 2050. The document “Waterproofing Adelaide”, prepared by the Government of South Australia, suggests a demand of about 230,000 ML in the city in 2025. This unexpectedly low value may include some allowance for a curb on demand as the result of future pricing policies. Even so, it is worrying when compared with the predicted supply figures for 2025 from present sources, which are about 190,000 ML in a dry year and 250,000 ML in a normal year. One reason for the relatively low supply figures is the reasonable assumption that climate change will continue and will, for example, reduce the available water from the Adelaide Hills by at least ten per cent.

It is clearly important to look for possible new sources of water in the feasibility study. Given the unavailability of new, plentiful and cheap sources of water in the surrounding hills, wider possibilities have to be considered. In anticipation of the steps in the feasibility study, we can see that background information will be useful on processes such as water desalination and recycling of used water, as well as methods for improving the efficiency of use of the existing sources of supply. Approximate (order-of-magnitude) information would be useful on more radical options such as the transport of icebergs (or iceberg water) from the adjacent Antarctic Ocean and measures to modify the location and quantity of rainfall. Another option requiring more accurate background information is the possibility of adjusting water usage in the south east of the country, perhaps through the marketing and trading of water and water rights.

Other relevant background information relates to the quality requirements of water when used in different ways. Clearly the quality of water needed for drinking and cooking is not the same as for irrigating public parks and gardens or for flushing toilets.

Regular updating of information

Predictions into the distant future for water supply are necessarily unreliable, concerning extremes of weather and other natural phenomena, are necessarily unreliable, but in the long term, extremes will undoubtedly occur in rainfall on catchment areas and flow in the River Murray. Procedures could be used to anticipate extreme events, including sensitivity studies and Monte Carlo simulation, but real data in the intervening years will be far more valuable than theoretical predictions. It is therefore essential that important data be gathered progressively so that the program can be revisited and updated regularly, perhaps every year or every two years. This updating must begin immediately and continue while the project is underway, and also after it has been completed, so that modifications can be introduced in a timely manner and at minimum cost.

It must also be remembered that planning is not being undertaken solely for the supply of water at some remote future time. Adelaide is a functioning city and requires an ongoing plan to provide adequate water throughout the intervening years up to the target year of 2050.

Underlying needs

While a useful picture of future domestic, industry and rural water needs can be built up for the planning period, the severe lack of water raises important questions that highlight the differences between needs and requirements. Specifically, considering the relatively high domestic use of water, with 40 per cent going to gardens and other outside applications, an important question is whether a distinction needs to be made between the quality and cost of water when used for different purposes. Such questions arise when we look at the underlying needs, as distinct from demands, for water. This leads to the further question of whether water of varying quality should be supplied with pricing policies related to cost of supply. At present, almost all water is supplied at the one quality level. The issue of water trading focuses attention on the difference between needs and demands. Such questions are relevant when we undertake the feasibility study. Our purpose here is not to try to answer such questions, but to illustrate the process.

An important step in the planning process could well be to place relevant questions before the community for discussion and debate. Already in the year 2006 in various country communities in Australia questions were being examined, such as the acceptability of recycling water as a way to boost the capacity of town water supplies.

Side effects

Increasing pollution of the waterways in the Adelaide Hills and pollution and salinity problems in the River Murray are already prevalent. Worsening pollution is an obvious side effect to be expected in the present project. The side effects of depleting ground water through aquifer pumping are not well appreciated in the community, but require investigation. Increased salt in the aquifer water and overall deterioration of quality could become serious side effects should there be a significant increase in the use of ground water supplies. Over-use of river water generally means reduced flow, with adverse effects such as silting, algae blooms and other adverse environmental effects on the biota. Cessation of river flow and closure of the river mouth is likely. Closure has already occurred at the Murray mouth, and the local community is well aware of the consequences.

Other side effects include increased salinity in the river water of the Murray due to increased pumping, and the need for correspondingly more expensive treatment procedures to achieve adequate water quality. Another consideration is the inhibiting feedback effect of poor quality water on industrial expansion, city development and the rural and wine industries.

Constraints

Constraints derive from state and federal regulations regarding water quality, environmental impact, land resumption, health regulations, city planning and zoning, and water use. World Health Organisation standards for water quality and health also serve as constraints. Current intrastate agreements on the allocation of water from the Murray-Darling basin impose severe constraints on the available options for the city of Adelaide. In this regard the possibility of modifying existing laws and regulations has to be considered, with the re-writing of state and commonwealth agreements.

Objectives

In the statement of objectives, account has to be taken of the demands and needs for water, and the likely limits on supply. Water scarcity and interstate and intrastate competition for this limited resource also have to be considered in the statement of objectives. A rather general statement is as follows:

> *To match, as closely as is economically and technologically feasible, the quantity and quality of water provided against the realistic needs of the various uses (domestic, industry and irrigation) and to provide security of supply over the planning period.*

This formulation treats both demand and supply as variables. On the other hand it skirts the problem of differentiating between need and demand. The prediction of realistic figures for demand versus need thus becomes a central task to be undertaken in this project. The above statement does not stipulate that water will always be available without restriction, for example in times of severe drought, although some minimum level of supply is clearly implied.

The linked concepts of technological and economic feasibility allow for changes and improvements in technology over the planning horizon. For example the economic feasibility of desalination plants is likely to improve substantially in the next decade, assuming that there is continued development and improvement in efficiencies in the relevant technologies.

Performance requirements and operating conditions

The water supply system will have to accomodate a progressive increase in the quantity of water to be supplied through the planning period, and not just in 2040. A wide variation in conditions in this period must be expected, including effects such as progressive climate change and alternating drought and flood throughout the state and in the upper reaches of the Murray-Darling system. Drought and flood can even occur at the same time in different parts of the system. In taking account of these possibilities, it will be necessary to specify minimum limits on the quantities of water to be supplied on a daily, weekly and yearly basis. Quality limits also have to be specified. Separate statements of the requirements for domestic use, industrial use and irrigation use may be needed. Above all, the performance requirements have to be realistic and achievable.

Measures of effectiveness

Total cost in dollar terms is one prime measure of effectiveness, but certainly not the only one. Another measure is needed in regard to the reliability of supply, which can be measured, for example, by the number of days per year when restrictions occur. Another measure may be needed for the quality of the water supplied, taking account of the average and peak salinity levels. Adaptability of the supply system to unexpected changes in demand over the medium term (in other words, system robustness) is another possible measure of effectiveness.

A combined measure of effectiveness might be made in terms of the dollar cost per unit of water supplied, using penalty costs to take account of reliability and adaptability. Such a measure of effectiveness can be broadened to allow for water being supplied at several different quality levels according to different uses. This leads to a separate measure of effectiveness for the different water quality levels considered. A single overall measure of effectiveness, again in dollar terms, could be obtained by differential costing of the different quality levels. The present comments are intended to show how a measure of effectiveness might be developed. They are not meant to indicate preferred options.

Feasibility study

We now look very briefly at the steps listed in Table 3.4, as they apply to the present example.

Checking available resources

In addition to the main resource, water, we have to consider the financial resources that will be available, and the technical resources and expertise that will be needed to undertake the project. When each specific option is considered, such as the construction of a new dam or the construction of a desalination plant, the specific resources and expertise have to be investigated.

Investigating and evaluating constraints

Some of the constraints to the problem have already been mentioned, such as the regulations and laws concerning water quality. It will be necessary in the feasibility study to investigate and study them to observe how they will affect the possible options and solutions.

Developing promising concepts

Some of the possible options for dealing with the problem have been foreshadowed in the previous discussion. Consideration has to be given to increasing the use of the available but presently unused water resources, such as the storm water which flows into the sea. An improved efficiency in the use of other existing resources is another important option, for example by reducing the losses that occur due to evaporation from open water surfaces and leakage from reticulation pipes. Another option is legislation to require the installation of domestic rainwater tanks. Further options that are more costly include desalination, recycling and reuse of treated wastewater. More radical approaches, which may appear at first sight not to be feasible, include the harvesting of icebergs from the Antarctic Ocean. The possibility of actively changing the climate to increase rainfall falls into this category.

A more obvious and potentially very effective approach is to curb the increase in demand for water, and possibly even reduce demand, through pricing policies coupled with education programs to alert the community to the real costs of water and to methods for effectively reducing its use.

Technical evaluation is needed to determine the relative costs, the feasibility and the quantities of water that can be delivered by each of the approaches. It is unlikely that any one approach will be adequate in itself. It is far more likely that a range of the cheaper and more effective measures will be used simultaneously so that, together, they produce the volume of water needed.

While some of the approaches listed might not be feasible today, we must remember that although the planning horizon is thirty years, the need for water will not suddenly stop in 2040. With improving technology and new scientific discoveries, we can confidently expect that some of the options that today seem unlikely will become quite realistic over time and may well be introduced at later stages as the century progresses.

Comparing alternatives and eliminating the non-competitive options

In this project we are dealing with a scarce resource, water, and, as already mentioned, a mixed strategy is likely to be appropriate. The employment of some options and not others will be decided on the relative costs and efficiencies, which will undoubtedly vary over the planning horizon.

The example being discussed here is different to many engineering projects, in that the best solution here will *not* be found by progressively eliminating all but one from a list of options. The nature of the problem means that a mix of options will be required, with only the clearly uncompetitive ones eliminated.

An evaluation of viable options can lead to their ranking in terms of initial set-up cost, cost per unit volume delivered, and total amount of water deliverable.

Such a ranking could then be used to make decisions on the mix of options to be employed initially and at subsequent stages during the planning period. Regular updating of the list, especially when new developments and discoveries are made, will allow decisions to be made on when specific options should be taken up. At the beginning of the century, unlikely options such as desalination, rain making and harvesting icebergs should not be discarded and forgotten; they may well become attractive in time, as conditions change, and as the options initially employed are exhausted.

Post script

Detailed information and numerical data on the supply of water to Adelaide, and to the state of South Australia, are available from two planning documents that were released by the South Australian Government. The earlier document, from 2005, was entitled "Water Proofing Adelaide", and presented a plan for ensuring water for Adelaide up until the year 2025. The second document, undated but apparently from the year 2010, was entitled "Water for Good", which extended the planning horizon to the year 2050".

The second plan was produced just 5 years after the first one. The reason for this may well be that in the years 2006 to 2009 the water inflows into the River Murray system were unprecedentedly low, and provided just a very small proportion of the average yearly inflow. The River Murray has been traditionally the main source of water for the city of Adelaide, together with local catchment areas. As a consequence of this apparent, sudden change in climate, a desalinisation plant was commisioned in 2014 with an annual capacity of 100 Gl, about half of the city's usage. Following the construction of the desalination plant, rainfall increased and the traditional sources of water have been more plentiful. In the period up to 2017, the desalination plant had not been used to capacity, except for short trial runs.

CHAPTER FOUR

Creativity and Creative Thinking

Creativity plays a crucial role in engineering and in particular in the planning and design of engineering projects. In this chapter we find that the brain works in some quite surprising ways; that much of its success comes from the use of simple heuristics (or search rules), and from hard-wired features that allow it to make reasonable decisions quickly. However, this form of thinking is convergent in nature and does not provide a sound basis for the divergent thinking we need for the creative aspects of planning and design. It is, therefore, important to develop techniques to enhance a divergent and creative approach to problem solving. We review some of the available techniques in the context of engineering problem solving.

4.1 INTRODUCTION

The aim of this chapter is not to turn everyone into an Einstein or Picasso but to demonstrate the need for a creative and open mind in engineering and to show how it can be achieved in practice by practise. We have seen in Chapters 2 and 3 that engineers are often charged with the responsibility of solving complex, open-ended problems; ones that have no single obvious solution and ones that will benefit from a novel and creative approach. This may seem easy initially, but, as will be shown in this chapter, our brains have evolved to be very efficient at making quick decisions that are often quite appropriate, using processes that have developed to ensure the survival of the species. However, these are generally not the processes needed for *creative thinking* in engineering planning and design.

Before we look at the operation of the brain, it is worth defining clearly the types of thinking that are important to engineers. On the one hand, there is *convergent or deductive thinking*, which aims to apply logic and take in just as much information as is necessary to develop a solution to a particular problem. We have seen in Chapter 2 that this is the sort of thinking that occurs when people are trying to solve an analytical problem, or making a decision by narrowing down the alternatives. Gilhooly (1982) quotes a finding that scientists working on the analysis of lunar samples as part of the Apollo mission tended towards convergent thinking: when presented with an open-ended question, they would quickly transform it into a more tightly defined one. *Divergent or inductive thinking*, on the other hand, is aimed at solving problems where there might be a large number of possible solutions or where there are no obvious solutions. Divergent thinking deliberately takes in as much information as possible without too much concern about relevance, and tries to develop solutions by making links that are not necessarily obvious or (apparently) sensible.

An example of divergent thinking is the case of the Swiss engineer, Robert Maillart (1872-1940). In 1901 he designed and constructed a bridge at Zuoz in Switzerland and noted a couple of years later that a section of concrete webbing near the main supports was cracking. The obvious solution was to repair and strengthen it, but Maillart instead investigated the behaviour of the structure and found after some study that the presence of cracks in those locations was not having a detrimental effect on the bridge's performance. On this basis he modified his design to take advantage of this finding, leading to a much more efficient cross-section for the girders (Ferguson, 1999). Many must have observed the cracking at Zuoz, but he was able to think about it in quite a different manner and come to a quite different and highly creative solution.

We will now explore creativity (or divergent thinking) in engineering – what it is, how it can be recognised, how it can be promoted, and why it is important in the planning and design of projects. It is worth noting at this stage that much of what we know about creativity comes from work undertaken by highly creative individuals and that when we consider creativity, we are looking not just at the solutions to engineering problems but also to problem formulation. As Findlay and Lumsden (1988) have argued, this "may be just as much a part of the creative process as its solution".

4.2 CREATIVITY DEFINED

Creativity is an elusive concept. Bohm (1996) believes that it is impossible to define in words, but others have tried. One definition that picks up many of the elements is by Gardner (1993) who defines the creative person as one "who regularly solves problems, fashions products, or defines new questions in a domain in a way that is initially considered novel but that ultimately becomes accepted in a particular cultural setting." He goes on to suggest that "initial rejection is the likely fate of any truly innovative work". Gardner was looking for much more than general creativity, and was attempting to pick out individuals who had shown significant creative output. As a matter of interest, he selected Albert Einstein (physicist), T.S. Elliot (writer), Sigmund Freud (psychologist), Mahatma Gandhi (politician and statesman), Martha Graham (dancer), Pablo Picasso (painter) and Igor Stravinsky (composer) for his list from a wide variety of fields, and justified each based on his definition.

We propose that, for engineering planning and design, creativity must demonstrate novelty, value and eventual acceptance in the posing or solution of engineering problems. The requirement for eventual acceptance means that it is not possible to dismiss ideas at the time they are generated. It is worth remembering the numerous examples of ideas that were ridiculed when first presented, only to be accepted some time later when it was agreed they came "before their time".

In the next section we look at the brain and its workings with the aim of showing just how constrained human thinking often is, and how much of what we do is governed either by features that are hard-wired into our brains or by the unconscious application of simple rules or heuristics. In the later sections, various procedures are outlined to demonstrate how it is possible to coerce the brain to ignore these rules, and to work in a way that promotes creativity.

4.3 THE BRAIN AND ITS WORKINGS

As recently as two hundred years ago little was known about the brain, and serious research into the functioning of the brain really only began in the first half of the 20th Century. The fact that many parts of the brain are named after their shape rather than function gives an indication of just how little was known. For example, in the brain there are areas called the “pons”, “amygdale” and “hippocampus” named after bridges, the shape of almonds, and seahorses, respectively (Rose, 1998). It is only quite recently that it has been possible to determine the actual functions of some brain regions using techniques such as functional magnetic resonance imaging (fMRI), which is able to measure oxygen uptake rates as a measure of neural activity, and positron emission topography (PET) which can measure, in addition to oxygen uptake rates, the release of neurotransmitters such as dopamine (Knutson and Peterson, 2005). Based on these studies it is now known, for example, that the hippocampus is important for short-term memory and the ability to make mental maps, while the amygdala handles fear and other emotions.

The basic building block of the brain is the *neuron*. Humans start with around one hundred billion neurons, but the number declines with age. Neurons (see Figure 4.1) have a cell body with input (dendrite) and output (axon) channels that connect to other cells. The number of input and output channels per neuron can be many thousands. The contact between neurons is made at synapses and each neuron has between a thousand and ten thousand synapses. A section of brain the size of a grain of sand contains about one hundred thousand neurons, 2 million axons and one billion synapses (Ramachandran and Blakeslee, 1998).

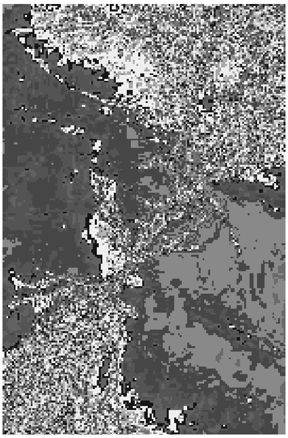


Figure 4.1 A neuron in a rat hippocampus showing the cell body and a small number of the connections to adjoining cells. Note that the staining technique highlights single neurons and so does not show the density of connections to adjacent cells. © Used with permission of Synapse Web.

The connections in the brain develop from birth (Klawans, 2000) and are strengthened by use. The infant brain starts with a high level of redundancy in the form of parallel paths. In fact, one of the strengths of the human brain is based on its parallelism which is measured as *fanout*: the number of direct contacts a neuron makes with other neurons (Holland, 1998). Typical computer elements have a fanout of about 10, whereas human central nervous system neurons have a value between 1,000 and 10,000. It is believed that this explains much of the potential for complex behaviour. Until more details are available on exactly how neurons store information, it is not possible to determine a measure of brain power based on a simple neuron count but a comparison with other animals is instructive. Camazine (2003) points out that honey-bees, with approximately one million neurons, have the capability to navigate by the sun, fly to a food source, make what appear to be decisions, communicate with other honey-bees, and perform other complex activities. Boden (2004) tempers any enthusiasm for reading too much into these sorts of abilities with a description of the hoverfly which, although capable of quite complex behaviour, has much of it hard-wired into its brain and nervous system. With this understanding, she believes that the hoverfly's intelligence "has been demystified with a vengeance".

There is still debate about the exact mechanism that the brain uses to store and process information, but it would appear that the synaptic connections are heavily involved. When messages are transmitted the synaptic gap (approximately 1/40,000 mm) is reduced, making further communication along that path easier. In this way pathways are established leading to patterns that the brain can use. Pathways that are not used retain a large synaptic gap and are therefore less likely to be used. Researchers in computing are using this model to add some 'intelligence' to the Internet. According to Brooks (2000), a mechanism for making active World Wide Web links has been developed so that links that are used most commonly are given prominence and those that are not used for a time are discarded. In this way the Internet may be able to optimise data collection by highlighting the most common paths and making them easier to follow (a convergent thinking approach). Whether this will actually assist people gathering information is yet to be seen. There could be an argument for not wanting to follow all the paths that everyone else on the Web has followed but allowing a few less-travelled ones to come up in searches (a divergent thinking approach). One way to achieve this might be to allow searchers to specify an equivalent acetylcholine level for the search. Acetylcholine in the brain is important in strengthening synapses. It is in particularly high levels during REM (rapid eye movement) sleep and it has been suggested (Phillips, 1999) that this might be a reason that dreams during REM sleep can be so bizarre.

One of the findings that has come from research is the level to which the brain manipulates raw data before presenting it to the consciousness. As an example, consider the sense of sight. It has been discovered that when a human 'sees', information is being passed to approximately 30 different areas of the brain, with each working on a specific aspect of the vision (Ramachandran and Blakeslee, 1998). For example, there are separate areas for colour, depth, motion, slopes of lines, edges, and outlines. According to Goleman (1995) studies have identified neurons that only fire in response to certain facial expressions such as a threatening opening of the mouth, a fearful grimace, or a docile crouch. There is also a group of neurons that fire in response to 'seeing' the shape of the back of a monkey's hand (Popper and Eccles, 1977). What this means is that many decisions are effectively made using hard-wired sets of neurons rather than conscious thinking. We do not look at someone's face and actually have to think about what it is showing – as soon as our eyes gaze on the face, neurons fire in response to aspects they have evolved and the brain is fed the information that the face is showing; for example, a threatening opening of the mouth. This results in extraordinary speed, but at a cost: it is automatic, and we do not necessarily realise that.

It is not commonly known, but every eye in every human has a significant defect, i.e. a blind spot. The fact that there is such a problem was first discovered centuries ago by a researcher studying the optics of the visual system, and is due to the geometry of the lens and retina. That it should be so well hidden says more about the brain than the eye. The brain actually assists the eye and 'fills in the missing details' automatically. An example of a test picture that can illustrate the effect is shown in Figure 4.2. To demonstrate the effect: hold the diagram close to the face with the left eye looking directly at the circle and the right eye closed, then move the page away from the eye. At a certain distance the cross should disappear from the peripheral vision and be replaced by the general background.

Once the effect has been observed, just stop and think for a second. The brain is 'making up' some of what it reports it is seeing! If this is true, then people cannot or should not necessarily believe what they are seeing, because everything that they pick up is filtered and modified by the brain, mostly in a harmless way, but modified nonetheless.

The human brain is the result of approximately 200,000 years of evolution since *Homo sapiens* split from earlier ancestors (Gould, 1991). For most of that time humans were living in a more primitive world where many of the decisions were ones of life and death and where survival led to an emphasis on quick reasonable solutions rather than slower well thought out ones. Cohen and Stewart (1995) make this point in relation to avoiding black and yellow stripy things. Often they might not be tigers, but it is not worth having a complicated and time-consuming procedure to work this out because this will not aid survival!

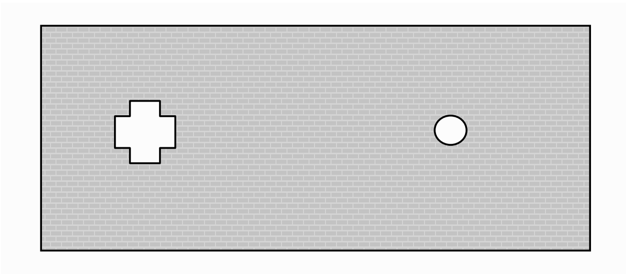


Figure 4.2 Example of a diagram that can be used to illustrate the optical blind spot. Hold it close with the left eye looking directly at the circle and the right eye closed, then move the page away until the cross disappears.

We now turn to conscious thinking and find that there are also issues with this activity, especially where creativity is concerned. Most high school graduates majoring in mathematics and science would consider themselves to be well trained in thinking. However, much of the training has been to develop skills of a very specific kind. There are various terms for the traditional type of *thinking*: convergent, analytical, vertical, and logical are just a few. Application of this type of thinking means that people start at a particular point and move one step at a time, justifying each step as they go and narrowing down the options until it leads to a clear solution. This type of thinking is particularly true in mathematics, for example, where there is generally a correct way of arriving at an answer. Even if there is more than one way, each follows a logical sequence of steps that lead on from the last. Although this type of thinking is important, and particularly so in engineering, there is another type of thinking that is equally important, namely, creative, divergent or lateral thinking. De Bono (1970) classifies the key differences between vertical and lateral thinking and these are listed in Table 4.1.

Table 4.1 Key characteristics of the types of thinking by de Bono (1970).

Convergent/Vertical Thinking	Divergent/Lateral Thinking
• moves forward in logical steps	• generation of new ideas
• selective	• generative
• rightness matters	• richness matters
• analytical	• provocative
• sequential	• jumps
• correct at every step	• does not have to be correct
• use of negative to block pathways	• no use of negative
• follows most likely paths	• follows all paths
• finite process (expects answer)	• probabilistic
• excludes that which is irrelevant	• welcomes intrusions

The concept of right brain and left brain

In the late 1960s Roger W Sperry, an American who was later to become a Nobel laureate in 1981, discovered that the human brain has two very different ways of thinking.

In popular psychology, “left-brain” and “right-brain” became terms associated with both personality traits and cognitive strategies, where a left-brain individual or cognitive style is typically verbal and processes information in an analytical and sequential way, looking first at the pieces then putting them together to get the whole, and a right-brain individual has a more creative style that is visual and processes information in an intuitive and simultaneous way, looking first at the whole picture then the details.

However, based on fMRI, Nielson et al. (2013) identified the broader left-dominant and right-dominant connectivity networks, with a more consistent scheme including left-dominant connections associated with language and perception of internal stimuli, and right-dominant connections associated with attention to external stimuli.

A student educated in science and mathematics might look to the right hand side of characteristics and be horrified at some of the entries. For example, “does not have to be correct” and “welcomes intrusions” seem like odd ways of tackling serious engineering problems, but these are in fact ideal when one considers that the brain has not been designed to tackle tasks where one wants lots of solutions and where the time taken to arrive at the solution is of limited importance.

4.4 THE BRAIN AND HEURISTICS

According to work by a number of researchers (e.g. Gigerenzer and Goldstein, 1996; Borges et al., 1999; Cross and Jackson, 2005) much of the convergent thinking we do can be described based on the application of very simple rules (*heuristics*). These heuristics are interesting for two reasons: firstly, to see how real decisions are made, and secondly, to highlight the need to force the brain to act otherwise in situations where quick and dirty decisions are not wanted.

Recognition heuristic

We start the description of the recognition heuristic with a simple example. In one experiment (Reimer and Katsikopoulos, 2004) a group of U.S. university students was asked to say which city was larger, San Diego or San Antonio. The correct answer was given by 66percent of the group. When a similar group of German university students was asked the same question, 100percent of them selected the correct answer. So how is it that German students seem to know more than U.S. students about American cities and their size? The explanation given was in terms of the recognition heuristic (Goldstein and Gigerenzer, 2002). In this type of question with only two alternatives, if someone recognises only one of two options, then they will assume that that one is likely to be the correct answer. In this case, it was likely that the German students had heard of San Diego but not San Antonio and so selected the former. The recognition heuristic is often referred to the "less is more effect". The less one knows, the better the chance of getting the correct answer! The method will of course only work in situations where there is a systematic bias in the options, but is important nonetheless. Of course, once people know more than the basic minimum that particular heuristic breaks down and people move on to other more complicated ones.

Minimalist heuristic

The next stage in decision making is to use a very small piece of additional information and the minimalist heuristic. Continuing the search for the larger of two cities: if one is forced to choose between two about which a little is known, it might be good to ask which has an airport, or a national soccer team – then choose that one. Again, this way of thinking gives a quick and easy answer and will be reliable under a wide range of situations.

One of the minimalist heuristics is 'Take The Last'. In this, a problem is tackled using the scheme that was used to solve a similar looking problem last time. This may seem a sensible approach and will often work, but it can have its problems. In one experiment reported by Gilhooly (1982), a group of people were given three water jugs of different volumes and asked to make up a particular volume of water by manipulating the quantities. For example, when given 18, 43, and 10 litre jugs and asked to make 5 litres people found, after some manipulation, that the answer was to fill the 43 litre jug, then use it to fill the 18 litre jug and then the 10 litre jug twice leaving the required 5 litres. Having gotter good at this and other similar problems, they were then given 28, 76, and 3 litre jugs and asked to make 25 litres. The subjects had difficulty because they were intent on using all three jugs because that had been necessary previously.

Take the best heuristic

The next stage in the decision-making process is to gather additional data, and to sift through it until one choice comes up as the best. At this stage the search stops. According to Gigerenzer and Goldstein (1999) the "Take the Best" heuristic works as well or better than more complicated decision-making methods such as linear and multiple regression in selected situations, and is faster as well. This conclusion was based on the study attempting to determine the larger of two cities where nine pieces of information were available for each including, whether it was a capital city, or had a soccer team, an intercity train or a university. It has been suggested that this is the process that humans go through when searching for a mate. They set up a list of requirements and search until someone matches enough of them, at which time they move to the next phase of decision making.

Availability heuristic

One of the aspects that has fascinated people studying human behaviour is the way people assess danger and its relative likelihood. Shark attacks and aeroplane crashes, for example, are relatively rare yet people will overestimate their perceived danger from such events. This has been put down to what is called the availability heuristic where "individuals estimate the frequency of an event, or the likelihood of its occurrence, by the ease with which instances or associations come to mind" (Wänke et al., 1995). As Tversky and Kahneman (1973) point out, the ease of recall should be related to relative frequency (and often is) but it should be understood that this short-cut is being used because there are times when it does not work. An example of where it works was given in an experiment by Tversky and Kahneman (1973) where subjects were given 7 seconds to estimate how many flowers, or four-legged animals or Russian authors they would be able to list in 2 minutes. Their estimates and subsequent performance was highly correlated and the authors suggested that a very quick assessment allowed people to make an accurate prediction of their recall ability.

However, the heuristic is open to abuse and can perform poorly in some circumstances. Tversky and Kahneman (1973) also described examples where it did not work. In one experiment, subjects were asked to estimate the relative frequency of words starting with the letter 'k' compared to words with 'k' as the third letter in typical text. Most could think of more words starting with 'k' so estimated that these were more frequent. In fact, words with 'k' as the third letter outnumber words starting with 'k' 2:1 in typical text. In a second experiment, subjects were asked to estimate (very quickly) the product: 1 x 2 x 3 x 4 x 5 x 6 x 7 x 8. Others were asked to estimate (very quickly) the product: 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1. The median guess for the first group was 512, while the median for the second was 2,250. The true answer is 40,320. Tversky and Kahneman suggest that in estimating the product, subjects looked at the first few terms and extrapolated based on that. They were unable to appreciate the exponential nature of the sum, particularly those who were considering where 1 x 2 x 3 ... would be going.

The use of the availability heuristic has important ramifications for engineers. As an example, Kates (1962) studied the decision-making process in relation to the provision of flood protection works:

> *A major limitation to human ability to use improved flood hazard information is a basic reliance on experience. People on flood plains appear to be very much prisoners of their experience ... Recently experienced floods appear to set an upper bound to the size of loss with which managers believe they ought to be concerned.*

First instinct fallacy

When students take multiple choice question exams and tests, conventional wisdom states that one should go with the answer that was chosen based on first instinct. In fact, some guides aimed at helping students recommend this course of action. However, it is very likely wrong despite people's opinion to the contrary. Why this is so can be explained very well by looking at how the brain reacts to being right and wrong and what regret might follow a wrong decision.

In a series of experiments with university students undertaking multiple choice exams, Kruger et al. (2005) tested student experience and came to a number of conclusions:

- the majority of answer changes were from incorrect to correct, and most people who changed their answers improved their test scores;
- changing the correct answer to an incorrect answer was likely to be more frustrating and memorable than failing to change an incorrect answer to the correct answer;
- the majority of changes were from wrong to right, but student's intuitions said the opposite;
- there is more regret associated with switching from a correct to an incorrect answer than failing to switch from an incorrect answer to a correct answer;
- there is a memory bias: students overestimate how many times they have switched from correct to incorrect and underestimate how often they failed to switch from a wrong answer;
- people who were observing the selection of answers from multiple possibilities were more critical of those who switched from a right answer than of those who did not switch from a wrong answer; and finally
- the first instinct fallacy strengthens in the face of mounting personal evidence to the contrary.

The important point in all of this is to realise that when humans make decisions they are pre-programmed to make "fast and frugal" ones (Gigerenzer and Goldstein, 1996). While this is good for survival in a more hostile world, or appropriate when pressed for time, it must be realised that if a team is developing a solution to an engineering problem it is important to be careful to avoid taking one of the quick solutions that the brain generates so easily, and to work in a more methodical way. It will not feel natural. It will be fighting against everything the brain has ever wanted to do, but it should be done.

4.5 CREATIVITY AND THE EUREKA MOMENT

When Francis Crick and James Watson realised that one of the implications of DNA's double helix was the potential to explain the passing on of genetic information, it seemed to be one of the many breakthroughs based on a brilliant flash of inspiration by another of the geniuses that fill the pages of history. But, it is not quite as simple as that. Crick's sudden insight occurred in 1952, but by that time he had been studying the problem of the structure of DNA for over a year – thinking of little else, discussing little else, working on little else. It may have been a brilliant flash, but he had prepared himself well for it.

The more one studies people that are regarded as creative geniuses, the more it becomes evident that there is much more to creativity than the sudden flash of inspiration. Archimedes in his bath, and Newton and the apple under the tree: in each case the real story is much more complicated than the popular folklore, and in each case the discovery came after years of hard work.

The original Eureka!

Many know that Archimedes discovered something while overflowing a bath and ran down the street shouting Eureka, but the details are not always as clear. Koestler (1964) gives a good description.

Eureka is Greek and translates as "I have found it!". The use of it to describe some sort of mental breakthrough dates from Archimedes who had been given the task of determining whether a crown given to his king was in fact pure gold, or gold mixed with a lesser metal. Knowing the density of gold the task was therefore to decide the volume of the crown and to check its density. A lower density would indicate that the gold had been mixed with a lighter element. Archimedes pondered the task of deciding the volume. If he could melt it down he could measure its volume but that was not an option. Whether it is apocryphal or not, the story goes that one day as he was getting into his bath, Archimedes noticed that as he got in his body displaced the water and it overflowed. He realised that if he submerged the crown it would displace its own volume and he would be able to measure it accurately. He then set off down the street overjoyed at his discovery.

There appears to be a general agreement that it takes approximately 10 years of consistent work before significant outcomes eventuate in scientific discovery, and the fact that there is little apparent progress does not mean there is no progress. According to Bronowski (1978), "the discovery is made with tears and sweat (at any rate, with a good deal of bad language) by people who are constantly getting the wrong answer". Gilhooly (1982) and Miller (2001) catalogue the stages involved in a creative leap:

- preparation, where the problem solver familiarises himself or herself with the problem by engaging in "conscious, effortful, systematic and usually fruitless work on the problem";
- incubation, where the problem is set aside and no conscious work is done on it;
- illumination or inspiration, where a break is made which, even if not the complete solution, leads to an advance; and
- verification, where conscious work is done to test the idea.

Several of these stages deserve further examination. The incubation stage is an interesting one and highlights a mode of thinking that many believe to be crucial to human thought: that carried out by the subconscious. Both Albert Einstein and the French scientist Henri Poincaré, for example, were firm believers in the power of this type of thought and deliberately structured their work to make the most of breaks from conscious thought. While sitting not thinking may sound appealing, it should be noted that it presupposes that work has been done beforehand so that the brain has something to work with! The act of creativity is most clearly associated with the third stage, where things come together in a novel and useful way. A key point that is made by a number of researchers is that the new idea does not appear from nowhere, but comes from the making of novel connections between information already stored in the brain. On this point, Poincaré believed that new mathematical ideas must be "beautiful" and that the brain actually used this beauty to identify the potentially valid ideas from the mass of others (Poincaré, 1908). This need for existing knowledge underlies many of the methods designed to enhance creativity, as we shall now see.

4.6 TECHNIQUES FOR STIMULATING IDEAS

At this stage it is worth reflecting on what we have found so far about human brains and thinking. Brains like short-cuts and have developed a whole range of them to allow us to make quick and often quite good decisions. Brains also hide much of their inner workings and are far more pre-programmed to work with anticipated situations than we realise.

Engineers and scientists are generally well trained in convergent thinking; making use of a minimum of information to come to a solution quickly. They may even resist divergent thinking and the sorts of operations necessary to carry it out. Yet, in many engineering planning and design problems this is exactly the type of thinking that is needed. Therefore, it may be necessary to use a series of seemingly odd and somewhat artificial procedures to force the brain to act in a manner that suits these types of problems. It will not feel natural. It will not make sense. However, with practice, it will allow the development of some quite innovative solutions that no level of convergent thinking could ever approach.

For many problems that are presented to engineers, there will be an obvious 'common sense' solution that most will see quite quickly. The value that can be attributed to the idea that led to the common sense approach is therefore minimal. However, if one engineer comes up with quite a different solution, one that no-one else had thought of, its value is much higher. That is the aim of creative thinking:

to come up with those different ideas. Much of the time they may not lead to a better approach, but every now and again they will and those times are worth striving for.

If any further encouragement was required to promote creative thinking it might come from knowing how many ideas it takes to come up with a truly successful one. Stevens and Burley (1997) carried out such a study by investigating the relationship between new ideas in an industrial setting and how many progressed on to success as commercial products. Based on a wide range of views and data they suggest that for every 3,000 raw ideas, about 150 are developed to the stage where a patent application is filed. Of these applications about 112 are issued with patents of which only 9 have commercial significance. Of those, only 1 leads to major commercial success. It seems we need ideas, and we need lots of them.

One of the common aims of the various methods is to force the brain to look at things in a different light and to make connections that it would not normally make. According to Findlay and Lumsden (1988) "the creative process ... is derived from the establishment of new links among already existing elements". In the following sections, a number of established and accepted methods for enhancing creativity are outlined. They are not in any particular order and it should be noted that each method will have situations where it is suited and some where it is not. Therefore, they should be read with an open mind. Nothing stifles creativity more effectively than a sneering cynic.

Attribute listing

Attribute listing works best where there is an existing solution or product and a better one is being sought. As Svensson (1974) has pointed out, this may be with the aim of eliminating defects from a product, improving its operation or reducing costs. The steps involved are: list all the components or elements of the existing solution; list all the attributes of the component (e.g. weight, size, colour, shape, material etc.); and then systematically change each attribute in every conceivable way.

As an example, taken from Miller (2000), consider an improved design for a heating system radiator using attribute listing. The elements of the existing product, their present attribute and other possibilities are listed in Table 4.2. There are two complementary ways the information at this stage can be used. Firstly, if improvements are required to the existing designs, then some of the possible attributes can be evaluated while leaving the others fixed. For example, it may be beneficial to consider the effect of changing the colour of the product. Superficial changes of different shades of white or cream are possible, but what about a paint that changes colour as it heats? This may enhance the design and improve safety by reducing the chance of accidental burns or scalds. There are similar options that could be tried for all the attributes. Alternatively, is there a quite different finish that does not involve paint? Or might the finish make the product essentially invisible?

Table 4.2 Product attributes and other possibilities for a house radiator.

Element	Present attribute	Possible attributes
material	metal	plastic, ceramic, paper, wood
shape	rectangle	circle, oval, sphere, ellipsoid, tube
mobility	fixed	mobile, float, sink, roll
liquid	oil	coloured water, gas, liquid metal
opacity	opaque	transparent, semi-transparent
finish	smooth	rough, carved, regions of texture
colour	white	red, blue, black, fluorescent, variable

Secondly, combining all present and possible attributes leads to 5 x 6 x 5 x 4 x 3 x 4 x 6 = 43,200 different possibilities and these could be evaluated. One of these is the current configuration, and many will not make any sense, but some surprising and challenging alternatives may come from this, and may prove valuable. For example, some of the variations include:

- plastic sphere that rolls around the floor filled with oil;
- transparent plastic tubing filled with coloured water;
- black metal tubing filled with hot oil;
- plastic tube that is mobile and filled with oil; it comes in a variety of colours.

The challenge is to see if and how these can lead to something new and viable. If nothing else, the various combinations can be used to foster discussion, and to explore what is really required for a solution. In any case, the main aim has been fulfilled in that the designer has been forced away from seeing all home radiators as flat metal components fixed to a wall. At the early stages of the design process, the rule should be quantity not quality of concepts. Another important consideration is that sometimes ideas that come up using this technique can be better applied in a quite different area. For example, consider the last of the four examples listed, the mobile plastic tube filled with oil. While this may not be suitable for warming a room, it may be ideal as a hand warmer or a foot warmer for cold nights. It was not what was asked for but it is worth being open to the unexpected. The mobility idea could provide the germ of further ideas: why for example should one heat a whole house if it is possible for a person to carry a heat source around and so be comfortable anywhere?

Morphological synthesis

Morphological synthesis (also called “matrix analysis”) is similar in some ways to attribute listing in that it involves the generation of unusual combinations of system functions; however, it is more focussed on generating completely new ideas rather than just improving existing ones. The steps involved are: describe the problem by its systems; list the major system parameters; list the alternative ways of satisfying each system parameter; draw up a matrix with each system parameter as one dimension; and then consider all possible combinations of ways of satisfying each parameter.

For example, it is desirable to establish innovative public transport systems in cities, and then a morphological synthesis can be used to generate alternatives. If the three major system parameters are considered to be the route alternatives, energy source and method of support the results can be generated as shown in Table 4.3.

Table 4.3 Morphological synthesis used on public transport.

System parameter	Possible solutions
route alternatives	rail, road, monorail, water, air
energy source	liquid fuel, electricity, gas, solar, wind, animal
method of support	wheels, air, magnetic, wires, rail

The method has generated 5 x 6 x 5 = 150 combinations, which include:

- an electric powered monorail with magnetic levitation;
- gas powered vehicles travelling over water with air suspension;
- solar powered buses with wheels on a road.

Many of the 150 may be infeasible; some will already exist; others will not make any sense at all. However, if there is one that is worth further consideration, or one that sparks a new idea of some sort, then the time spent using the method has been well spent.

Setting a minimum quota

One of the main problems in developing novel or creative solutions is the tendency to apply one of the brain's shortcut heuristics and go with the first viable solution that comes to mind. In many cases this may appear to work, but it is a convergent thinking technique and it is possible that other equally viable (and possibly superior) solutions may be overlooked. To overcome this, it is considered good practice to set a minimum quota of ideas which must be generated. In this way the tendency to follow the best idea as soon as it occurs is overcome. De Bono (1970) suggests a limit of around four or five as a minimum. Better advice is to estimate how many solutions there are, and then seek double or triple that number.

As with many of the methods that are used to develop creativity, the setting of a minimum quota works on a number of levels. Firstly, it forces a number of solutions to be generated. Secondly, and perhaps more importantly, it forces the engineer to think about the problem more deeply and to consider much more carefully what might be considered a solution.

The idea of setting a quota does not actually help generate new ideas, but it forces the search to continue and is certainly worth the effort. McCormick (2004) warns against the practice of setting rituals where a certain number of ideas must be generated for a 'correct' solution, so it is important to keep in mind the importance of the exercise rather than the idea that a certain number of ideas will produce the correct answer (as there is no single correct answer).

Challenge assumptions and implied solutions

One of the issues with problem solving is the tendency to impose limits on the solution due to assumptions which are made even without realising it. For example, consider for a moment the task of creating a new design for a toothbrush. Even with no further instructions it is likely that one would assume that the solution being pursued would have to be portable, cheap, hand-held, able to use standard toothpaste, taste and odour free, able to be purchased easily, and instantly recognisable as a toothbrush. With all these constraints in place it is no wonder that any innovative design that is unlike what currently exists is virtually impossible. But are these really necessary? A truly innovative approach might come from altering the question to “design a new way of protecting teeth” and this in turn could lead to the search for a much wider solution than the current toothbrush affords. Whatever the outcome, it is important to realise that humans are very likely to self-impose creativity-limiting assumptions as part of a convergent approach.

Suspended judgment

The purpose of thinking creatively is not to be right, but to be effective. In general thinking there is a tendency to judge each and every aspect of new ideas and to try and discard them on any particular or perceived flaw. History, however, provides ample examples of people who were quite wrong about aspects of their work or thoughts and yet made great advances. According to Kolb (1999), Kepler, the astronomer who formulated the three famous rules of planetary motion developed them based from a question that came to him in the middle of a lecture to undergraduates: why are there six planets? From that poor starting position the excellent work that demonstrated the properties and consequences of the planets’ elliptical orbits followed. In fact, in the field of astronomy good results have come from ideas and calculations containing basic errors. According to Thornton (2000), Neptune was discovered based on calculations that showed there was something (possibly a new planet) affecting the orbit of Uranus. The calculations contained an error but, as luck would have it, one that had little effect at the particular time the search for the cause of the disturbance was undertaken. Had the search been undertaken at a different time, the error could have made it a futile one.

According to de Bono (1970), there are several advantages with suspending judgement on ideas:

- an idea will survive longer and may breed further ideas;
- other people will offer ideas which they themselves may have discarded but which others can make something of;
- ideas can be accepted for their stimulating effect; and
- ideas which may be judged wrong in the current frame of reference may survive long enough to show that the frame of reference was wrong.

The way in which suspension of judgement is used (de Bono, 1970) is that:

- one does not rush to judge or evaluate an idea, one explores;
- some ideas which are obviously wrong can be used to explore, not why they are wrong but how the idea could be useful;
- even if one knows that an idea will eventually be thrown out, one keeps it for as long as possible to get as much out of it as possible; and

- instead of forcing an idea in a given direction, one follows the idea to see where it will lead.

Marconi and radio transmission

There have been occasions when people have done things which were impossible or wrong, and it turned out that despite everything being against it the final result was what they were seeking. Guglielmo Marconi succeeded in transmitting radio waves across the Atlantic believing that the waves would follow the curvature of the earth. He had been working on transmitting radio waves and had been able to achieve it over increasing distances. First in metres, kilometres, and across the English Channel. He was convinced he would be able to do a transatlantic transmission so he set up a transmitter in England and set up an aerial held aloft by a large kite in Newfoundland. He waited. People told him radio waves would travel in straight lines so he would not be able to do it. But he did it, and in that respect was correct. The people were also right when they said radio waves travel in straight lines, but did not understand that the earth had an ionosphere from which the waves could bounce. If Marconi had listened to others he would not have made the breakthrough.

Gilhooly (1982) reports on a study where subjects were asked to generate solutions to a particular problem for 5 minutes. One set of subjects was asked to work in a conventional manner, evaluating ideas as they went. In the other they were asked to defer judgment. A team of judges then evaluated the ideas. The first group produced 2.5 ‘good’ ideas (on average) while the second produced 4.3 ‘good’ ideas. The higher productivity was explained in terms of suspending judgement.

Brainstorming

Brainstorming is a technique that was developed in the 1940s and 1950s by a businessman, Alex Osborn (Gilhooly, 1982). When it was first introduced it took off quite quickly, and has now been applied in industry and evaluated in the laboratory. According to Gilhooly (1982) “numerous studies support the hypothesis that groups using brainstorming produce more ideas than similar groups that work along conventional lines”. There are also reports of not only more ideas, but better ideas too, coming from brainstorming sessions.

A brainstorming session usually involves 6 to 12 people. Two officials are needed; a chairperson and a recorder. The chairperson introduces the topic, enforces the rules and is responsible for maintaining a continuing flow of ideas. The recorder makes a note of all ideas. The topic must be formulated in such a way that it is not too vague. Since the brainstorming session will be straining the limits of creativity starting with a vague problem could lead to an over-wide collection of ideas. At the same time the problem must not be over-formulated so that some exploration around the topic is possible. This comes with experience. The rules of a brainstorming session are (de Bono, 1970; Perkins, 2000):

- all ideas must be recorded;

- no idea may be criticised or evaluated during the session (suspend judgment);
- it is permissible to build on the ideas of others; and
- the emphasis is on quantity of ideas; quality follows from quantity.

It is better to record the ideas on a whiteboard or similar so that all participants can see them. At the start of the session the chairperson introduces the topic and sessions tend to last around 30 to 45 minutes. As a way of enforcing the rules and maintaining the flow of ideas the chairperson should:

- stop people trying to evaluate or criticise other people's ideas;
- stop people all talking at once;
- make sure the recorder has written the idea down;
- fill in gaps by offering ideas;
- suggest different ways of tackling the problem;
- define the central problem and keep pulling people back to it;
- end the session either at a designated time or when things are flagging; and
- organise the evaluation session.

Culp and Smith (2001) suggest that it may be beneficial if, at the start of a session, everyone spends a couple of minutes writing down their own ideas with no discussion. Then the session starts with the ideas being read out in turn. This is designed to ensure everyone gets an initial word in and that the more dominant personalities do not take over the session.

Often, in the course of a brainstorming session the flow of ideas can wane and there is a prolonged silence. A 'trigger' is a very useful device to restart the ideas. A number of triggers have been suggested. These include:

- a checklist;
- wildest fantasy;
- juxtaposition.

One checklist is provided by the acronym SCAMPER and is listed in Table 4.4. The use of 'wildest fantasy' as a trigger encourages all participants to think of the wildest idea. Some of these may lead to other more practical solutions. 'Juxtaposition' involves the following steps:

- select three words at random;
- discover a connection between each word and the problem being studied.

The best words to use are familiar ones such as water, meat, cup, or pen. For example, in trying to suggest ideas for the design of a vehicle which does not use fossil fuels a group has run out of ideas. They decide to use juxtaposition as a trigger, and select the words pencil, water and chicken. The word 'pencil' then may lead to the suggestions of using wood as a fuel, making the vehicle long and thin to reduce drag, using graphite as a fuel or lubricant, and a vehicle which walks on stilts. The other words will also generate discussion in a similar way.

Table 4.4 The checklist SCAMPER.

Substitute	Who? What? Other material?
Combine	Units? Purposes? Ideas? Other place?
Adapt	What else is like this? What can I copy?
Modify	Change meaning? Shape? Colour?
Put to other uses	What is different about this use? What is similar?
Eliminate	~ parts of
Rearrange	Interchange components? Other layouts?

Water supply in the developing world

According to New Civil Engineer (2002), the World Summit on Sustainable Development held in Johannesburg in 2002 developed an aim to halve the number of people without access to proper sanitation and to have all people with access to clean water by 2020. In a surprising twist, the World Bank and the United Nations called for no new infrastructure but a program of education and water supply improvements. Although large schemes seem the logical way to go, experience has shown that unless there is full subsidy of the schemes they are rarely sustainable and the result is a flawed system of broken pipes and pumps and a lack of local skilled personnel to effect repairs.

"The role of the engineer in these countries will need to change, along with their perception of what constitutes success."

There must be no evaluation of ideas in the session. That is a task for later. Statements that must be discouraged include: "that would never work because ... ", "it is well known that ... ", and "how would you get to do that ... ". The ideas generated should be evaluated several days later. Evaluation is better carried out without consultation between participants. One possibility is to give several people the list and ask them to identify the best 10%. These could then go on and be further evaluated and developed. In the evaluation session which follows some time later, the central tasks are to

- pick out the ideas which are directly useful;
- extract the ideas which are wrong or ridiculous;
- list functional ideas and new ways of considering the problem;
- pick out those ideas which can be tried out with relative ease even if they seem wrong at first;
- pick out those ideas which suggest that more information could be collected in certain areas;
- pick out those ideas which have already been tried.

Although there are supposed to be some benefits in having a number of people in a group feeding each other ideas and benefiting from the group dynamics there is also evidence that a single-person brainstorming session can be beneficial.

Here the main benefit comes from the deliberate suspension of judgement and the idea that ‘anything goes’. Studies have shown that a nominal group made up of a number of people working individually can combine outputs to produce a better set of solutions than a group of the same size working as one (Gilhooly, 1982). One *Nominal Group Technique* (Delbecq et al.,1975) can be as simple as individuals first thinking of as many ideas as possible to solve a particular problem in a designated time. Then pairs are formed and all ideas are combined and further ideas are added building off the individual lists. No ideas are to be rejected. Then two pairs are combined with all ideas listed and any more ideas added within a designated time period. Assessment of the ideas can be done by combining all the groups' ideas and then everyone having a certain number of votes such as 5, 4, 3, 2 and 1 to allocate for the ideas that they favour.

Reading and contemplation

A common theme among many individuals who have made significant breakthroughs is the place reading and study played in their lives. Charles Darwin was a voracious reader and his natural selection breakthrough came to him following his reading of a work by Thomas Malthus on population dynamics. Albert Einstein’s theory of relativity came after reading David Hume’s *An Enquiry Concerning Human Understanding* and he made no secret of the benefit he had obtained from reading it (Edmonds and Eidinow, 2006). Isaac Newton and numerous other scientists of note were all well known for their reading that covered a wide range of topics from science to philosophy. The power of reading also comes to lesser mortals. James Ellis was a leader in the field of code-breaking and code-making. Singh (1999) suggests that one of the reasons for Ellis’ success was his breadth of knowledge:

> *He read any scientific journal he could get his hands on, and never threw anything away. ... if other researchers found themselves with impossible problems, they would knock on his door in the hope that his vast knowledge and originality would provide a solution.*

The same must be true for much of life and life’s experiences. A wider knowledge base provides different ways of looking at things and a much greater chance of seeing something from a different perspective (Ferguson, 1999):

> *More important to a designer than a set of techniques (empty of content) to induce creativity are a knowledge of current practice and products and a growing stock of first-hand knowledge and insights gained through critical field observation of engineering projects and industrial plants.*

The benefits of reading come not only from the additional knowledge that may be gained, but also from the stimulating effect that reading has on the brain. As has already been noted, creativity comes from making novel connections between pieces of information stored in the brain, and the active thinking that goes on as one reads assists in this process.

More than a bit player: Howard Hughes Snr.

The movie “The Aviator” (starring Leonardo DiCaprio and Cate Blanchett) tells the story of the early years of Howard Hughes Jnr. who made a name for himself producing Hollywood movies, building and flying aeroplanes, setting speed and distance flying records, and generally making the most of his money, youth and talent. The source of his early money was based entirely on a single invention that his father, Howard Hughes Snr, had patented: a drill bit for the oil industry that was particularly good at drilling through rock.

Hughes Snr. had been in the mining industry, but was lured to Texas with the discovery of oil near Beaumont in 1901. He worked in a number of locations and by 1907 had a partner in a small business although Bartlett and Steele (2003) note that he was “usually too independent-minded to work with anyone”. In 1908 Hughes met a millwright, Granville Humason, who had designed a new drill bit and who showed him a wooden prototype. Humason had come up with the idea for the drill one morning as he ground his coffee and had shown it to a lot of miners but no-one had been interested. Hughes offered him $150 on the spot for it, which was accepted. Hughes then rushed to work up a patent application and had the device patented in the U.S. and overseas. The final drawing was carried out on the kitchen table. Again, according to Bartlett and Steele (2003):

He emerged from the family dining room with the Archimedean cry of ‘Eureka’ and the picture of a bit that had no less than 166 cutting edges.

And the rest, as they say, is history. Hughes went on to be awarded over 70 patents and founded a company that generated tens of billions of dollars. The interest in the story is to consider what it was that made Hughes Snr. successful? According to the biographers, he was a stickler for details. They also contend that he would have no doubt been aware of the previous attempts to patent such a drill bit and would have recognised the significant breakthrough it represented. Perhaps the real success was based on a good notion of what the idea was really worth and the need to protect that with good patents. Hughes spent at least $1500 on that first patent and worked hard to maintain and protect it.

Fostering creativity

While one of the aims of this chapter is to promote creativity in individuals, a related aim is to ensure that those who are creative will have their ideas received in a positive environment. People should be able to recognise and appreciate creativity, and remember that it will not always be logical, immediately useful or even sensible under current thinking. According to a co-worker (quoted in Singh, 1999) of James Ellis, the cryptographer mentioned earlier:

> *He was a rather quirky worker, and he didn’t really fit into the day-to-day business of GCHQ. But in terms of coming up with new ideas he was quite exceptional. You sort of had to sort through some rubbish sometimes, but he was very innovative and always willing to challenge the orthodoxy. We would be in real trouble if everybody in GCHQ was like him, but we can*

tolerate a higher proportion of such people than most organisations. We put up with a number of people like him.

Creative individuals are not necessarily those that are highly intelligent (there is only a moderate correlation between creative ability and IQ) and so many will not stand out in an academic sense. In many cases their ideas will be of little value. Note above that the workers with Ellis "had to sort through some rubbish sometimes", but in the end having creative people around is well worth it because when virtually everyone comes up with the same solution, the truly creative person will be the one who has a different idea that no one else had thought of or considered. That is when they come into their own. And that is why everyone should try to be more creative and to appreciate and nurture those who are. Goel and Singh (1998) have a similar message when they urge businesses to "understand that creative ability of most people is fragile and could be seriously suppressed by destructive criticism" and to design a company structure that "enhances rather than detracts ... from useful creative activity".

4.7 SUMMARY

The brain is a complex, power-hungry, secretive, manipulative and vital organ that defines human beings. It gives us the *sapiens* in *Homo sapiens* and is the product of 200,000 years of evolution since our species split from earlier ancestors. The brain has developed a particular skill in coming up with reasonable solutions to many problems in the minimum of time, often without any formal decision-making process, often using a large number of hard-wired techniques.

If the brain is to be used for creative thinking a conscious effort must be made to force those with convergent thinking away from natural methods of solution and towards ways that ensure that unusual or unexpected results can follow. There are a range of methods that can be employed including attribute listing, morphological synthesis, the use of a quota, brainstorming and extended reading and thought. Each takes practise and time, but it will be time well spent.

PROBLEMS

4.1 Apply attribute listing to the problem of trying to generate at least 50 new ideas for a different type of ball point pen.

4.2 Use morphological synthesis to generate at least 20 new ideas for a new way of drying clothes in a typical suburban house.

4.3 In 2 minutes write down as many uses that can be thought of for a paperclip. Start now!

4.4 Use a one person brainstorming session to devise ways of reducing fuel consumption for the average family car. In 10 minutes try and generate as many ideas as possible.

4.5 Assume the government has put out a tender for development of a mobile kitchen that could be used for the army in exercises in remote and arid areas. The unit must be suitable for preparing hot meals for a group of 20 soldiers. Generate at least five quite different solutions to the problem.

4.6 Suggest five creative solutions to the problem of an existing bridge that vibrates excessively when people run over it.

4.7 The people who prefer not to define creativity often resort to the argument that they can recognise it when they see it. Consider your group of friends and acquaintances and identify creative ability in them. What are the main distinguishing characteristics?

4.8 If the Eureka moment comes after much concentrated study and thought, what are the implications for students wishing to master a subject at university?

4.9 If truly creative people really do see life differently and think about problems and solutions in very novel ways, what fraction of the population do you think they should be? What are the issues to be considered?

REFERENCES

Bartlett, D.L. and Steele, J.B., 2003, *Howard Hughes, His Life and Madness*. (Penguin), 687pp.

Boden, M.A., 2004, *The Creative Mind. Myths and Mechanisms*. 2nd Edition, (Routledge), 344pp.

Bohm, D., 1996, *On Creativity*. (Routledge), 153pp.

Borges, B., Goldstein, D.G., Ortmann, A. and Gigerenzer, G., 1999, Can Ignorance Beat the Stock Market? In: *Simple Heuristics that Make Us Smart*. (Eds. G. Gigerenzer, P.M. Todd and the ABC Research Group), (Oxford University Press), pp. 75-95.

Bronowski, J., 1978, *The Origins of Knowledge and Imagination*. (Yale University Press), 144pp.

Brooks, M., 2000, Global Brain. *New Scientist*, Vol. 166, No. 2244, pp. 22-27.

Camazine, S., 2003, Self-Organizing Systems. Encyclopaedia of Cognitive Science, www.cognitivescience.net, 4pp.

Cohen, J. and Stewart, I., 1995, The Possibilities of Evolution In: *The Collapse of Complexity*, (Penguin), pp. 97-133.

Cross, F.R. and Jackson, R.R., 2005, Spider Heuristics. Behavioural Processes. 69(2), pp. 125-127.

Culp, G. and Smith, A., 2001, Understanding Psychological Type to Improve Project Team Performance. *Journal of Management in Engineering*, ASCE, 17(1), 24-33.

de Bono, E., 1970, *Lateral Thinking*. (Penguin), 260pp.

Delbecq A. L., VandeVen A. H., and Gustafson D. H., 1975. *Group techniques for program planning: a guide to nominal group and Delphi processes*, Glenview, Illinois: Scott Foresman and Company.

Edmonds, D. and Eidinow, J., 2006, *Rousseau's Dog. Two Great Thinkers at War in the Age of Enlightenment*. (Faber and Faber), 405pp.

Ferguson, E.S., 1999, *Engineering and the Mind's Eye*. (The MIT Press), 5th printing, 241pp.

Findlay, C.S. and Lumsden, C.J., 1988, The Creative Mind: Toward an Evolutionary Theory of Discovery and Innovation. *Journal of Social Biology*, 11, pp. 3-55.

Gardner, H., 1993, *Creating Minds*. (Basic Books), 464pp.

Gigerenzer, G. and Goldstein, D.G., 1996, Reasoning the Fast and Frugal Way: Models of Bounded Rationality. *Psychological Review*, 103(4), pp. 650-669.

Gigerenzer, G. and Goldstein, D.G., 1999, Betting on One Good Reason. In: *Simple Heuristics that Make Us Smart*. (Eds. G. Gigerenzer, P.M. Todd and the ABC Research Group), (Oxford University Press), 75-95.

Gilhooly, K.J., 1982, *Thinking. Directed, Undirected and Creative*. (Academic Press), 178pp.

Goel, P.S. and Singh, N., 1998, Creativity and Innovation in Durable Product Development. *Computers & Industrial Engineering*, 35(1-2), pp. 5-8.

Goldstein, D.G. and Gigerenzer, G., 2002, Models of Ecological Rationality: The Recognition Heuristic. *Psychological Review*, 109(1), pp. 75-90.

Goleman, D., 1995, *Emotional Intelligence*. (Bloomsbury), 352pp.

Gould, S.J., 1991, *Bully for Brontosaurus*. (Penguin).

Holland, J.H., 1998, *Emergence from Chaos to Order*. (Oxford University Press), 258pp.

Kates, R.W., 1962, Hazard and Choice Perception in Flood Plain management. *Department of Geography Research Paper No. 78*, University of Chicago.

Klawans, H., 2000, *Strange Behavior. Tales of Evolutionary Neurology*. (W.W. Norton and Company).

Knutson, B. and Peterson, R., 2005, Neurally Reconstructing Expected Utility. Games and Economic Behaviour 52 (2005), pp. 305-315.

Koestler, A., 1964, *The Act of Creation*. (A Laurel Edition), 751pp.

Kolb, R., 1999, *Blind Watchers of the Sky*, (Oxford University Press), 338pp.

Kruger, J., Wirtz, D. and Miller, D.T., 2005, Counterfactual Thinking and the First Instinct Fallacy. *Journal of Personality and Social Psychology*, 88(5), 725-735.

McCormick, R., 2004, Issues of Learning and Knowledge in Technology Education. *International Journal of Technology and Design Education*, 14, pp. 21-44.

Miller, L., 2000, Idea Generation. *Engineering Designer*, Sept Oct 2000, pp. 26-27.

Miller, A. I., 2001, *Einstein, Picasso. Space, Time and the Beauty that Causes Havoc*. (Basic Books), 357pp.

Nielsen J.A., Zielinski B. A., Ferguson M.A., Lainhart J. E., Anderson J. S., 2013, An Evaluation of the Left-Brain vs. Right-Brain Hypothesis with Resting State Functional Connectivity Magnetic Resonance Imaging. PLoS ONE 8(8): e71275. doi:10.1371/journal.pone.0071275

New Civil Engineer, 2002, "Debate Water Supply", 12th September, https://www.newcivilengineer.com/debate-water-supply/803760.article, accessed 28th July 2017.

Perkins, D., 2000, *The Eureka Effect. The Art and Logic of Breakthrough Thinking*, (W.W. Norton and Company), 292pp.

Phillips, H., 1999, They Do It With Mirrors. *New Scientist*, Vol. 166, No. 2243, pp. 27-29

Poincaré, J.H. (1908) L'Invention Mathématique, L'Enseignement Mathématique. 10, 357-371. Translated by editor as Mathematical Invention. In: Musings of the Masters. An Anthology of Mathematical Reflections, Ayoub, R.G. (Ed.), 17-30.

Popper, K. and Eccles, J..C., 1977, *The Self and Its Brain*. (Routledge), 597pp.

Ramachandran, V.S. and Blakeslee, S., 1998, *Phantoms in the Brain*. (Fourth Estate), London, 328pp.

Reimer, T. and Katsikopoulos, K.V., 2004, The Use of Recognition in Group Decision-Making. *Cognitive Science*, 28, pp. 1009-1029.

Rose, S., 1998, Brains, Minds and the World. In: *From Brains to Consciousness, Essays on the New Sciences of the Mind*. Ed: S. Rose, (Penguin), pp. 1-17.

Singh, S., 1999, *The Code Book*. (4th Estate).

Stevens, G.A. and Burley, J., 1997, 3,000 Raw Ideas = 1 Commercial Success! *Research Technology Management*, 4093, pp. 16-27.

Svensson, N. L., 1974, *Introduction to Engineering Design*, (NSW University Press), 129pp.

Thornton, C., 2000, *Truth from Trash*. (MIT Press), 204pp.

Tversky, A. and Kahneman, D., 1973, Availability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology*, 5, pp. 207-232.

Wänke, M., Schwarz, N. and Bless, H., 1995, The Availability Heuristic Revisited: Experienced Ease of Retrieval in Mundane Frequency Estimates. *Acta Psychologica*, 89, pp. 83-90.

CHAPTER FIVE

Project Scheduling Techniques

When a large engineering project is undertaken, there are many activities that have to be coordinated, and careful planning is needed if the project is to be completed on time and within budget. Scheduling techniques to assist in the planning of large projects, including the critical path method (CPM) and the Gantt Chart, are considered here. The use of these techniques assists in identifying and scheduling activities so that the project can be completed in the minimum feasible time. The rescheduling of activities taking account of limitations on certain critical resources is also discussed.

5.1 INTRODUCTION

Project scheduling deals with the timing and sequencing of the many activities that comprise a large project. It requires a comprehensive understanding of the project and informs other project management and control processes. Project scheduling will be used at different times within a project and will become more specific as the project progresses. For example, a schedule developed after the detailed design will be more comprehensive than the schedule developed during the preliminary design of the same project. Scheduling entails identification of the activities required within the project, estimation of the duration of each activity, identification of the precedence relations between activities (i.e. which ones need to precede others) and development of an organisational network or schedule that represents this information accurately. Such an organisational network can be used to provide the following information:

- the minimum time to complete the project if all activities run on time;
- the activities that are critical to ensure that the project is finished in the minimum time;
- the earliest start time and the latest finish time for each activity, if the project is to be finished in the minimum time; and
- the amount of time by which each activity can be delayed without delaying the project as a whole.

An organisational network can also be used to examine the times when resources (human resources, cash and equipment) need to be supplied and whether limits on these resources are likely to cause delays in the project completion. Furthermore, an organisational network can be used to identify which activities should be accelerated if the project needs to be completed in a shorter time than the current estimate.

This chapter serves as an introduction to project scheduling techniques. It provides a brief overview of their historical context and explains the steps required to develop and use the resultant organisational network and/or charts. Software is available to assist project managers and engineers develop and control project schedules. Such software is essential for large and complex projects, with potentially thousands of interconnected activities. However, it remains important for an engineer to develop an understanding of the underlying concepts and approaches upon which the software has been developed. This enables engineers to adapt their use of the software to suit the project being considered.

5.2 HISTORICAL BACKGROUND

Although project planning is sometimes considered as a recent development, the construction of structures such as the great Pyramids of Egypt and the Americas, Stonehenge and the statues on Easter Island undoubtedly required considerable organisational, resource management and scheduling skills. Unfortunately, there are no records of how these projects were planned, scheduled or managed.

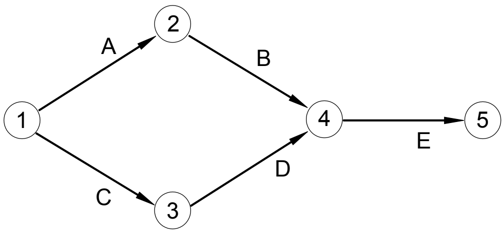
One of the first modern scheduling techniques developed to assist in project planning was the *Gantt Chart*. This was created in 1917 by Henry Laurence Gantt (1861–1919), an American mechanical engineer. The first Gantt Chart was developed for planning the building of ships during World War I. A Gantt Chart is a graph in which the horizontal axis represents time. All activities are listed down the page with each activity having a horizontal bar representing the planned timing of its completion. Gantt Charts can have varying levels of complexity and are discussed in Section 5.4.

In the late 1950s several techniques were developed to assist in the planning of complex projects. These include the critical path method (CPM) and the program evaluation and review technique (PERT). CPM was developed in the late 1950s by Morgan Walker of E.I. Du Pont and James E Kelly of Remington Rand Univac Corporation and was first used to schedule maintenance shutdowns in chemical processing plants. PERT was developed by Booz, Allen & Hamilton and the US Navy with the aim of coordinating the many thousands of contractors who were working on the Polaris missile program (Griffis and Farr, 2000; Shtub et al. 2005). Both techniques use a network of arrows and nodes to represent the activities in a project. Calculations can then be carried out to determine important information, such as that listed in Section 5.1. The basic difference between the two techniques is that CPM assumes that the durations of all activities are known, whereas PERT represents the durations of activities as random variables with optimistic, pessimistic and most likely estimates of their durations. In the 1980s, Gantt Charts were modified to include links between tasks so that they could also include certain attributes of CPM.

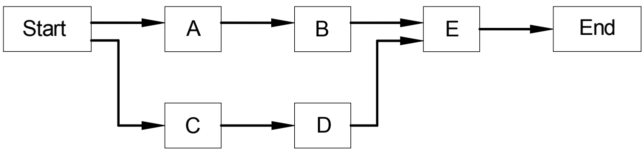
This chapter contains a description of CPM and Gantt Charts. As noted above, Gantt Charts were developed before CPM; however, they are introduced after CPM in this text to allow a more logical development of the concepts. It is these techniques, including PERT, which form the basis of modern project scheduling and control software. Progress can be tracked, assessed and updated at the task level, work package level or project level.

5.3 THE CRITICAL PATH METHOD

As already mentioned, CPM uses a network of arrows and nodes to represent all activities in a project. Two different types of notation are commonly used. These are called activity on node (AON) and activity on arrow (AOA). Either notation can be used to represent the precedence relationships between activities. To illustrate the two notations, consider a simple project consisting of five activities (A, B, C, D and E). Suppose that Activity A must be completed before Activity B can commence, Activity C must be complete before Activity D can commence and Activities B and D must be complete before Activity E can commence. An organisational network for the project using activity on arrow notation is shown in Figure 5.1 (a), while the network using activity on node notation is shown in Figure 5.1 (b).



(a) Activity on arrow notation



(b) Activity on node notation

Figure 5.1 Two types of network notation.

In the AOA notation in Figure 5.1 (a), the arrows represent the activities and the nodes are used to represent the precedence of activities. The nodes are arbitrarily numbered, so that each activity has a unique designation in terms of its start and end nodes. For example, Activity A is designated 1-2, Activity B is designated 2-4 and so on. Starting at node 1, Activities A and C have no preceding activities, so both may commence immediately. Activity B cannot commence until activity A is completed, so its start node is the end node of Activity A. For example, Activity A may represent the construction of footings for a house and Activity B the construction of the walls. In this case, Activity B cannot commence until Activity A is complete.

Similarly, Activity D has its start node (3) as the end node of Activity C. Now, why do both Activities B and D end at the same node? This is a notational convenience that also achieves a compact network, as Activity E requires them both to be completed before it can commence. There are other equally valid ways to draw this diagram that will be described later. Finally, as shown, Activity E can only start after both Activities B and D are complete and it finishes at Node 5. In this network, Node 5 also represents the completion of the project.

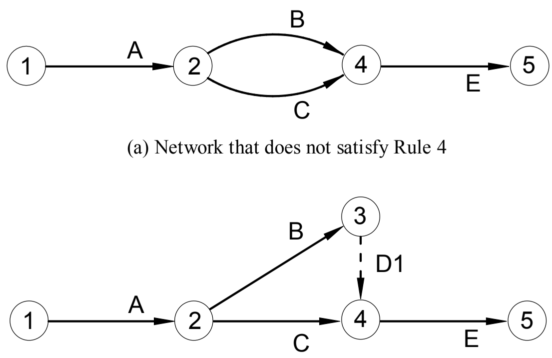
Figure 5.1 (b) shows an organisational network for the same project using AON notation. For this notation, a box represents an activity. There is also a box for the start and finish of the project. Arrows then represent the precedence relationship between activities. From the information provided above, Activities A and C do not need any activities to be completed before they can start. Therefore, they can both commence at the beginning of the project as shown in Figure 5.1 (b). Activity A must be completed before Activity B can commence, so an arrow goes from A to B. Similarly, Activity C must be completed before activity D, so an arrow runs from C to D. As both B and D must be completed before Activity E commences, they both have arrows running into Activity E. Finally the project can end once Activity E is finished.

AOA notation will be used for the remainder of this chapter, as it is more commonly used in practice.

There are four basic rules to observe when constructing AOA organisational networks. These are:

1. the network must have one starting node and one finishing node representing the start and finish of the project, respectively;
2. each activity is represented by a single arrow in the network;
3. before an activity can start, all activities leading to its starting node must be complete; and
4. there can be, at most, one arrow between any pair of nodes in the network.

Rule 4 is required to ensure that each activity is uniquely defined by its starting and ending nodes. On occasions this may require the use of *dummy activities*. These are artificial activities of zero duration that are used purely to maintain the logic of the network. An example of the use of dummy activities is shown in Figure 5.2.



(a) Network that does not satisfy Rule 4

(b) Modified network using a dummy activity

Figure 5.2 Two ways of representing the same organisation network using AOA notation.

For the project shown in Figure 5.2, Activity A must be completed before both Activities B and C can commence. Activities B and C must both be completed before Activity E can commence. Figure 5.2(a) shows both Activities B and C passing from Node 2 to Node 4. This satisfies the precedence rule (Rule 3 above), but does not satisfy Rule 4. Figure 5.2 (b) shows an alternative representation in which a dummy activity, D1, has been introduced. This network satisfies all the given rules.

Dummy activities can also be used to represent the precedence logic in a network rather than specifically to satisfy Rule 4. An example is where there are four activities that form part of the project. These are designated A, B, C and E. Activity A must be complete before Activity C can commence, but Activity E requires both Activity A and B to be complete before it can commence. An organisational network for this project is shown in Figure 5.3. Note that a dummy activity, D1, has been introduced to satisfy the precedence logic.

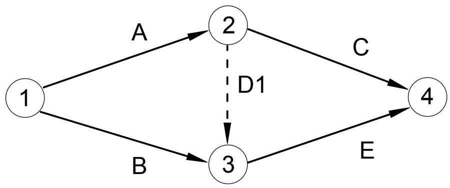


Figure 5.3 Use of a dummy activity to represent precedence logic

Application of the Critical Path Method

The CPM process for project planning involves the following steps:

1. list all activities that form part of the project;
2. estimate the time to complete each activity (called the *duration* of the activity);
3. identify the *precedence* relationships, i.e. for each activity, which other activities must be completed before it can commence;
4. construct the network;
5. analyse the network to identify the earliest start time (EST) and latest finish time (LFT) for each activity;
6. identify the critical path(s) for the network;
7. use the EST and LFT to estimate the latest start time (LST) and earliest finish time (EFT) for each activity;
8. use the EST, LST, EFT, LFT and duration for each activity to estimate its total float, free float and interfering float.

The critical path method will be demonstrated by applying it to a case study. This is the construction of a single span highway bridge over a creek. Simplified design drawings of the bridge are shown in Figure 5.4.

Step 1: List all activities that form part of the project.

An experienced engineer typically prepares the list of activities for the project. The complexity of the project and the level of control required will dictate how the list is provided. The activities can be presented as a sequential list or a hierarchical list. A hierarchical list is referred to as a *Work Breakdown Structure* (WBS) and is used to group similar tasks or activities undertaken by the same contractor. For example, all off-site activities may be grouped together and all on-site activities could also be grouped together. The hierarchical structure can provide added flexibility in relation to the level of detail presented and reported. A basic sequential list is used for this example, with the list of activities given in Table 5.1 The project involves fabricating the steel girders and handrails off-site and then transporting them for installation on site. Concrete will be produced off-site and delivered to the site ready to pour. All other activities will take place on-site.

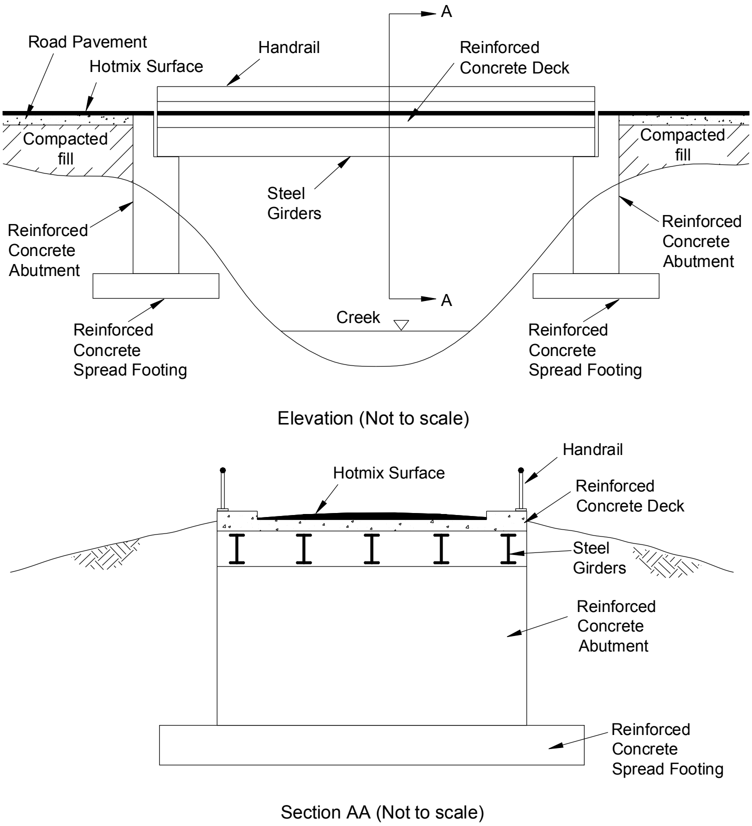


Figure 5.4 Simplified design drawings for a highway bridge.

Step 2: Estimate the duration of each activity.

The durations estimated are given in Table 5.1. These are based on the assumption that adequate equipment and human resources are available on-site to complete the activities without delays. In this simple example many simplifying assumptions have been made. For example, the construction of the foundations, abutments and embankments on each side of the creek would be considered as separate activities, as they may be scheduled to occur simultaneously or in sequence, depending on the available resources. In this case, these items are assumed to be undertaken on both sides of the creek simultaneously. A further practical consideration is the planning involved in transporting major items of equipment to the site and scheduling their usage. Such considerations have been ignored in this simple example in order to demonstrate how organisational networks are developed and analysed.

Table 5.1 Activities involved in the construction of a highway bridge.

Activity	Description	Duration (days)	Preceding Activities
A	Establish site office and transport equipment to site	3	None
B	Remove topsoil	4	A
C	Excavate foundations	3	B
D	Order reinforcement and have it delivered to the site	5	None
E	Place formwork and reinforcement for footings	2	C, D
F	Pour concrete footings	1	E
G	Cure concrete footings	7	F
H	Place formwork and reinforcement for abutments	2	G
I	Pour concrete abutments	2	H
J	Cure concrete abutments	14	I
K	Fabricate steel girders off-site and deliver to the site	21	None
L	Fabricate steel handrails off-site and deliver to the site	10	None
M	Place steel girders	3	J, K
N	Place formwork and reinforcement for concrete deck	3	M
O	Pour concrete deck	1	N
P	Cure concrete deck	14	O
Q	Install handrails	2	L, P
R	Place and compact fill on approaches	21	None
S	Construct pavement on approaches	10	R
T	Place hotmix on approaches and on the bridge deck	3	P, S
U	Replace topsoil and re-vegetate the embankments	5	R
V	Paint the handrails	3	Q
W	Clean up the site	7	T, U,V

Step 3: Identify the precedence relationships.

The relationships determined for this project are given in Table 5.1.

Step 4: Construct the network.

The construction of the network is a relatively straightforward task once steps 1-3 have been completed. However, steps 1-4 can involve several iterations as the process of drawing the network can facilitate thoughts about the activities involved and their inter-relationships and dependencies. The completed network is shown in Figure 5.5.

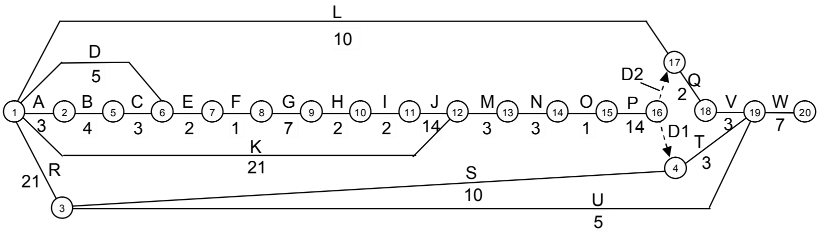


Figure 5.5 Organisational network for construction of the highway bridge.

The network satisfies the four basic rules given in Section 5.3 and includes two dummy activities. The dummy activity, D1, between Nodes 16 and 4 ensures that Activity T (place hotmix on approaches and on the bridge deck) cannot commence until Activity P (cure concrete deck) is complete. The other dummy activity, D2, between Nodes 16 and 17 ensures that Activity Q (install handrails) cannot commence until Activities P (cure concrete deck) and L (fabricate steel handrails off-site and deliver to the site) are complete. Two dummies are required in this instance because Activity Q does not depend on the completion of Activity S (construct pavement on approaches) being complete and Activity T does not depend on Activity L being complete. Some work can be undertaken on the abutments before the concrete of the foundations has reached its design strength so the duration for the curing of the concrete in the foundations has been set at 7 days.

Step 5: Analyse the network to identify the earliest start time (EST) and latest finish time (LFT) for each activity.

The earliest start time (EST) for an activity is the earliest time that it can commence assuming all preceding activities and the overall project starts on time. The latest finish time (LFT) for an activity is the latest time that it can finish without increasing the minimum completion time for the overall project. The EST and LFT are shown on Figure 5.6 with the EST shown in the left-hand box and the LFT in the right-hand box. Typically, the EST and LFT would be documented for each node. In Figure 5.6 this information is presented only at significant nodes for clarity.

The ESTs are calculated in the following way: Begin at the first node (i.e., the node that has no arrows leading into it). In this case this is Node 1. Set the EST at this node to be 0. For each subsequent node the EST is determined by examining all activities leading into the node. For each activity determine the sum of the EST

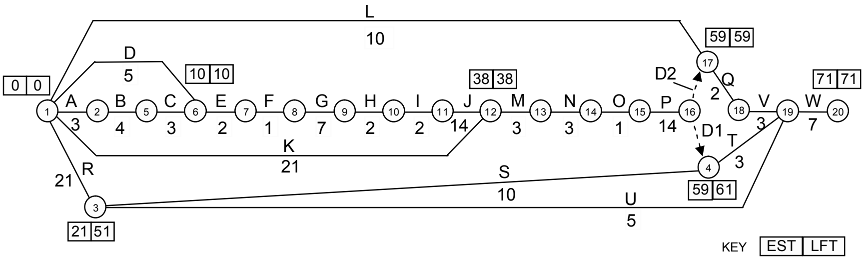


Figure 5.6 EST and LFT for activities involved in the highway bridge construction.

of its start node plus its duration. Take the largest of these values to give the EST at the new node.

For example, Node 2 has only 1 activity (A) leading into it. Clearly the EST at Node 2 is the EST at node 1 plus the duration of Activity A (3 days), thus giving an EST of 3 days at node 2. Similarly, the EST at Node 5 is the EST at Node 2 (3 days) plus the duration of Activity B (4 days) giving 7 days. Node 6 has 2 activities leading into it (C and D). The EST at Node 6 is the larger of the EST plus duration for these 2 Activities. For Activity C, its EST is 7 days and its duration is 3 days giving a total of 10 days. For Activity D, its EST is 0 days and its duration is 5 days, giving a total of 5 days. The larger of these two is 10 days which is, therefore, the EST at Node 6. This clearly must be the case, because no activities starting at Node 6 may commence until all activities feeding into this node have been completed. That is why it is the larger of the two times.

Mathematically, EST at Node j (EST_j) can be determined using the following equation:

$$EST_j = \max_i \{EST_i + D_{ij}\} \quad \text{all } i \in I_j \tag{5.1}$$

where, D_{ij} = the duration of Activity ij; and I_j is the set of all starting nodes that have activities that finish at node j.

The determination of ESTs requires the nodes to be considered in a certain order. For example, the EST at node 4 requires the EST at nodes 3 and 16 be determined first. This can be accomplished by starting at the node that has no activities leading into it and working one at a time to nodes that only have activities with known ESTs coming into them. This can be carried out by observation when working simple examples (such as this one) by hand and is programmed into software used for analysing more complex organisational networks.

Latest Finish Times (LFTs) are determined in a similar fashion by working backwards in time in the following manner: Start at the final node (the one with the largest EST. The LFT at this node is the same as the EST. Then work backwards one node at a time and considering all Activities that commence at the node. The LFT is the *smallest* of the LFT of any Activities that start at the node *minus* their duration. For example, the LFT at Node 20 equals its EST of 71 days. The LFT at Node 19 is the LFT of Activity W (71 days) minus its duration (7 days) and is therefore 64 days. The LFT at Node 4 is the LFT of Activity T (64 days) minus its duration (3 days). It is therefore 61 days.

There are two activities leading from Node 3. Therefore, the LFT at this node is the smaller of the LFT minus the duration for each of these activities. Activity S has a LFT of 61 days and a duration of 10 days and therefore a difference of 51 days. Activity U has a LFT of 64 days and a duration of 5 days and therefore a difference of 59 days. The smaller of the two (51 days) then becomes the LFT at Node 3. Mathematically, the LFT at Node *i* can be determined using the following equation:

$$LFT_i = \min_j \{LFT_j - D_{ij}\} \qquad \text{all } j \in J_i \tag{5.2}$$

where, J_i = the set of all ending nodes that have activities that commence at Node *i*.

As for the EST, the determination of the LFT involves careful selection of the sequence of nodes to be considered. A partial check on the calculations of EST and LFT is provided in that the LFT should be 0 at the first node. Although, if this value is obtained it does not necessarily mean that all values of EST and LFT are correct.

Step 6: Identify the critical path(s) for the network.

The *critical path* is the set of all activities that cannot be delayed without delaying the entire project. There is always at least one critical path from the first node to the last node. There may be more than one critical path in some cases. The minimum time to complete the project is given by the length of the critical path. This is also given by the EST and LFT at the last node.

In the example illustrated in Figure 5.6, the critical path consists of the following sequence of activities: A, B, C, E, F, G, H, I, J, M, N, O, P, Q, V, and W. The minimum time to complete the project is 71 days. It can also be seen that the EST and LFT are the same for all nodes on the critical path.

Step 7: Use the EST and LFT to estimate the latest start time (LST) and earliest finish time (EFT) for each activity.

The latest start time (LST) for an activity is the latest time that it can start without increasing the minimum completion time for the overall project. The earliest finish time (EFT) for an activity is the earliest time it can finish if it and all preceding activities start at their earliest start times.

It should be clear that the LST for each activity is simply its LFT minus its duration. Likewise the EFT for each activity is simply its EST plus its duration. LSTs unlike ESTs are not necessarily the same for all activities starting at the same node in the network. Likewise, EFTs are not necessarily the same for all activities that finish at the same node in the network. The EST, LST, EFT and LFT for all activities in the example are given in Table 5.2.

It should be clear from the definitions that the EST and LST are equal for all activities that are on the critical path. Similarly, the EFT and LFT are also equal for these activities. This is verified by the values given in Table 5.2.

Step 8: Use the EST, LST, EFT, LFT and duration for each activity to estimate its total float, free float and interfering float.

The *total float* (TF) for an activity is the amount of time that the activity can be delayed from its EST without affecting the time to complete the overall project. The total float must be zero for all activities on the critical path. For other activities, it may be determined by computing the difference between the LST and EST. Alternatively, the TF may be determined by computing the difference between the LFT and the EFT for each activity. The TFs for all activities in the example problem are shown in Table 5.2.

Table 5.2 Important times and floats for all activities.

Activity	Duration days	Critical (Y/N)	EST days	EFT days	LST days	LFT days	TF days	FF days	IF days
A	3	Y	0	3	0	3	0	0	0
B	4	Y	3	7	3	7	0	0	0
C	3	Y	7	10	7	10	0	0	0
D	5	N	0	5	5	10	5	5	0
E	2	Y	10	12	10	12	0	0	0
F	1	Y	12	13	12	13	0	0	0
G	7	Y	13	20	13	20	0	0	0
H	2	Y	20	22	20	22	0	0	0
I	2	Y	22	24	22	24	0	0	0
J	14	Y	24	38	24	38	0	0	0
K	21	N	0	21	17	38	17	17	0
L	10	N	0	10	49	59	49	49	0
M	3	Y	38	41	38	41	0	0	0
N	3	Y	41	44	41	44	0	0	0
O	1	Y	44	45	44	45	0	0	0
P	14	Y	45	59	45	59	0	0	0
Q	2	Y	59	61	59	61	0	0	0
R	21	N	0	21	30	51	30	0	30
S	10	N	21	31	51	61	30	28	2
T	3	N	59	62	61	64	2	2	0
U	5	N	21	26	59	64	38	38	0
V	3	Y	61	64	61	64	0	0	0
W	7	Y	64	71	64	71	0	0	0
D1	0	N	59	59	61	61	2	0	2
D2	0	Y	59	59	59	59	0	0	0

The *free float* (FF) for an activity is the amount of time that the activity can be delayed from its EST without affecting the starting times of subsequent activities. Once again, the FF is zero for all activities on the critical path. For non-critical activities, the FF is determined by subtracting its EST and its duration from the EST of its ending node. This can be calculated for all activities using the information contained in Figure 5.6. The FFs for all activities are given in Table 5.2. For example the FF for Activity K is based on the EST at Node 12 (38 days) and the EST at Node 1 (0 days). The FF is the difference between these (38 days) less the duration of the activity (21 days) giving 17 days.

The *interfering float* (IF) for an activity is simply the difference between its TF and FF. These values are also given in Table 5.2. The use of interfering float could delay subsequent activities, although it will not affect the time to complete the overall project.

The TF for Activity K is all FF, as the use of this time will not affect the timing of subsequent activities. On the other hand, the TF of Activity R (30 days) is all IF, as it will reduce the float available to Activities S, T and U if it is utilised by delaying Activity R.

Summary

The construction and analysis of an organisational network can provide very valuable information that is needed to manage major engineering projects. This information includes the following: The minimum total time required to complete the overall project, the activities that cannot be delayed without affecting the total time to complete the overall project, a summary of activities that need to be completed before any particular activity can be started, the earliest start, earliest finish, latest start and latest finish times for all activities and the float time for all non-critical activities (i.e., by how much time they can be delayed without affecting the time to complete the overall project or without affecting the timing of subsequent activities.

Because of its great value, it is common to use an organisational network for all major engineering projects. As previously discussed, CPM provides the basis for modern scheduling software packages.

5.4 GANTT CHARTS

As noted in Section 5.2, a Gantt Chart is an organisational network that represents the timing of activities that make up a project. In a Gantt Chart all activities are listed down the page with each activity having a horizontal bar representing the planned timing of its completion.

Figure 5.7 is a Gantt Chart for the construction of the highway bridge considered in Section 5.3. The shaded bars in Figure 5.7 represent the activities occurring over time (based on their EST), while open bars represent total floats for the corresponding activities. Each activity that has float can be rescheduled within the times represented by the open bar without delaying the overall project (assuming that all other activities run to schedule).

One advantage of a Gantt Chart compared to a critical path network is that the former shows which activities should be running at a particular time (by noting which shaded bars are intersected by a vertical line through the corresponding time). On the other hand it is not usually possible to draw the Gantt Chart without first analysing the relationships between activities using a critical path network in order to determine the EST, LFT and floats for all activities. Furthermore, it is not easy to depict the precedence relationship between activities in a Gantt Chart for complex projects. Although, in theory, vertical lines can be drawn from the end of preceding activities to subsequent activities, this can become very messy and hard

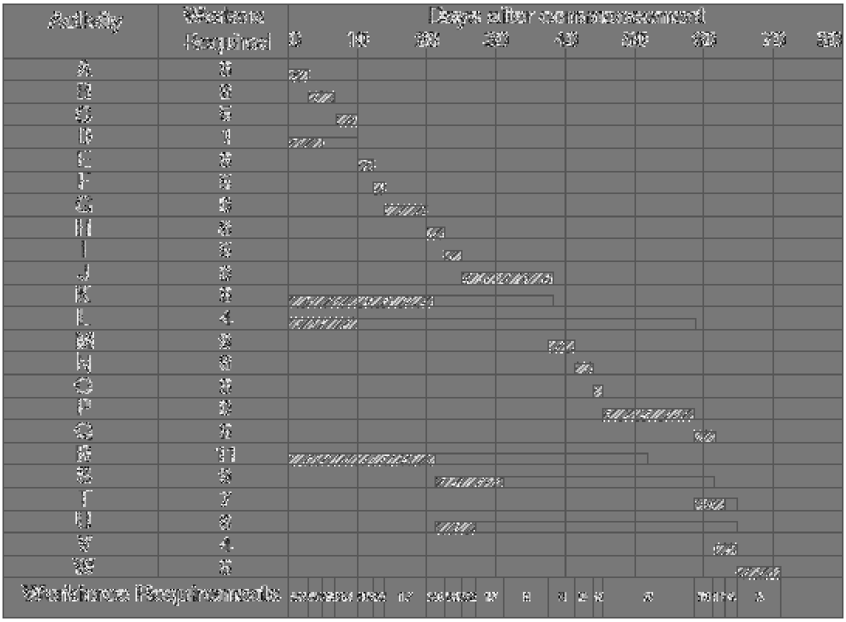


Figure 5.7 Gantt Chart for the construction of a single span highway bridge.

to follow for real projects. Nonetheless, Gantt Charts are commonly used to track the progress of projects (often in combination with critical path networks).

5.5 RESOURCE SCHEDULING

In drawing the organisational networks to date it has been assumed that the only factor that constrains the start and finish of activities is the precedence relationships, i.e., certain activities must be completed before a particular activity can commence. In practice, the availability of critical resources can constrain the timing of activities. These critical resources may be human resources such as the total workforce, skilled labour in particular areas (e.g. steel riggers or electricians) or critical items of equipment (e.g. cranes, bulldozers, graders). Project managers must schedule activities taking these critical resources into account, and, in appropriate cases acquire additional resources by purchase, lease, rent or redeployment from other projects.

A Gantt Chart can be used to indicate the allocation of critical resources on a project (or set of projects). A Gantt Chart together with a critical path network can be used to reschedule activities so that a project can be completed within the resource constraints.

Example

The example of the construction of the single highway bridge (considered earlier) will be used to demonstrate the use of a Gantt Chart to keep track of resources. The critical resource considered is the total labour force. For simplicity it is assumed that all workers on the project are interchangeable and have the same skill set. In reality, there are many different tradesmen and professionals employed on engineering projects. The principles demonstrated in this section can be used for projects with multiple resources.

Table 5.3 shows the estimated workforce requirements for the activities that make up the highway bridge project. Note that Activities G, J and P involve allowing the concrete to cure and do not require significant worker involvement.

The information given in Table 5.3 has been summarised in the bottom row of the Gantt Chart (Figure 5.7) to show the workforce requirements on each day of the project assuming that all activities start at their EST. From the bottom row in Figure 5.7 it can be seen that the workforce for the project (based on EST) starts at 28 workers and builds up to a peak of 30 workers but drops to 9 workers by day 26 and, in fact is 0 workers for some periods of the project. This is unlikely to be optimal. The company carrying out the construction would want to move its workers between projects so that all are fully employed and that no project runs overtime. A common approach is to attempt to "level" or smooth the resources required for any particular project. A trial-and-error process can be used by moving non-critical activities within their allowable times so as to smooth the workforce requirements as much as possible.

An alternative approach is to reschedule the activities so the workforce used on the project stays within a defined cap. For example, suppose that the maximum workforce available for this project is 20. It is desired to reschedule activities so that the workforce requirements stay within this limit and, if possible, the time to complete the overall project does not increase beyond the minimum 71 days. A heuristic approach to reschedule activities is presented by Meredith et al. (1985). The steps presented by Meredith et al. (1985) are as follows:

1. start with the first day and schedule all activities possible, then do the same for the second day and so on;
2. if several activities compete for the resource, then schedule first the one with the smallest float;
3. then, if possible, reschedule activities not on the critical path in order to free resources for the critical path activities.

This heuristic will be applied to the highway bridge project using the data contained within Table 5.3 and Figures 5.6 and 5.7. For simplicity it will be assumed that once an activity commences it cannot be interrupted in order to redeploy workers to other activities. The heuristic can be extended to the case where interruption of activities can occur, although there are usually inefficiencies in redeploying workers between activities due to the start-up time of activities and the need to re-familiarise the workers with the task.

Table 5.3 Workforce requirements for the activities involved in the construction of the highway bridge.

Activity	Description	Workers Required
A	Establish site office and transport equipment to site	6
B	Remove topsoil	8
C	Excavate foundations	5
D	Order reinforcement and have it delivered to the site	1
E	Place formwork and reinforcement for footings	6
F	Pour concrete footings	5
G	Cure concrete footings	0
H	Place formwork and reinforcement for abutments	8
I	Pour concrete abutments	5
J	Cure concrete abutments	0
K	Fabricate steel girders off-site and deliver to the site	6
L	Fabricate steel handrails off-site and deliver to the site	4
M	Place steel girders	9
N	Place formwork and reinforcement for concrete deck	8
O	Pour concrete deck	6
P	Cure concrete deck	0
Q	Install handrails	6
R	Place and compact fill on approaches	11
S	Construct pavement on approaches	9
T	Place hotmix on approaches and on the bridge deck	7
U	Replace topsoil and re-vegetate the embankments	8
V	Paint the handrails	4
W	Clean up the site	6

The process commences at the start of the first day. Five activities could start (Activities A, D, K, L and R). Activity A is on the critical path and so will be scheduled. The total floats for the remaining activities are given in Table 5.4.

Table 5.4 Total floats for non-critical activities that could start on day 1.

Activity	Workforce Requirement	Total Float (days)
D	1	5
K	6	17
L	4	49
R	11	30

As Activity A requires 6 workers, there are 14 available for other activities. The order of priority is Activity D, K, R and L. Activities D and K require a total of 7 workers thus leaving 7. This is not sufficient for Activity R so it will not be started at this time, but L can be, so the scheduled Activities on day 1 are A, D, K and L with a total workforce requirement of 17 workers as shown in Figure 5.8. This process has been set up in the first rows of Table 5.5 and will be repeated as we work through the time schedule of the project.

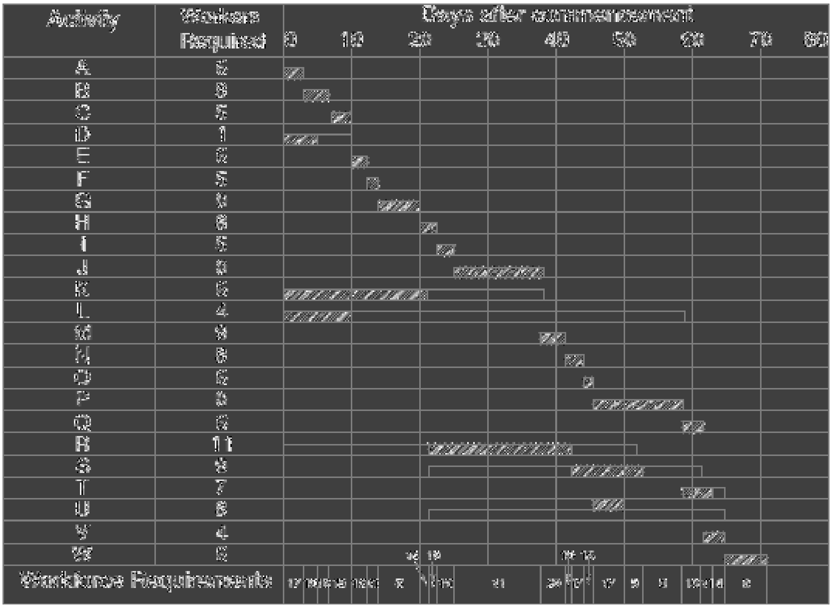


Figure 5.8 Revised Gantt Chart for construction of the highway bridge.

At the start of day 4, Activity A is completed and Activity B (a critical one) is due to start. Activities D, K and L are running and require a total of 11 workers. Activity B (requiring 8 workers) Activity R (11 workers) could be started. As Activity B is on the critical path it is scheduled to start immediately. The total workforce at this time is 19. This process has been continued for the total project. It is summarised in Table 5.5. The revised Gantt Chart is shown in Figure 5.8.

It should be noted that the project can still be completed in 71 days with a workforce of 20 workers compared with a peak of 30 workers using the EST (shown in Figure 5.7). Also note that the organisation network (Figure 5.6) needs to be considered when carrying out the scheduling, for example, Activity S cannot start until Activity R is complete.

Rather than blindly following the steps outlined above, some judgement has been applied in developing the above schedule. For example, Activity R could be scheduled to start at the end of day 13. However, it if did start then Activity H (which is on the critical path) could not start at the end of day 20 as only 3 workers would be available (Activity K and R would be active). This would delay the completion of the project as a whole. Therefore, Activity R has been delayed until the end of day 21.

Table 5.5 Scheduling of activities to match a resource constraint.

Time (end of day)	Activities in Progress	Workforce Committed (Available)	Possible New Activities	Workers Required	Total Float (days)	Activities Scheduled
0	None	0 (20)	A	6	0	A
			D	1	5	D
			K	6	17	K
			L	4	49	L
			R	11	30	
3	D, K, L	11 (9)	B	8	0	B
			R	11	27	
5	B, K, L	18 (2)	R	11	25	None
7	K, L	10 (10)	C	5	0	C
			R	11	23	
10	K	6 (14)	E	6	0	E
			R	11	20	
12	K	6 (14)	F	5	0	F
			R	11	18	
13	K	6 (14)	G	0	0	G
			R	11	17	
20	G	6 (14)	H	8	0	H
	K		R	11	10	
21	H	8 (12)	R	11	9	R
22	R	11 (9)	I	5	0	I
24	R	11 (9)	J	0	0	J
38	R	11 (9)	M	9	0	M
41	R	11 (9)	N	8	0	N
42	N	8 (12)	S	9	9	S
			U	8	17	
44	S	9 (11)	O	6	0	O
			U	8	15	
45	S	9 (11)	P	0	0	P
			U	8	14	U
50	P, S	9 (11)	None			None
52	P	0 (20)	None			None
59	None	0 (20)	Q	6	0	Q
			T	7	2	T
61	T	7 (13)	V	4	0	V
62	V	4 (16)	None			None
64	None	0 (20)	W	7	0	W

Resource Smoothing

In the above example, the peak workforce was limited to 20 workers. The question arises as to what is the minimum peak workforce requirement that does not result in the project being delayed. This can be determined using the above procedure and progressively reducing the limit on the workforce until the time to complete the overall project is increased. Careful examination of Figure 5.8 indicates that a workforce of 20 workers is required from the start of day 39 until the end of day 41. This cannot be reduced without delaying the overall project, so 20 is the minimum workforce to complete the project in minimum time. Of course, if the available workforce is less than this (say 17 workers), the above procedure can be used to schedule activities so that the project finishes in the minimum possible time (which in this case will be more than 71 days). Then a decision will need to be made as to whether it is better to accept this delay in the completion of the project or to hire additional workers.

5.6 SUMMARY

Project planning may be defined as "the process used to implement a plan and hence to achieve a designated objective, through the efficient use of available resources. It involves the regular monitoring of progress and the scheduling of activities as appropriate to achieve the objective in the required time frame." It has existed as a science since the late 19th Century.

A number of techniques exist to assist in scheduling the activities of complex engineering projects. One of the most commonly used techniques today is the critical path method (CPM). CPM can be used to estimate the total time to complete a complex engineering project as well as the earliest start time for each activity that forms part of the project. CPM can also be used to identify which are the critical activities (i.e., those whose delay will result in a delay to the overall project). The total float for each activity is defined as the amount of time that the activity can be delayed without affecting the duration of the overall project. The total floats for all activities can be determined using the CPM technique. Free float is another attribute of each activity that can be determined using the CPM. It is defined as the amount of time that the activity can be delayed without affecting the timing of subsequent activities.

An older technique for scheduling activities that is still used today is called a "Gantt Chart." It is difficult to depict all of the precedence relationships between activities on a Gantt Chart. A Gantt Chart together with a CPM diagram for a project can be used to reduce the peak workforce requirement of a project.

PROBLEMS

5.1 A swimming pool is to be constructed in the backyard of a house as shown in Figure 5.9. The pool will be 3 m by 1.5 m in plan and range in depth from 2.4 m at one end to 1.0 m at the other. It will be constructed of reinforced concrete and covered in ceramic tiles on the top and inside. Access to the site will be via a driveway, but the carport will need to be dismantled to allow access for earthmoving equipment. Assume that you are the contractor who will construct the pool from a complete set of plans and specifications. Make a list of the activities involved and draw up an organisational network that shows the interdependencies between them.

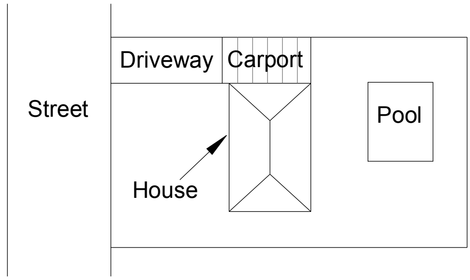


Figure 5.9 Site of a proposed swimming pool.

Table 5.6 Activities involved in constructing a steel frame building.

Activity Number	Description	Duration (days)	Preceding Activities	Number of People
1	Establish site office	3	None	3
2	Clear topsoil from site	3	1	6
3	Order steelwork	2	None	1
4	Fabricate steelwork offsite	14	3	8
5	Deliver steelwork to the site	2	1,4	4
6	Order steel reinforcement	2	None	1
7	Fabricate steel reinforcement offsite	7	6	6
8	Deliver steel reinforcement to site	1	1,7	2
9	Deliver cladding to site	1	1	2
10	Excavate for foundations and floor slab	4	2	7
11	Place reinforcement for foundations and floor slab	4	8,10	6
12	Pour concrete in foundations and floor slab	2	11	6
13	Cure concrete	7	12	0
14	Erect steelwork	8	5,13	10
15	Place cladding on steelwork	4	9,14	5
16	Complete internal fit out of the building	6	15	8
17	Clean up site	2	16	4
18	Remove site office	1	17	3

5.2 Table 5.6 contains a list of activities associated with the construction of a steel frame industrial building.

(a) Draw an organisational network for this project.
(b) Determine the earliest start time (EST) and latest finish time (LFT) for each activity.
(c) Determine the critical path and the minimum time to complete the project.
(d) If Activity 4 is delayed by 7 days, what will be the delay in the total project?

5.3 The organisational network for an engineering project using arrow notation is shown in Figure 5.10. The duration of each activity in days is indicated in the figure. Compute the following for each activity: earliest start time, latest finish time, latest start time, earliest finish time, total float, free float and interfering float. Also determine the critical path and the minimum time to complete the project.

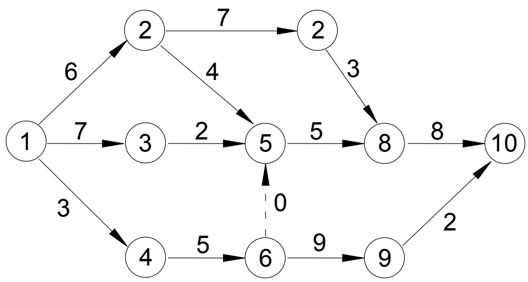


Figure 5.10 Organisational network for an engineering project.

5.4 Draw a Gantt Chart for the project described in Question 5.2 assuming that each activity commences at its earliest start time. Determine the total float, free float and interfering float for each activity.

5.5 Table 5.6 contains an estimate of the number of people required to carry out each activity.

a) Using the Gantt Chart developed in Question 5.4, estimate the peak workforce requirement for the project assuming that all activities commence at their earliest start times.
b) If the available workforce is limited to 16 people who can carry out any of the activities associated with the project, reschedule activities so that the project can be completed in the minimum possible time.

REFERENCES

Archibald, R.D. 2003, *Managing High-Technology Programs and Projects.* Third Edition (New York: John Wiley & Sons).

Burke, R. 2001, *Project Management. Planning and Control Techniques* (Chichester, UK: Wiley & Sons Ltd).

Faniran, O. 2005*, Engineering Project Management* (Pearson Education Australia).

Griffis, F.H. and Farr, J.V. 2000, *Construction Planning for Engineers* (Boston USA: McGraw-Hill).

Hyman, B. 1998, *Fundamentals of Engineering Design* (Upper Saddle River, New Jersey: Prentice –Hall, Inc.).

Meredith, D.D., Wong, K.M., Woodhead, R.W. and Wortman, R.H. 1985, *Design and Planning of Engineering Systems, Second Edition* (Englewood Cliffs, New Jersey: Prentice –Hall, Inc.).

Shtub, A., Bard, J.F. and Globerson, S. 2005, *Project Management. Processes, Methodologies and Economics* (Upper Saddle River, New Jersey: Pearson Education Inc.).

CHAPTER SIX

Management Processes and Skills

Management is about making the best use of available resources, including money, people, time and materials, in order to achieve designated goals. The management of engineering projects is a very broad and important field in itself. In this chapter we provide a broad overview of the field, and concentrating on the management processes and skills that are relevant to the work of planning and design. Of prime importance is teamwork and the full effective use of the skills of all individuals. Skills to manage oneself and also a team are discussed. The importance of personality characteristics of individuals and the roles individuals take in teams are also outlined. Skills that are required in managing day-to-day activities and provide the basis for undertaking many basic engineering activities such as planning, design and construction are also described. Managing conflict, time management and effective communication are skills for successfully undertaking an engineering project.

6.1 INTRODUCTION

This chapter addresses some of the concepts of *management* and outlines skills required by engineers to manage the resources of people and time. An engineering manager has to be capable of dealing with the processes related to the planning, design, construction and operation of engineering systems. At the same time the engineer's managerial responsibilities include the allocation of *human and financial resources* to enable tasks to be performed. Because of the nature of the engineering profession, engineers are members of teams and may take on management roles soon after graduation. Engineers therefore need to develop good communication and interpersonal skills in order to perform in teams and manage people. Throughout history many major engineering accomplishments have changed the way societies work. The design and construction of the Pyramids, the development of transport systems that crossed continents, the planning and construction of major dams such as the Hoover Dam and the Three Gorges Dam, the development of the space shuttle and of the personal computer are examples of engineering projects that are very different in their nature, but all have changed their societies. Such projects were undertaken using various leadership and management styles, varying from the one extreme of an autocratic and military nature to the *laissez faire* creative atmosphere in which personal computers were developed. While the management structures varied in these examples, the success of the project always rested on individuals with special skills who could work together in teams to achieve an effective solution using the resources at hand.

6.2 MANAGEMENT HISTORY AND PROCESS

The earliest management theories developed from the desire to manage workers and organizations more efficiently. *Engineering management* as a discipline has its origins in the industrial revolution of the 18th and 19th Centuries. During this period, cottage industries were replaced by large industrial organisations in the mining, minerals processing, manufacturing and construction sectors. In the late 19th and early 20th Centuries, engineers and managers attempted to develop a more rigorous approach to managing people, machinery and resources to achieve a defined end.

The discipline of Scientific Management commenced in the 1880s with the work of F.W. Taylor and the time and motion studies of Taylor and Gilbreth in the early 1900s. This was followed by Gantt's analysis of project management and the use of charts as outlined in Chapter 5 of this book. Henri Fayol (1841-1925), a French mining engineer, described management in terms of planning, organisation, command, coordination and control. Fayol's Administration Industrielle et Generale (published in 1916) was translated into English by Constance Storrs in 1949. From the translation, Fayol says "To manage is to forecast and plan, to organise, to command, to coordinate and to control. To foresee and provide means examining the future and drawing up the plan of action. To organise means building up the dual structure, material and human, of the undertaking. To command means maintaining activity among the personnel. To coordinate means binding together, unifying and harmonising all activity and effort. To control means seeing that everything occurs in conformity with established rule and expressed command" (Fayol, 1949). In 1913, Henry Ford adapted time and motion studies to the concept of an *assembly line* for the production of cars.

The Venetian Assembly Line

In the 1400s in the city of Venice, which was known for its naval production facilities, assembly-line techniques were used to fit out galley ships for war. A Spanish traveller in 1436 described the Venetians' process (George, 1968): "And as one enters the gate there is a great street on either hand with the sea in the middle, and on one side are windows opening out of the house of the arsenal, and the same on the other side, and out came a galley towed by a boat, and from the windows they handed out to them, from one the cordage, from another the ballistics and mortars, and so from all sides everything which was required, and when the galley had reached the end of the street all the men required were on board, together with the complement of oars, and she was equipped from end to end. In this manner there came out ten galleys, fully armed, between the hours of three and nine." This was an example of an assembly line 500 years before Henry Ford (George, 1968).

Shtub et al. (2005) defined *management* as "the art of getting things done through people." They identified the following seven functions of management: planning, organising, staffing, directing, motivating, leading and controlling. These activities are described in more detail in Table 6.1. The similarities to Fayol's list of management functions are obvious. The modern emphasis on motivating and leading staff, as distinct from commanding them, should be noted. It undoubtedly reflects a more egalitarian perspective than that which existed in the early part of the 20th Century.

Max Weber described bureaucratic management in the 1920s and emphasized that order, system and rationality in management leads to equitable treatment for employees. Throughout the 1920s and 1930s an effort was made to understand human behaviour in the workplace through studies such as that led by Elton Mayo at the Hawthorne plant of Western Electric Co. (US). The workers were tested to see how their work environment affected their production. Many other human relations researchers contributed to the field, including Mary Parker Follett, Abraham Maslow, Kurt Lewin, and Renais Likert. In the 1950s, a further group of researchers including Douglas McGregor, Chris Argyris, Frederick Herzberg, Renais Likert, and Ralph Stogdill proposed behaviour theories. There was a mix of psychologists and managerial academics working on organizational development and restructuring in the 1960s and 1970s, including people such as Fred Emery and Peter Drucker (Ullman, 1986; McShane and Travaglione, 2003).

Table 6.1 Functions of management (adapted from Shtub et al. 2005).

Function	Description
Planning	Setting goals for the organisation. Identifying a course of action or plan that will lead to the achievement of these goals.
Organising	Assigning people and resources to activities, delegating appropriate authority and establishing a structure for reporting.
Staffing	Ensuring that appropriate human resources are available for the desired activities. Ensuring that adequate training and reward structures are in place.
Directing	Orientating staff and resources towards achieving the goals of the organisation.
Motivating	Encouraging individuals to achieve their best regardless of the tasks undertaken.
Leading	Setting an example for others to follow. Encouraging the development of group pride and loyalty.
Controlling	Monitoring performance relative to the plans. Taking action when the desired outcomes are not being achieved.

The development of understanding an organization as a system began in the 1950s, with an emphasis in this approach on explaining outputs in terms of transformed inputs, taking into account interaction with the surrounding environment. Ludwig von Bertalanffy, a biologist, developed a general systems model which contributed to this thinking. Other early systems contributors included Kenneth Boulding, Richard Johnson, Fremont Kast, and James Rosenzweig. Following World War II, a contemporary School of Management evolved which included Deming and Juran, who were the proponents of *Total Quality Management* which transformed many industries in both Japan and the United States.

The Contingency School of the 1960s emphasized that management processes were dictated by the unique characteristics of each situation. It was resolved that the complexity of organizations was such that no single management strategy supplied the complete answer.

In summary, it has been seen that ideas on management have come from many different sources, with a range of tools proving to be useful for solving managerial problems. Over time it has been resolved that managerial actions must be decided on a situational basis. It is essential that organizations are considered as open systems and designed to consider individual needs for harmonious and continuing survival.

Human needs

All engineers require some management expertise to undertake engineering work. The planning, design and construction of projects, the innovative design of engineering systems, practical problem solving, managing operations of large engineering water and energy utilities and decision-making at all levels of an organisation require the management of the needs of a large number of individuals. Many engineers will work as project managers where the role is to manage people and resources committed to a project to ensure that it is completed on time and within budget. The needs of each individual in a project group affect the operation of the group, so much so, that the needs and tasks of both the individual and the group need to be considered together.

Intrinsically, under an optimistic view of life or the McGregor X style, it is assumed that individuals want to do their best (McShane and Travaglione, 2003). Research into what makes people work harder and achieve greater output was, and still is, of paramount importance to management in organizations. Maslow (1970) developed his Hierarchy of Needs, shown in Table 6.2, upon which many management researchers based their work.

Abraham Harold Maslow

Abraham Harold Maslow (1908-1970) was born in Brooklyn, New York, one of seven children born to his Jewish immigrant parents from Russia. His parents pushed him hard for academic success. He was lonely as a boy, and found refuge in books. Maslow's thinking was surprisingly original because he researched positive mental health. Maslow became the leader of the humanistic school of psychology, that emerged in the 1950s and 1960s, which gave rise to the idea that people possess inner resources for healing and growth (Public Broadcasting Service, 1998; Boeree, 2004)

"I was awfully curious to find out why I didn't go insane" - Abraham Maslow.

Maslow and researchers who followed developed models of the reasons for an individual's work behaviour. Note that these are only models and, in reality, human beings are very complex creatures who have many different aspirations that drive them. Maslow's theory suggests that, as a person satisfies one level of need, then behaviour is motivated to meet the next level of need. This theory has been built upon to explain people and organizational needs of today with various alternative variations of the five levels evolving. These models include Alderfer's

ERG (Existence Relatedness Growth) theory, Herzberg's motivator-hygiene theory and McClelland's learned needs, all of which are used to explain employee motivation (McShane and Travaglione, 2003).

Table 6.2 Abraham Maslow's hierarchy of needs (adapted from Maslow, 1970).

Basic Needs	Elements
Physiological	Food, air, water, sleep, comfort
Safety/Security	Physical, emotional security, fairness and justice, absence of threat, consistency and predictability
Social/Love and belonging	Love (both giving and receiving; separate from sex), affection, friendship
Esteem	Personal achievement, adequacy, confidence, freedom, independence, and the desire for reputation, prestige, recognition, appreciation from others
Self-Fulfillment (Actualization)	Realising a person's potential, basically satisfied people who expect the fullest creativity

6.3 WORKING IN GROUPS AND TEAMS

A *group* is a term which is vague in concept and can be any number of individuals who interact together. A *team* may be a group of people who work well together to achieve a common goal. Effective teams must have members who are willing and able to complete the task set, as well as work in a team environment. The size of effective teams has always been somewhat subjective because large tasks need large teams, but generally teams should be small enough to maintain efficient communication and coordination among the team members. Larger groups will always break into smaller informal groups to allow effective contributions from all members.

The three circles model of Adair (1983), shown in Figure 6.1, illustrates the three areas of needs in any group or team. Adair does not take credit for the model and its origin is unknown. It has been proposed (Johnson and Johnson, 2000) that one of the keys to human development has been the ability to form and work in small effective groups. Groups are central to much of human life and engineering is no exception.

Whether as a member of tutorial or practical groups as part of undergraduate study or as part of a large consulting firm working on multi-million dollar projects, engineering students and engineers will generally find themselves part of a group and in many cases will take, or be expected to take, a leadership role.

Effective groups are a force to be reckoned with; however, not all groups are effective and this can lead to significant problems. According to Johnson and Johnson (2000), an effective group

- achieves its goals (task);

- maintains good working relationships among members; and
- adapts to changing conditions in the surrounding organization, society and world.

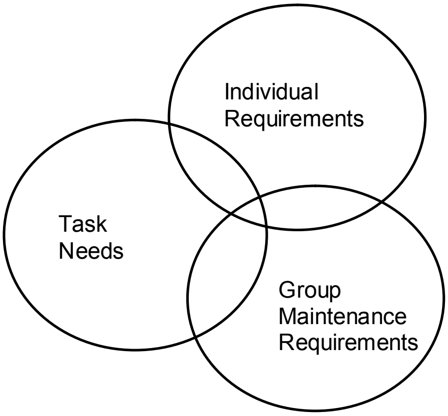


Figure 6.1 Three Circles Model (adapted from Adair, 1983).

What is missing here is attention to the needs of individuals in the group which is a factor that can make or break a group's activity. At present the trend is for more group work rather than less. A Fortune 1000 survey in 1993 found that 91% of companies had implemented some form of team working (up from 70% in 1987), and in Australia in 1991, 47% of manufacturing companies had employees in teams (up from 8% in 1988).

This means that engineers will not only work in groups and teams but in all likelihood take some form of leadership role in them as well, given that engineers are more likely to be in senior positions. For this reason it is important to know how teams work and, more importantly, to know what to do when they start to break down. Groups are often formed for a particular project and the members may not have worked together before. Members, therefore, can feel a little awkward with one another. Since the group will probably be working within a time limit, it is important for individuals to understand how groups function in order to improve the effectiveness of the team as soon as possible. This will be discussed in Section 6.6. Figure 6.2 shows some of the group activities undertaken in the problem solving processes and the interpersonal processes that involve group maintenance and self oriented behaviours.

6.4 LEADERSHIP

The question "What makes a leader?" has been posed many times. In business, sport, industry and academia there are facilitators, coaches, management academics and practitioners trying to explain a *leadership* concept that is broad enough to include leaders varying from Mohandas (Mahatma) Gandhi, Margaret

Thatcher, Elizabeth I, Joseph Stalin, Bill Gates, John F Kennedy, Martin Luther King, Mao Zedong, Nelson Mandela, John Monash and Aung San Suu Kyi. We will see later that different leaders have different personality styles.

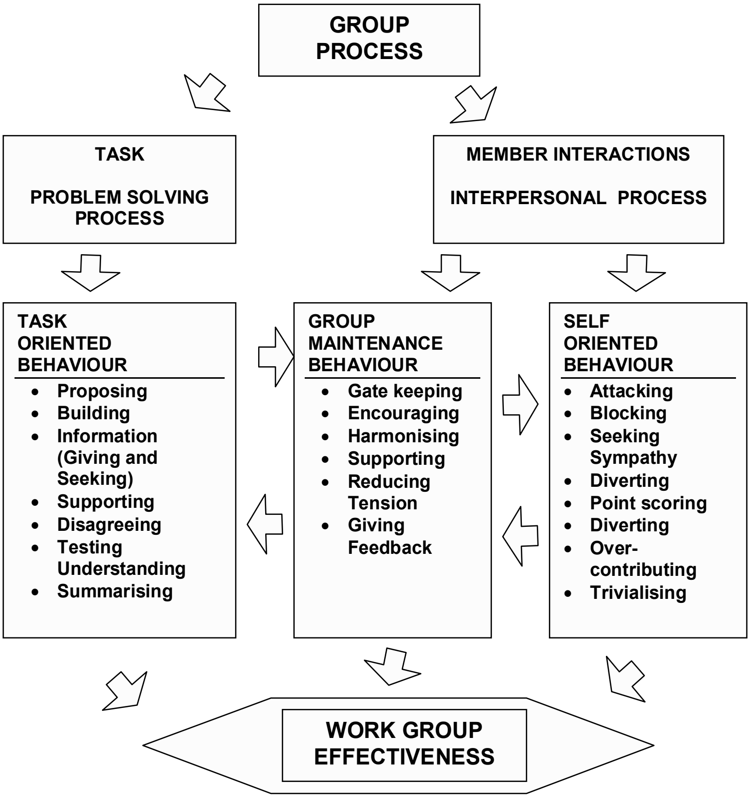


Figure 6.2 Behaviours undertaken for group effectiveness (adapted from Adair, 1983).

Developing leadership skills in the workforce is paramount to the success of organizations. Books, materials, training courses and conferences on the subject are plentiful. The challenge is to have the capacity to try out individual actions that improve the way things are done through development of an individual's style. Adair (1983) extended the simple three circles model for the functioning of groups and teams to show how leaders of teams need to function by maintaining processes in each of the three areas of task, group and self. The leader has the responsibility to ensure that all areas of need are addressed, the task is achieved, by building and maintaining the team and by ensuring the development of the individual.

Teams need a leader to succeed. An effort should be made to designate someone to play that role at or before the first meeting. Sometimes a person is appointed to the leadership role because of their place in the organisation. However, there is no need for that leader to be permanent and omnipotent. The leader needs to ensure that the task and maintenance functions of the three circles model are performed. The roles of the members within the group can be decided by natural aptitudes for different functional roles or different people can be assigned to take on functions akin to their particular styles. Belbin (1981) outlined eight roles that any member of a group prefers to operate in, as shown in Table 6.3. This was later extended to nine roles with a specialist type added.

Table 6.3 The proposed team roles of Belbin (adapted from Belbin, 1996).

Type	Characteristics
Coordinator/Chairperson	Respected, mature and good at ensuring that talents are used effectively
Shaper/Driver	Dynamic and challenging, usually leads
Plant	Very creative, the ideas person
Resource Investigator	Extrovert, good at making outside contacts and developing ideas
Monitor Evaluator	Shrewd and prudent, analytical
Implementer	Practical, loyal and task orientated
Completer/ Finisher	Meticulous and with attention to detail, also full of nervous energy
Team Worker/Supporter	Caring and very person orientated
Specialist	High technical skill and professional, operates narrowly

The Coordinator /Chairperson takes the responsibility of keeping people on track, and pays attention to group processes. This person, with the help of the Team Worker/Supporter, ensures that all members participate and notices when someone is upset. However, it is apparent that when both a Shaper and a Coordinator work in a team, one of them has to adopt a secondary preferred role.

The Resource Investigator can also act as a person who serves as liaison between the team and the rest of the world. In the context of groups at university, this person interacts with the academic supervisor and other groups. We all learn from others but mostly from our own experience. It can be said that someone with a modest amount of natural ability, who works hard at observing the member interactions of group maintenance and self oriented behaviours, as well as the task and problem solving process, will forge ahead of a person of high natural ability who relies on instincts and never addresses his or her faults. To understand the principles of leadership and to work hard at them will ensure success. As Adair (1983) stated: “Good leadership is often so silent, so self-effacing, that you are hardly aware of it, but bad leadership always shouts at you.”

Laws of Leadership

A leader needs vision, discipline and wisdom (Newman, 1994). Vision shows that the leader knows what the long-term goal is and discipline is used to ensure that energy, time and resources are directed to achieve the goal. Wisdom can be considered to be the ability to apply knowledge and experience to any situation.

The leader also shows courage when different situations demand it. How does one create courage? The answer is as Mark Twain said: "Courage is a resistance to fear, mastery of fear—not absence of fear." To make some difficult decisions requires courage and this decision-making ability is an attribute that is needed by a leader. The leader empowers others to make decisions, as the team depends on all members contributing to the process of achieving the designated goal.

Friendships and humility are developed throughout one's career. Being a good listener and trusted confidante helps in mentoring the team, in avoiding rivalry and in producing the best outcomes. In each person's career there are times when it is necessary to have someone to open up to and ideally that person is the leader. No one survives alone. In developing such relationships a leader must exercise tact and diplomacy, and show impartiality. The leader must be prepared to learn from those within the team and those outside the team to produce the best outcome. Ideally a leader will exude inspirational power and enthusiasm and encourage team members to do their best. A leader's role is to serve the team so that the team will be able to say, "We did this ourselves".

6.5 BEHAVIOURAL STYLES OF INDIVIDUALS

One important consideration when thinking about groups and teams and how individuals can be melded into a team is the personalities of all members. There are many possible ways of classifying individual personalities. One way is according to the kinds of tasks that the person likes to take on: (leader, innovator, or keeper of the peace among many others). A search in the literature and the Internet can yield many free tests to discover what preferences of behavioural style an individual has. It is important to note here that there is no 'right' personality. The main task is to recognize the differences among team members and to work with the team members, rather than have them work against the team.

We will consider just two personality models from the many typologies that are available. The first, which has the advantage of simplicity, is called DISC, which is an acronym for: Dominance, Influencing, Steadiness, and Compliance. It parallels the writings of the Greek Hippocrates who established some terms in 370 BC for four temperaments: Sanguine, Choleric, Phlegmatic, and Melancholic. Plato in 340 BC also identified and labelled 4 categories: Guardian, Artisan, Scientist and Philosopher (McShane and Travaglioine, 2003). The second model is that based on Jungian theory and adopted by Isabel Myers and Katharine Briggs for their MBTI model, which now comes under other guises such as Keirsey's Temperament Sorter (Keirsey, 1998).

DISC behavioural styles

DISC follows a theory developed by Dr W.M. Marston in 1928 and further developed in the 1940s (Cole and Tuzinski, 2003). This theory suggests that a person's preferred behavioural style falls into one of four categories. Although all four styles will usually be displayed by a single person, one style tends to describe the person's behaviour better than the others. DISC does not measure skills, experience, values, intelligence, beliefs or knowledge. DISC behavioural styles are shown in Table 6.4.

Table 6.4 DISC behavioural styles.

	Tasks and Results	Ideas and People
Direct Style	D: Direct, Dominant, Doer	I: Influencer, Inspired, Persuader
Indirect Style	C: Conscientious, Cautious, Critical	S: Steady, Supporter, Stable

The four styles are determined from whether behaviour is direct or indirect and whether behaviour is oriented towards tasks and results, or people and ideas. All styles are necessary and valuable. The descriptors help us to clarify differences in people and remove barriers to improve communication. By determining whether another person is direct or indirect, task-oriented or people-oriented, one can develop a better relationship and better communication with that person. Tests that can be used to assess behavioural style can be found on the Internet so that an allocation to one of Direct (Eagle), Influencing (Parrot), Supportive (Dove) and Conscientious (Owl) can be made.

Direct style? Faster or slower paced

Some people are direct and work quickly, take risks, are forceful, talkative and tend to make decisions quickly. The Dominant Director and Influencing Persuasive types fit into the category of direct styles, as shown in Fig. 6.3. The indirect styles are quieter, patient, cooperative, more cautious and easy-going.

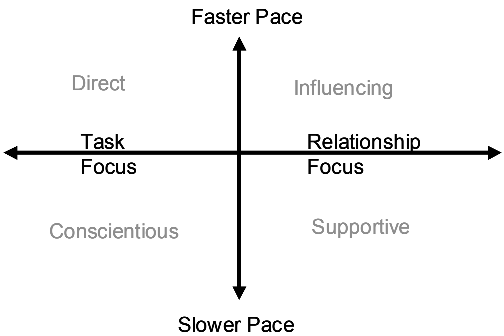


Figure 6.3 DISC Personality types and their simple characteristics.

They are good listeners and tend to take their time in making decisions and take fewer risks. Examples of indirect types include the Supporter and Conscientious styles.

People and ideas or tasks and results?

People who are relationship-oriented are generally open, appreciative and supportive, and find it easy to make friends. These people can be enthusiastic and can share their feelings. They tend to go with the flow in a relaxed manner. The Steady Supporter (S) and Influencing (I) styles stress people and ideas. They tend to promote harmony. The other types of people seem to be biased to task and results. This type likes structure, procedures, guidelines and facts. They prefer to get to the point and do not like their time wasted. They usually take a considerable time to show their warm side and tend to keep their feelings to themselves.

Note that no one operates entirely in one quadrant and we all tend to have a mixture of all the styles in different percentages. Each style has weaknesses and it is important to be aware of our own weaknesses so that we can improve in those areas. One style will be the dominant style of an individual, although in different situations other styles may well be used. Is the dominant style Direct or Indirect? Is there a preference for tasks and results, or people and ideas? Are people who have similar or different characteristics easier to get along with? Why?

Eagle (Direct)

The eagle personality is decisive and strong willed. Eagles are keen to get to the point and quickly. They are adventurous, take risks and are forceful. At work eagles are results orientated, but to be easier to work with, eagles should do more listening and improve their consideration of others. In various other typologies, this personality has been labelled as a Bear, Lion, Guardian and Controller.

Parrot (Influencing)

The parrot is enthusiastic and likes to express emotions. Parrots are talkative, optimistic and confident. In groups they are persuasive and gregarious. Parrots need to pay more attention to detail and improve their follow-through on tasks. In various other typologies, this personality has been labelled as a Monkey, Otter, Artisan and Promoter.

Dove (Supportive)

The dove is dependable, diplomatic, patient and a team player. Doves like to get results, are stable, good listeners and are sincere and patient. However, they could work at coping with change and improve their decision-making ability. In various other typologies, this personality has been labelled as a Dolphin, Golden Retriever, Philosopher and Supporter.

Owl (Conscientious)

The owl personality is orderly, cautious and reserved. Owls have high standards, are careful, analytical, diplomatic and accurate. They could, however, open up more and attempt to move out of their comfort zone if they are to make more impact on their team. In various other typologies, this personality has been labelled as a Beaver, Scientist and Analyst.

Myers-Briggs type indicator

In 1920 Carl Jung theorised that people were all different, having various degrees of four characteristic functions: Thinking, Feeling, Sensation and Intuition, and two attitudes: Introversion and Extraversion (Jung, 1923). He proposed that people's psychological orientation was formed by a dynamic mix of attitudes and functions. Isabel Myers used Jung's typology to establish a procedure for determining personality type in individuals and added the dimension of Judging and Perceiving to Jung's original typology. The Myers-Briggs Type Indicator (MBTI) was developed from decades of research by accumulating information on individuals' behaviour and attitudes of people in all spheres of work and life. The MBTI made available the theory of Jung to a much wider audience and was popularised through the 1980s and 1990s and is now embedded in many organisational management programs. A modified version has been developed by Keirsey and Bates (1984) and Keirsey (1998). Boeree (2004) has reviewed the work of Jung and has developed a questionnaire to map the elements of the MBTI as well, and this is freely available on the Internet.

The test has four scales. The first is the *Extroversion (E) – Introversion (I)***.** This scale demonstrates how we interact with others. Overall about 50 percent of the population is extroverted and 50 percent introverted.

The second scale is *Sensing (S) - Intuiting (N)* with approximately 73 percent of the general population sensing and 27 percent intuitive. This scale shows our preference for how we deal with understanding the world either through step-by-step approaches, or through using vision and insight. A sensing person tends to assimilate a series of facts in a linear fashion, while the intuiting person absorbs the same information through conceptual jumps and development of patterns from abstractions. The S types dislike solving new problems without prior experience on how to solve them. On the other hand, the N types prefer to solve new problems and dislike carrying out the same thing over and over again. Of course, people tend to share both sets of qualities to some extent.

Thinking (T) - Feeling (F) is the third scale and these functions are distributed in a proportion 40 percent thinking and 60 percent feeling, with 60 percent of men being thinking types, while 70+ percent of women are feeling types. This scale is related to how we make decisions. This is the only scale that has a gender bias.

The last scale is *Judging - Perceiving (J-P)* and was included by Myers and Briggs to help determine the superior function of an individual. Judging people tends to be more cautious, whereas Perceiving people generally tends to be more spontaneous.

The Judging – Perceiving continuum also determines the superior function. Those people who have a "J" and are extroverts have the superior function on the thinking feeling continuum. Alternatively, extroverted and "P" means that the superior function is on the sensing feeling scale. An introvert deemed to be judging will have the superior function of senser or intuiter, while an introvert with perceiving will be superior on the thinking feeling scale. It has been found that J and P types are approximately evenly distributed throughout the population.

One function from each continuum is combined to identify a type represented by four letters such as ESTJ. Table 6.5 shows the various types for the general population and Table 6.6 the various types for engineers. The tables could be rolled to make a cylinder which puts ISTP and INTP side by side and ESTJ and ENTJ side by side. We are more likely to favour the adjacent styles when not operating out of the main style. Each of the functions is a continuum, so situations will determine how much along that continuum we are likely to be.

Tables 6.5 and 6.6 suggest that engineers have a different personality profile from the general population and we find that most engineers are driven by thoughts and ideas, make decisions based on logic and facts, and tend to be organised and punctual (STJs). On the other side of the table, there are the more creative NTs who like to solve new problems, have an insight into the future and do not like routine. As with all these groupings, it is necessary to appreciate that different people see things differently, make decisions differently, interact with others differently and have different preferences.

Table 6.5 Myers-Briggs types summarised (% for general population).

Type Preferences	Sensing		Intuiting	
	Thinking	Feeling	Feeling	Thinking
Introversion Judging	ISTJ 11.6% Archivist	ISFJ 13.8% Devoted Carer	INFJ 1.5% Counsellor	INTJ 2.1% Builder
Introversion Perceiving	ISTP 5.4% Artisan	ISFP 8.8% Reticent artists	INFP 4.4% Idealist	INTP 3.3% Analyst
Extraversion Perceiving	ESTP 4.3% Negotiator	ESFP 8.5% Performer	ENFP 8.1% Enthusiast	ENTP 3.2% Pragmatic Politician
Extraversion Judging	ESTJ 8.7% Executive	ESFJ 12.3% Loyalist	ENFJ 2.5% Empathic leader	ENTJ 1.8% Visionary Commander

The % of each type comes from Myers et al. (1998)

Table 6.6 Myers-Briggs types summarised (% for engineers) (after Culp and Smith, 2001).

Type Preferences	Sensing		Intuiting	
	Thinking	Feeling	Feeling	Thinking
Introversion Judging	ISTJ 23%	ISFJ 5%	INFJ 2%	INTJ 14%
Introversion Perceiving	ISTP 6%	ISFP 2%	INFP 5%	INTP 6%
Extraversion Perceiving	ESTP 5%	ESFP 1%	ENFP 4%	ENTP 5%
Extraversion Judging	ESTJ 8%	ESFJ 4%	ENFJ 2%	ENTJ 7%

A fuller description of the Myers-Briggs types and characteristics are given in Appendix 6A and Appendix 6B.

Diversity – an important dimension of a team

"We know that a group of intelligent, motivated men and women of many different backgrounds and experiences makes an ideal engineering team. These individuals will see the world through different eyes and bring unique perspectives to the engineering task at hand. Their diversity will yield a diversity of solutions, which ultimately leads to the best solution." – *Dean of Engineering, University of Illinois at Urbana-Champaign (quoted in Daniel, 2002).*

When people talk of diversity, it is common to classify differences into a number of categories. These are

- professional - level of expertise, career stage;
- demographic - gender, age, nationality, language;
- psychodynamic - attitudes, personality type, sexual orientation;
- physiological - energy level, health; and
- world view - values, tendency to prejudice, tendency to bias, ethics stance.

As an example, the Adelaide Car Component Company had a workforce with the following statistics:

- 1160 employees;
- 58 managers, 161 professionals, 80 sales and service, 861 on shop floor;
- 71 percent male, 29 percent female; and
- 52 ethnic groups.

One would expect diversity based on professional, demographic, psychodynamic, physiological and world view – in fact over all categories. There is a tremendous potential here for great things (and for disasters!). A characteristic of teams with diversity is that they are susceptible to splitting into subgroups along gender, ethnic or other dimensions (Lau and Murnighan, 1998), but they do have the capability of having a great synergy to get better solutions to problems.

6.6 GROUP AND TEAM DEVELOPMENT

Many common-sense strategies for working in groups are now discussed and, if used in conjunction with knowledge of the stages of *team development*, can assist in moving the team forward.

Getting to know other group members

This appears to be elementary, but many groups never get to know the other team members. Generally, different team members have very different values, motivations, abilities and personalities. The first thing to do as a group is to make sure everybody introduces themselves. Make sure everyone has written down all other members' names. Suggest that everyone uses each person's name once at the initial meeting. Many people will immediately forget names if heard once or never quite hear them the first time. Ensure that everyone gets each other's name (and do not be afraid to ask for the spelling). Following this, get to know what they like doing, where they are from, show interest and generally find out about them. Some members will be totally involved and others will be apathetic if they are allowed to be. Try not to let people remove themselves from the group because they then become dead weight, leading to frustration and resentment in some group members.

Something that is hard to accept is that when team members do not participate it is not entirely their fault. It is also the fault of the leader and all other team members. One cannot be responsible for all team members all the time but some simple skills will enable fuller participation by all.

Our style and approach is adapted to the situation, depending on whom we are dealing with; from being patient to being very direct. The key to our understanding is knowing those we are dealing with, and learning to change our individual style so that better understanding is achieved. While each person has a dominant style of interaction it can be advantageous to understand each others' styles and to adapt when necessary. Keep a watch on the process and practise this skill.

From the MBTI we know that approximately 50 percent of people are introverted. These people find it against their nature to immediately talk in a group situation as they find it stressful. They tend to think things through before talking. Some people consider others' feelings and will not tell them when they have made mistakes. In a group with many extroverts, the introverted person will find it difficult to participate. To speak up to gain attention is not in his or her nature. It is the extrovert's responsibility to make an effort to include the introvert, to not dominate the conversation with them, and to not take the floor away from them. If the team wants to succeed, the team must actively manage the process of inclusion of others.

Stages in team development

When teams come together to undertake a specific task, it has been found that they tend to go through a series of quite well-defined stages. In 1965, Tuckman developed a model that had four stages of forming, norming, storming and performing. Some stages are very productive, some less so. It is important to realise that this is standard behaviour and that, if one is part of a group, these are the different stages that should be expected to happen and need to be worked through. Two extra stages of pre-group and adjourning can be added, as shown in Figure 6.4 and described below.

Stage 1 – Pre-group

At this stage the task is usually broadly defined, the group is undefined, the resources are undefined and there might appear to be little that anyone can do. However, there is. Members can decide on commitment at this early stage, and show it by, for example, arriving on time, or behaving in a way that demonstrates their willingness to be part of the group.

Stage 2 – Forming

As the groups are formed, inclusion or exclusion is of paramount importance as members try to work out their place in the group. There may be superficial conversation and people orientate to each other and the task. It is suggested that even at this early stage there will be a search for direction and people will be looking for someone to provide strong leadership. The culture of the group is established during this stage. When teams are beginning, each member considers his or her identity within the group. Identity can be considered to be a combination of personality, behaviour, competencies, and position in the social structure of the group (Borgatti, 2002). Certain members will fight for dominance, others will like to be seen as being smart, others will play a comedian role, and some just want to be liked. Knowing Belbin's roles helps a team through this stage by allowing the preferred styles of individuals to be incorporated in planning roles for the group by the leader.

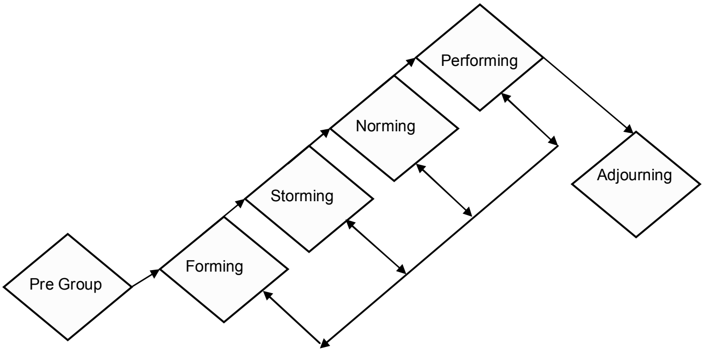


Figure 6.4 Stages of group development.

Stage 3 – Storming

Once the initial nerves are overcome, it is time for people to be confident enough to cause trouble! There may be the first conflict with personal agendas revealed. The interactions between members may be quite uneven and there is the tendency to rebel against the leader. The interactions allow a pecking order to be established. It is important to resolve conflict, otherwise apathy can set in, and doubts may surface about the ability to cope with the task.

Stage 4 – Norming

Following the early disruptions, there may emerge a sense of renewed hope. Members can become more tolerant of each other and cohesion increases, and harmony becomes important. The roles of the members are established and they can become less dependent on the leader, although the leader is still necessary.

Stage 5 – Performing

The group then enters its most productive phase. Members 'get on with the job', although minor problems may still occur. The role of the leader has changed and he or she is now seen as a peer and resource person. The roles have all been determined and there is consolidation of the status hierarchy.

Stage 6 – Adjourning

The final stage of the team is its disbandment. As things wind down, there is time to check that the goals have been achieved and to try and cope with the end of the group. Many will make plans for future meetings and there may even be some sadness. Most importantly, there will often be a level of excitement based on what was achieved.

As teams evolve, if new members join the group, there is a tendency to revert to an earlier stage of development. This can be seen in Figure 6.4 where a number of cycles are generally needed, especially if all members do not attend meetings or carry out the work allocated to them. It is important that teams assess their performance from time to time. Most teams start out well, and then drift away from their original goals and eventually fall apart. This is less likely to happen if, from time to time, the team facilitator or leader asks everyone how they are feeling about the team, and does a public check of the performance of the team against the mission/vision statement.

6.7 TEAM MEETING SKILLS

The skills required by an engineer are many and varied. These include the management of time and people. Managing people requires the manager to know his or her inner workings. Once an individual knows how others see him or her and how he or she interact with others, he or she will become better in leading groups and teams. Managing the resources that are available is paramount for any engineer carrying out projects. The ability to facilitate meetings and to manage one's own time are two skills that are part of being an efficient manager and leader. Meetings can be useless if there is no control of the discussion.

Effective Meetings

To have effective *meetings*, it is important to set an agenda, start and finishing times, the location for the meeting and ensure the chairperson keeps control of the process and time schedule. One simple thing that helps a lot is having an agenda. Brainstorming sessions have their own format. Meetings should have a designated outcome but if the meeting is just for a one-way distribution of information, this can be done prior to the meeting and the *agenda* set for discussion. Some simple processes are

A Attendees assign a chairperson who should encourage active participation to have the best possible decision-making process. Remember the meeting is for the benefit of all attendees. Listen to other points of view and do not be afraid to offer your own opinions.

G Group dynamics are important to allow all members to have a say. Ensure quieter members contribute, as they often have excellent ideas. Be firm with dominators who talk all the time, not allowing others to contribute. Remember the goals of the meeting.

E Expected outcomes are achieved by explaining the purpose of the meeting, exploring ideas and allowing all members to contribute to the discussion. Each member should feel that a contribution is expected and will be valued.

N Note-taking for an accurate record of the meeting is essential. This is a skill to practise and apply in meetings. The more that one does note-taking, the more adept one becomes at it. Rotating the position of the official recorder of the meeting is important for all members to get experience.

D Designate an action against each item. Decide "who is doing what" by what date so that everyone is clear on what is required. Create an action list.

A Announce and advise what is on the designated **Action** list, by circulating the minutes of the meeting as soon after the meeting is completed and set the next agenda, including time and place! The first item on any agenda should be a “status check,” which is where the facilitator asks each person how things are going and whether actions have been undertaken and, more importantly, whether there is anything that needs to be discussed.

As an individual attending a meeting, check the contribution that can be made by asking: Should this be a brainstorming session? See guidelines on brainstorming and creativity in Chapter 4. What preparation is required for the meeting? Has the work that has been allocated on the action list been completed? Are there other members who were not at the meeting who need to be involved?

Some goals that successful chairpersons try to undertake are to establish the vision in conjunction with the group; ensure the team focus on the task; actively pursue participation from all members of the team; protect individuals from direct personal attack and establish conflict resolution strategies; suggest alternative processes when the team is stalled; and summarize and clarify the team's decisions. The chairperson accomplishes these goals by doing the following:

- Stays neutral and ensures that good seating arrangements are used (e.g., best in a circle);
- Keeps the meeting on time, even if it is going well (or people will try to avoid coming next time);
- Expresses out loud what seems to be happening (e.g., “George and Paul (in a side conversation) can you please give your opinion on the issue at hand”; "nobody seems to be saying much since Belinda suggested ... ");
- Ensure conflicts, snide remarks and put-downs are addressed immediately to help foster a team spirit even referring the perpetrators to reading material on group behaviours; and
- after a person has not contributed for a while, ask for his or her opinion.

Most importantly, before it is decided that a meeting needs to take place, examine if there is a better method to achieve a similar outcome. Email, telephone and/or video-conferencing may be more efficient alternatives.

Handling conflict in teams

Too much conflict may be a bad thing, but some is considered good because it may lead to more ideas being considered. The important issue in conflict is to ensure it is controlled and does not get out of hand. Correct decisions need to be taken if a wrong engineering decision will lead to a bad outcome, even if this is bad for team morale. This is not as easy as it sounds because there are many sources of conflict, many different situations in which it occurs and therefore many different ways of handling it. The five ways of handling conflict shown in Figure 6.5 depend on the behaviour and attitude of the person causing the conflict, and the managerial style of the person attempting to resolve it. These will now be discussed in a little more detail.

Coercion

With coercion the leader essentially imposes his or her will on the person who has caused the conflict. It is a power-orientated approach and is useful if quick action is needed, or for unpopular decisions. While it works well in these situations, it may cause problems later because, although quashed at the moment, the conflict may not be resolved and the losers may become angry because there is competition for their own concerns. At this point, a negotiated resolution becomes less likely.

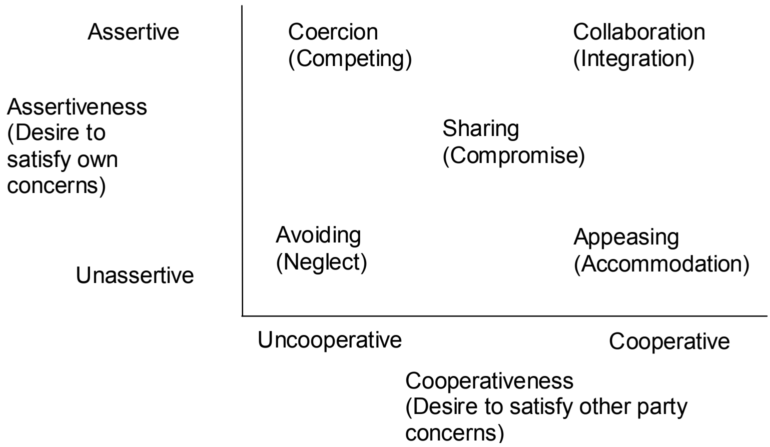


Figure 6.5 Conflict types and resolution strategies (adapted from Thomas, 1976).

Accommodation

In accommodation, an unassertive leader works with a cooperative team member. Both neglect their own needs to suit others (to a certain extent). It is a useful conflict strategy if a team needs to preserve harmony, or when one of the two realises he or she is wrong. It has the advantage in that it allows others to learn by their mistakes, and is useful in building credit for later. On the down side, because they have given way, the feeling of self-worth may be diminished, and winners may press for further domination.

Avoidance

Conflict from an uncooperative team member can also be handled by avoiding the issues altogether. Avoidance does not address the conflict directly but sidesteps and attempts to postpone any resolution. As a strategy, avoidance is useful when the issue is trivial, or the leader's power is low. It may also be applicable when the potential damage is greater than the benefits of resolution, or if there is a need to allow people to cool off. On the other hand, it generally gives low satisfaction and may even increase the opponent's frustration.

Compromise

Compromise seeks a mutually acceptable solution to the problem, or at least a temporary solution. It can be used where two opponents have equal power, or to achieve a temporary settlement in complex matters, or when goals are only moderately important. The main disadvantages with it are that the conflict is thought likely to re-emerge, and both sides may be dissatisfied with the outcome.

Collaboration

If there is time, working together using successful negotiating behaviours can bring an excellent result to conflict. Collaboration can be used to find a solution when both parties' concerns are too important to ignore, to maximise commitment, and when it is important to learn from others. It can, however, take time and energy to work through it.

Elephants and ants analogy for choosing priorities

If ants were on the menu for dinner, how many would be needed for a feed? A great many! A lot of hunting would be required. On the other hand, if an elephant was on the menu for dinner how many would be needed? One would satisfy everyone's appetite.

The activities that we perform in our daily lives can be related to the elephants and ants analogy. When work is about killing ants there is a confusion of activity with accomplishment. What kind of sustenance would be obtained from a large number of ants? Choosing to carry out easy small tasks is done because they can be done quickly, so as to achieve the illusion that a lot has been accomplished, when in reality productivity has been poor. One falls further and further behind because the elephant hunting has been overlooked.

Elephant hunting or focussing on high payoff activities leads to productivity in the longer term. If one is busy stomping ants all day long, then the elephants will be totally ignored.

If the situation exists of constantly being busy, spending time answering unimportant emails or dealing with minor interruptions, then there is a need to change from these "ant stomping tasks" to work on higher payoff activities. The ants still need to be dealt with but at a lower priority and with less time allocation. Try to spend less time with ants each day. **Don't Confuse Activity with Accomplishment**. (Adapted from Vance, 1993)

6.8 PERSONAL TIME MANAGEMENT SKILLS

Vilfredo Pareto (1848-1923), an economist, made the well-known observation in 1906 that 20 percent of the population owned 80 percent of the property in Italy which was later generalised into the so-called Pareto Principle (for many phenomena, 80 percent of consequences stem from 20 percent of the causes). This principle was further generalised in 1937 by Joseph Juran (1904-2008), the quality management guru of the 20th Century. Applied to time management, this translates to 20 percent of the time being expended on "the vital few" situations or problems to account for 80 percent of the results.

People can waste much time if they do not have direction or know how to recognise what is important. Time is irreversible and can never be regained. Engineering managers and leaders have many demands in any one day, and therefore time is required to be managed to ensure effective use of it.

Vance (1993) emphasises the need to focus on important activities rather than activities which just keep everyone busy and being reactive to distractions or so-called urgent needs. This is also the basic premise of Covey (1989) in "The Seven Habits of Highly Effective People". The quick way to improve work productivity, family and social activities is to identify the 20 percent of the activities that will reap 80 percent of the benefits. It is necessary to reduce the time allocated to activities which give us small benefits, the so-called "ants". Table 6.7 lists some possible time wasters suggested by Richards (1982).

Table 6.7 List of time wasters.

Waiting	Travel
Attempting too much at once	Unrealistic time objectives
Lacking objective/priorities and planning	Telephone interruptions/conversations
Inability to say "No"	Indecision
Jumping in	Bureaucratic processes and form filling
Lack of self-discipline	Television watching/DVD
Internet/computer games	Idle conversations
Failure to listen	Poor organisation/filing system
Duplicating effort from not finishing	Unanticipated interruptions or visitors
Communication problems with others	Ineffective meetings
Micromanagement/too much supervision	Environment with visual/noise distractions
Confused responsibility	Doing urgent rather than important tasks
Other……	Other……

Source: (adapted from *"The Executive's Guide to Modern Management Skills"* by Richards, 1982).

Procrastination

One must overcome the state of mind that is procrastination, which is the delaying of planned activities. What we have to do in most cases is just get started—make a plan and do it. Do not make excuses for putting off doing activities or even not planning to do them at all. Analyse how time is spent, for example, by jotting down what is done for a week for each quarter of an hour, and then implement a few methods to eliminate the identified bad habits that will gain the most time.

Rank these according to level of importance. Do this so that there are four groups of three activities. Once this has been done, undertake actions to operate on the top three as a priority, to help manage time.

First of all, we need to allocate time to important tasks rather than urgent but unimportant tasks. This can be achieved by developing a plan with priorities which is updated along with a daily to-do list and weekly and monthly goals. To accomplish good time management needs self-discipline. This will require changing old habits and developing new ones. One has to be persistent in using the time saving hints developed in the time management plan. Once plans are put in place and gains in time have been achieved, the next areas for review can be considered.

A Simple Time Management Plan

Effective time management at work is crucial to accomplishing tasks and goals, so that there is personal time left for personal enjoyment. Organization and task focus are the primary requirements. There needs to be persistence and self-discipline, an awareness of overcoming procrastination and a long-term vision.

- **Plan activities**. Take time to plan activities. It is important to allow time to plan wisely. Establish priorities for the day, the week, short-term, mid-term, and for the long-term. **Dividing large tasks** into a series of small manageable tasks will let the large task be easily accomplished. Each small task should have a deadline.
- **Set time aside to do high priority tasks.** Maintain accurate calendars; abide by them and adjust priorities as a result of new tasks. This can be done by using checklists and to-do lists and in some cases doing the most difficult task first.
- **Do one task at a time, if possible.** When starting a task, try to complete it before starting another task-this can be done by allocating enough time for the task.
- **Set start and stop times for activities.** This will need estimates, but these will improve with practice. Challenge the theory, "Work expands to fill the allotted time." Therefore, establish deadlines for all tasks. Ensure meetings have a specified purpose, have a time limit, and include only essential people. Do not waste other people's time.
- **Do it, delegate it, and dump it.** (Vance, 1993) Do not put unneeded effort into tasks and activities which do not require perfection. Handle correspondence quickly with short letters and memos. Save time for other activities. Learn to know when to stop a task, using the Pareto Principle. Delegate as much as possible and empower subordinates. Throw unneeded things away.
- **Learn to say no.** By making the mistake of saying yes to too many things, priorities are then decided by others.
- **Avoid committing to unimportant activities, no matter when they are.** Ask what can be planned for this time slot, such as a holiday, a weekend hiking or camping or a fun weekend with family.

- **Get Started.** The classic time waster is avoiding starting a project. The most important item is to do it now. “A journey of a thousand miles starts with one step.”-Lao Tzu. The ability to start work quickly results in achievement and satisfaction. Five minutes now can achieve the start and the journey is begun. Give rewards for finishing tasks, as this tends to stop procrastination and helps starting.
- **Develop a routine** to do certain tasks like answering emails when there is time but not the energy to do other tasks. Set aside time for reflection.

Time Management—Long-Term Goals

Keep long-term goals in sight. Have checklists with items such as: "department meeting at 2:00" and "ring so and so, write letter/memo on…,” but ensure time is put aside for the relationship tasks. Other examples include:

- meet with staff on a regular basis both formally (interviews) and informally (morning coffee at least once a week);
- develop a plan for the organisation to use only recycled paper;
- enroll to study Italian because in 4 years I want to be fluent because (insert your goal); and
- exercise each day (e.g., at least 15 minutes walk no matter what the weather).

Planning

As a planner, memorising this poem of Rudyard Kipling (1865-1936) will assist in developing a focus on the crucial facets of any plan. In planning, responses need to be made to each of these questions. (Adair, 1986)

I keep six honest serving-men
  (They taught me all I knew);
Their names are What and Why and When
  And How and Where and Who.
I send them over land and sea,
  I send them east and west;
But after they have worked for me,
  I give them all a rest.

(From *The Elephant’s Child*)

Try not to get caught up in short-term demands which put pressure on things that are more important that one should be doing but cannot find the time for. Approximately 30 percent of items on the “to-do list” should be long-range items that would normally be put aside for when there is enough time! These long-term goals are very much embedded in students earning a degree but also goals for fitness, exercise and sports, relaxation and enjoyment need to be developed. Spontaneity can be fun, but make sure that there is time available for the things which will enable long-term goals to be accomplished.

6.9 SUMMARY

The resources of time, money, materials and people are used to make things happen. In this chapter the need to manage people and time has been addressed. First of all, learning how to manage oneself can make it easier to manage others. To manage others it is imperative that one manages oneself in a way that promotes the desire in others to do well. In this way teams and groups will achieve desired outcomes and some fun may be had along the way.

As an engineer, you will not only work in groups and teams but in all likelihood take some form of leadership role as well, as many engineers rise to senior positions. The importance of knowing how these roles work and, more importantly, to know what to do when they start to break down cannot be stressed enough.

It has been shown that effective time management is crucial to accomplishing tasks and goals as well as providing time for personal activities.

PROBLEMS

6.1 Write down five groups/teams that you have been part of in the last year and determine if they were really a group or a team. Is it possible to classify them easily or are there elements of both in some of them?

6.2 Break into groups, as designated by lecturer. These can be groups of different sizes ranging from 3 to more than 10. Discuss "What are the important elements of having a successful meeting?" Prepare a set of hints for running meetings.

6.3 Use the results of the personality test (http://www.keirsey.com/) to assess the attributes of the people in an engineering group project that you are involved in. What are the advantages of doing this in a formal way?

6.4 Working with an engineering project group of which you are/were a member, list the range of diversity that the members exhibit. Take some care with this because, due to cultural differences, some may not be happy to discuss various aspects of themselves. You may find sexual orientation is one area that, quite reasonably, many are keen to avoid discussing, and this should be taken into account.

6.5 In groups where you are involved through the year, make a careful observation of conflict and how it is resolved. Are there people who seem to always want to use the same resolution strategy no matter what the situation? Does this work?

6.6 Irrespective of the nature of the conflict or the feelings of the leader, what strategies of conflict resolution might be appropriate for a problem that arises on the day before a report is due?

6.7 Choose three of the leaders listed in Section 6.4, and find out what their main attributes of leadership were against the 10 leadership laws.

6.8 Activity for group review: assess what might be going wrong in your team and think how to remedy it.

How is the group functioning?	**Score from 1 to 7: 1 is agree, 4 is sometimes, 7 is disagree**
Group clarifying what the task or objective is	
Group continuously checking on progress against timeline and seeks reasons for non-compliance	
Group clarifying or recording what has been decided	
Group clarifying who is going to do what	
Group clarifying what has to be done by when	
Group establishing procedures for handling meetings	
Group keeping to agreed procedures	
All members listening to each other	
Not allowing individuals to dominate and others to withdraw	
Not compromising individual needs for the sake of the team	
Group recognising the feelings of members of the team	
Members contributing equally to the progress of the team	

Sum the 12 scores. A score of 12 would be perfect, 12 to 29 shows team is working well, 24 to 50 requires more effort, 50 to 84 would be disastrous and requires a meeting to change attitudes in the group.

REFERENCES

Adair, John, 1983, *Effective Leadership*, Aldershot: Gower, 228p.

Adair, John, 1986, *Effective Team Building*, Aldershot: Gower, 212p.

Ancona, Deborah, Kochan Tom, Scully Maureen, Van Maanen, John and Westney Eleanor, 1999, *Managing for the Future: Organizational Behaviour and Processes.* 2nd ed. 3rd ed., Cengage Learning Inc.

Belbin, M., 1981, *Management Teams*, *Why they Succeed or Fail*, Heinemann: London.

Belbin, M., 1996, *Team Roles at Work*, Butterworth.

Berens, Linda V., Ernst, Linda K. and Smith, Melissa A., 2004, *Quick Guide to the 16 Personality Types and Teams: Applying Team Essentials to Create Effective Teams,* Telos Publications.

Boeree, C. G., 2004, *Personality Theories*- Abraham Maslow, Retrieved 16/8/2005. http://www.ship.edu / ~cgboeree/maslow.html.

Boeree, C. G., 2004, Personality Theories- Carl Jung, Retrieved 16/8/2005, http://www.ship.edu/%7Ecgboeree/jung.html.

Borgatti, Stephen P., 2002, Introduction to Organizational Behaviour, last revised April 2002, August 2005. http://www.analytictech.com/mb021/mbtidim.htm, Retrieved.

Bynner, Witter, 1944, Tao Te Ching, Translation, The Way of Life, http://home.switchboard.com/ TaoTeChing.

Cleland, David I. and Kocaoglu Dundar F., 1980, *Engineering Management*, McGraw-Hill.

Cole, Pamela and Tuzinski, Kathleen, 2003, *DiSC* Indra Research Report, Inscape publishing, 31pp.

Covey, S.R., 1989, *The Seven Habits of Highly Effective People*, Simon & Schuster: NY.

Culp, G. and Smith, A., 2001, Understanding psychological type to improve project team performance. *Journal of Management in Engineering, ASCE, 17*(1), 24-33.

Daniel, D.E., 2002, Letter from the Dean. *Engineering Outlook.* Vol. 42, (1&2), p2.

Duch, Barb, 2000, *Working in Groups,* University of Delaware, USA, Retrieved Aug 2005. http://www.physics.udel.edu/~watson/scen103/colloq2000/ workingingroups.html.

Fayol Henri, 1949, *General and Industrial Management*, translated from the French edition (Dunod) by Constance Storrs, Pitman, London.

George, Claude S., 1968, *The History of Management Thought,* Englewood Cliffs, N.J.: Prentice Hall.

Hofstadter, Douglas, 1979, *Gödel, Escher, Bach: an Eternal Golden Braid* (commonly *GEB*), Basic Books.

Johnson, D.W. and Johnson, F.P., 2000, *Joining Together. Group Theory and Group Skills*. Seventh Edition, Allyn and Bacon, 643pp.

Jung, C.G., 1923, *Psychological Types*, Routledge and Kegan Paul, London.

Keirsey, David and Bates, Marilyn, 1984, *Please Understand Me, An Essay on Temperament Styles*, Prometheus Nemesis Book Company.

Keirsey David, 1998, *Please Understand Me II*, Prometheus Nemesis Book Company. Retrieved 03/04/2017. http://www.keirsey.com/.

Lau, D. C. and Murnighan, J. K., 1998, "Demographic Diversity and Faultlines: The compositional Dynamics of Organisational Groups", *Academy of Management Review*, 23, April, pp. 325-340.

Maslow, Abraham H., 1970, *Motivation and Personality*, Harper & Row.

McShane, Steven and Travaglione, Tony, 2003, *Organisational Behaviour on the Pacific Rim*, McGraw-Hill.

Myers, I.B., McCaulley, M.H., Quenk, N.L. and Hammer, A.L., 1998, *MBTI Manual*. Palo Alto: Consulting Psychologists Press.

Newman, Bill, 1994, *The Ten Laws of Leadership*, BNC Publications.

Pinker, S., 2002, *The Blank Slate*. Penguin, 509pp.

Public Broadcasting Service, 1998, A Science Odyssey, People and Discoveries- Abraham Maslow Retrieved 31 August 2005. http://www.pbs.org/wgbh/aso/databank/entries/bhmasl.html.

Richards, Chris, 1982, *The Executive's Guide to Modern Management Skills,* Rydge Publications, 197p.

Samson, Danny and Daft, Richard L., 2003, *Management*, Pacific Rim Edition, Thomson.

Schein, Edgar H., 1970, *Organisational Psychology*, 2nd Edition, Prentice Hall.

Shtub, A., Bard, J.F. and Globerson, S., 2005, *Project Management. Processes, Methodologies and Economics,* Upper Slade River, New Jersey: Pearson Education Inc.

Thomas, Kenneth, 1976, "Conflict and Conflict Management", Chapter 21 in Dunnette, M.D. *Handbook of Industrial and Organizational Psychology*, pp. 889-935.

Tuckman, B.W., 1965, Developmental Sequence in Small Groups, *Psychological Bulletin*, vol. 63, 1965, pp. 384-399.

Turla, Peter (no date) Time Management Tips for Getting Results by Peter "The Time Man" Turla,Retrieved 16/08/2005. www.TimeMan.com.

Ullmann, J. E. (ed), 1986, *Handbook of Engineering Management*, John Wiley & Sons.

Vance, Greg, 1993, Getting Real, Kangaroo Press, 136p.

APPENDIX 6A THE MBTI DESCRIPTIONS

The following list comes from a number of sources including Keirsey and Bates (1978), Keirsey (1998) and Boeree (2004).

ENFJ (Extroverted feeling with intuiting): Mikhail Gorbachev, Mao. These people are the conversationalists of the world. They tend to idealize their friends. They make good parents, but have a tendency to allow themselves to be used. They make good therapists, teachers, executives, and salespeople.
ENFP (Extroverted intuiting with feeling): Leon Trotsky. These types of people love novelty and surprises and tend to be imaginative. They are big on emotions and expression. They tend to find reasons to do whatever they want. They tend to improvise rather than spend time preparing. They are good at sales, advertising, politics, and acting.
ENTJ (Extroverted thinking with intuiting): Bill Gates, Margaret Thatcher, Napoleon Bonaparte. In charge at home, they expect a lot from spouses and kids. They like organization and structure and tend to make good executives and administrators.
ENTP (Extroverted intuiting with thinking): Richard Feynman, Walt Disney and Nikola Tesla. These are lively people and tend to be outspoken, not humdrum or orderly. As mates, they are a little dangerous, especially economically. They are good at analysis and make good entrepreneurs. They do tend to play at one-upmanship.
ESFJ (Extroverted feeling with sensing): Michael Palin. These people like harmony. They tend to be very active committee members and work best with encouragement and praise. They may be dependent, first on parents and later on spouses. They wear their hearts on their sleeves and excel in service occupations involving personal contact.
ESFP (Extroverted sensing with feeling): Picasso. Very generous and impulsive, they have a low tolerance for anxiety and are sometimes labelled as performers or artisans. These people like public relations, and they love the phone. They tend to know what is going on and join in eagerly.
ESTJ (Extroverted thinking with sensing): Joseph Stalin, Harry S. Truman. These are responsible mates and parents and are loyal in the workplace. They are realistic, down-to-earth, orderly, and love tradition. The majority of engineers fall in this category. They often find themselves joining civic clubs!
ESTP (Extroverted sensing with thinking): Theodore Roosevelt, Henry Ford. These are action-oriented people, often sophisticated, sometimes ruthless - our "James Bonds." As mates, they are exciting and charming, but they have trouble with commitment. They make good promoters, entrepreneurs, and con artists.
INFJ (Introverted intuiting with feeling): Carl Gustav Jung, Mahatma Gandhi. These are serious students and workers who really want to contribute. They are private and easily hurt. They make good spouses, but tend to be physically reserved. People often think they are psychic. They make good therapists, general practitioners and ministers.
INFP (Introverted feeling with intuiting): Albert Schweitzer, Audrey Hepburn. These people are idealistic, self-sacrificing, and somewhat cool or reserved. They

are very family and home oriented, but do not like to relax. These people are found in psychology, architecture, and religion, but never in business.

INTJ (Introverted intuiting with thinking): Stephen Hawking, Dwight Eisenhower, Isaac Asimov. These are the most independent of all types. They love logic and ideas and are drawn to scientific research. They can be rather single-minded.

INTP (Introverted thinking with intuiting): Albert Einstein, Marie Curie. Faithful, preoccupied, and forgetful, these are the bookworms. They tend to be very precise in their use of language. They are good at logic and mathematics and make good philosophers and theoretical scientists, but not writers or salespeople.

ISFJ (Introverted sensing with feeling): Mother Teresa. These people are service and work oriented. They may suffer from fatigue and tend to be attracted to troublemakers. They are good nurses, teachers, secretaries, general practitioners, librarians, middle managers, and housekeepers.

ISFP (Introverted feeling with sensing): Mozart, Auguste Rodin. They tend to be shy and retiring, not talkative, but like sensuous action. They can be good at painting, drawing, sculpting, composing, dancing and they like nature. They are not big on commitment.

ISTJ (Introverted sensing with thinking): John D. Rockefeller. These people are dependable pillars of strength. They often try to reform their mates and other people. They make good bank examiners, auditors, accountants, tax examiners, supervisors in libraries and hospitals, business and boy or girl scouts! Engineers are a greater proportion than average in this category.

ISTP (Introverted thinking with sensing): Michael Jordan, Lance Armstrong. These people are action-oriented and fearless, and crave excitement. They are impulsive and dangerous to stop. They often like tools, instruments, and weapons, and often become technical experts. They are not interested in communication and are often incorrectly diagnosed as dyslexic or hyperactive.

APPENDIX 6B THE MBTI DIMENSIONS

This list is compiled from Keirsey (1998); Ancona, et al. (1999); and Borgatti (2002)).

Interacting with Others (E/I)

Extraverts	Introverts
Prefer variety and action Communicate freely Often impatient with long, complicated jobs Like having people around Are interested in the activities of their work and in how other people do it Are often impulsive Develop ideas by discussion Like greeting people Learn new tasks by talking and doing Enjoy meeting new people Seek out social gatherings When speaking publicly, will often improvise	Prefer working alone without interruptions Prefer quiet for concentration Tend not to mind working on one project for a long time uninterruptedly Are interested in the ideas behind their work Tend to think before they act Develop ideas by reflection Dislike intrusions and interruptions No strong need to meet regularly with others When speaking publicly, will prepare in depth and speak from a plan Consider consequences before acting socially Sometimes have problems communicating

Understanding the World (S/N)

Sensing	Intuition
Prefer practical problems Prefer established systems and methods Like using experience and standard ways to solve problems Enjoy applying what they have already learned: like to work with tested ideas May distrust and ignore their inspirations Seldom make errors of fact Like to do things with a practical bent Like to present the details of their work first Prefer continuation of what is, with fine tuning Usually proceed step-by-step Patient with routine detail Like to have schedule for working Search for standard problem solving approach	Dislike doing the same thing over and over Like solving new complex ambiguous problems Are impatient with routine details Like to float new ideas Enjoy learning a new skill more than using it See possibilities and implications Tend to follow their inspirations May ignore or overlook facts Like to do things with an innovative bent Have creative vision and insight Like to present an overview of work first Prefer change, sometimes radical, to continuation of 'what is' Usually proceed in bursts of energy Like innovative approaches

Making Decisions (T/F)

Thinking	**Feeling**
Try to establish objective decision criteria Measure decisions against payoffs Can be seen as hard hearted, detached and cold Decide according to situation Tend to relate well only to other thinking types Negotiate on the evidence Concern for fairness based on rules Like analysis and clarity Situation oriented Use logical analysis to reach conclusions Want mutual respect among colleagues May hurt people's feelings without knowing it Tend to decide impersonally, sometimes paying insufficient attention to people's wishes Tend to be firm-minded and can give criticism when appropriate Look at the principles involved in the situation Feel rewarded when job is done well	Personal subjective decision criteria Measure decisions against beliefs Can appear overcommitted to a point of view Believe in deciding on personal considerations Nostalgic Negotiate on rights and wrongs of issues Fairness comes from values and beliefs Like harmony based on common values Objectives emerge from beliefs Principles oriented Use values to reach conclusions Want harmony and support among colleagues Enjoy pleasing people Often let decisions be influenced by their own and other people's likes and dislikes Tend to be sympathetic Dislike telling people unpleasant things Look at the underlying values in the situation Feel rewarded when people's needs are met

Allocating Time (J/P)

Judging	**Perceiving**
Like clarity and order Work best when they can plan their work and follow their plan May decide things too quickly Concerned with resolving matters Dislike ambiguity Can be inflexible once decision is made Emphasize decision taking over information getting Like to get things settled and finished May not notice new things that need to be done Tend to be satisfied once they reach a decision on a thing, situation, or person Reach closure by deciding quickly Feel supported by structure/schedules Focus on completion of a project	Enjoy searching and finding Procrastinate decisions while searching for options Can tolerate ambiguity Concerned to knowing, not resolving problems Open-minded and curious Emphasis on diagnosing over concluding Enjoy flexibility in their work Like things open for last-minute changes Tend to procrastinate Tend to be curious and welcome a new light on a thing, situation,. or person Adapt well to changing situations and feel restricted without variety Focus on the process of a project

CHAPTER SEVEN

Communication

Effective communication is needed in all facets of engineering work. In undertaking a project, engineers have to listen to and understand others, question and discuss ideas, give feedback, write reports, take an active part in meetings, engage in conflict resolution, and develop and manage teams. In order to communicate effectively with others, it is necessary to be able to present information clearly and simply and to interpret the words, emotions and non-verbal messages of others. This chapter discusses the nature of, and strategies for, achieving effective and clear communication.

7.1 INTRODUCTION

Communication is the process by which humans interact with each other. Language, both written and spoken, enables us to develop ideas and plans, and then to communicate them to others. Language and communication together make social interaction possible. Without them, there would no society and it would be impossible to engage in engineering work. Communication is the basic ability of humans to make and interpret vocal utterances to make known their needs, wants, ideas and feelings, and to change these spoken words into written records for posterity; that make us different from other animals on earth.

Human beings have many *channels of communication* available to them. Like other animals, they can employ all of their five senses to communicate; however, social communication occurs predominantly through the use of eyes, ears and voice. For example, in face-to-face *inter-personal communication* between two people, one listens and watches while the other speaks, until the roles change. However, the message sent out by the speaker is not contained solely in the words used. *Non-verbal communication* also takes place, both intentionally and unintentionally, through the speaker's hand movements and facial expressions. The tone of voice, independently of the words used, can also transmit information. Likewise, the listener sends non-verbal messages back to the speaker, intentionally and unintentionally, through facial expressions which may contain a variety of signals such as agreement or disagreement, confusion, disbelief and non-comprehension. This is an instance of feedback. Various channels of communication are thus used to send information back and forth, even in this most common of situations.

Interpersonal communication can also take place through other channels that employ sight, with or without sound. For example, written notes and letters can be transmitted by post or by email, or even by carrier pigeon. Devices such as telephone, the internet, radio and television provide other potential communication channels. Throughout history, humans have devised a wide range of ingenious ways to communicate with each other using sight and sound to overcome the barriers of both space and time.

In the discussion to date we have focused on a specific form of communication involving two people. There are obviously other possibilities and various *forms of communication* to consider. Perhaps the simplest is *self communication,* which occurs, for example, when we talk to ourselves or write notes and messages to ourselves as reminders of things we want to remember and to record information we need to keep. The case of *person-to-person communication* has already been discussed. In *group communication* more than two people talk and listen to each other. It may be that everyone in the group speaks; alternatively, several may speak while the others listen. Meetings allow direct communication to occur in a group of people. Group communication can also occur through indirect channels such as written messages. In *mass communication* one or several people send messages to large numbers of people. In modern society, mass communication typically occurs through television, radio and the Internet. The communication in such circumstances is often one-way, in that the receivers do not have any opportunity to send messages back to the senders. There is often no *feedback* in such situations. Feedback can play an important role in communication: it provides one means for checking whether the message received is the same as the one sent. Mass communication can also take place through books, newspapers and other printed documents. An instance of mass communication occurs when a lecturer makes a presentation to an audience, with or without visual aids, in a hall. A play, staged in a theatre, is yet another form of mass communication. In the last-mentioned instances, it is interesting to reflect on the extent to which feedback can occur.

In human communication, the message that is sent is not always the same as the one received. Poor communication and miscommunication can occur in many ways. The fault may be that the original message is ambiguous or misleading. Alternatively, the error may be in a misreading of the received message. Messages can also be distorted and corrupted during the process of transmission, perhaps as a result of the channels used to send the messages. Non-verbal messages are susceptible to misinterpretation. Some people seem to be naturally good and effective communicators and almost always get their message across. Others do not communicate clearly and effectively, even in the simplest circumstances. Some people are good listeners, others less so. Depending on the channel of communication used, the message can become distorted during transmission.

In all forms of successful human communication, irrespective of the channels employed, there must be at least one person sending information (the sender) and at least one person receiving the information (the receiver). Some of the problems of communication are illustrated in the model shown in Figure 7.1, where there are two parties who send and receive information in turn. Each message is encoded and sent using voice (words and tones), eyes and body language, but is susceptible to *noise* which may result in loss of clarity and distortion. The noise depends on the channel used for communicating. When received, the message is decoded. The presence of noise, coupled with the possibility of faulty encoding and decoding, means that the message received may not be the same as the message sent. The

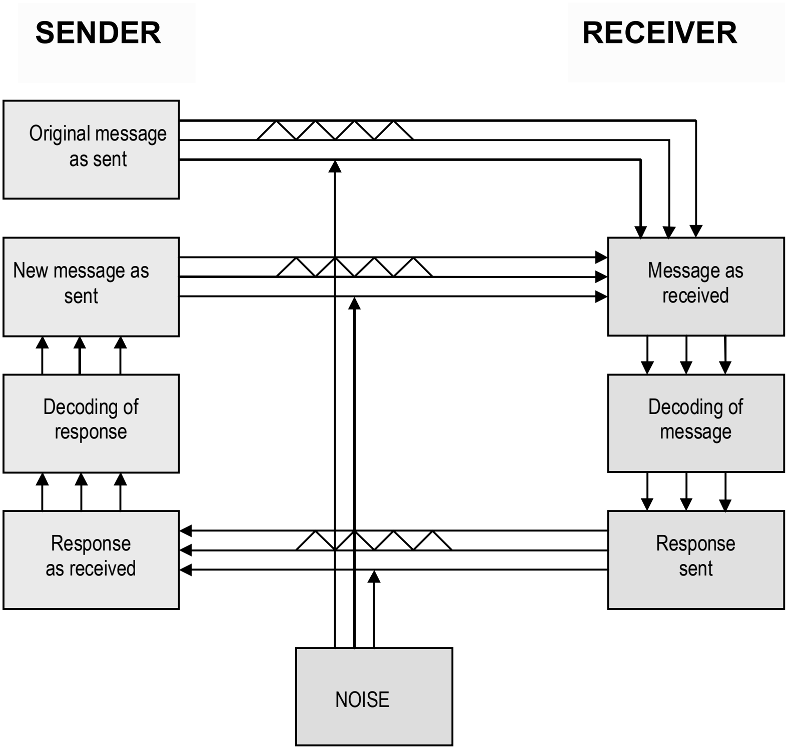


Figure 7.1 Model of the communication process.

model stems from Shannon's original model of communication in a telephone network (Shannon and Weaver, 1949).

Communication plays a particularly important role in all engineering work. It dictates how problems are formulated and how they are solved. The quality of the communication among team members and between client and engineer determines the quality of engineering plans and design solutions. Good communication is necessary in order to carry good designs and plans through to successful implementation. Frequent and effective communication must occur among the members of any engineering team if it is to be successful. Lack of communication, poor communication and miscommunication in an engineering project can lead to catastrophe, as was the case in the Space Shuttle Challenger disaster where different companies' management and NASA were told of critical safety indicators by engineers but allowed the Shuttle to be launched against many people's objections (Boisjoly, 2006).

In *engineering communication*, whether direct or indirect, a great deal of information can be transmitted clearly and unambiguously through the use of sketches and drawings. These provide an added communication channel.

Experienced engineers, when talking to each other, usually resort to simple sketches to get their ideas across clearly. The ability to sketch and the conventions used in sketching are therefore of great importance to engineers. It is very important to learn these skills. Sketches and figures are also used in more formal channels of communications, such as reports, design drawings, plans, and even books, and lecture notes. Engineers must be able to communicate effectively in activities in order to

- manage teams;
- participate in team work;
- counsel staff under their supervision;
- give feedback to team members;
- resolve conflicts among team members;
- present concepts to other engineers;
- present proposals to clients;
- interact with the community in consultations and public meetings; and
- prepare and describe clearly documents and design details.

Good engineering is of necessity based on good and effective communication. Fortunately, the ability to communicate effectively can be learnt. Even those who are naturally good communicators should not rely on their native talents. In the following sections, various aspects of communication that are relevant to engineering work are discussed.

7.2 PREPARING FOR COMMUNICATION

One of the first and most important keys to successful communication is adequate preparation. Before an effective and successful engineering message can be sent, the content of the message has to be clearly worked out and fully understood by the sender. This is particularly important if the message is complex, which is often the case for engineers.

Even when the content of the message has been worked out, it is necessary for the potential sender to put time and effort into choosing the best way in which to present the information so that the persons receiving the message will most easily understand it. In preparing for sending, it is important to evaluate the potential receivers, their backgrounds and expertise, and their level of understanding of the technical matters involved in the engineering message. For example, the communication of design information to non-specialist clients and the interested public will require a different form of presentation from that used for engineering colleagues who have a good technical understanding of the field. It may well be that a broad overview is better for non-specialists than detailed data which will not be understood.

The sequencing of information also needs careful attention. While a logical sequence is usually preferable, it will not always be the best. A logical but long and complex argument can rapidly become boring and hence lead to inattention and misunderstanding.

Methods for organising thoughts and for presenting ideas and concept.

Before an idea can be communicated, it must be developed. There are a number of simple tools which can assist in the development of ideas, ranging from hierarchical lists, using word outline for ease of developing hierarchy, *mind mapping*, and picture and concept mapping (Buzan, 1988; Checkland, 1981; Horan, 2002). These methods advocate a systems approach to thinking by using images to link key ideas and hence enable transference of ideas between thinking and communicating. The mind mapping terminology was popularised by Buzan (1988) to be used for problem solving but also for learning, communicating and remembering. It appears that the future will include more tools for using these techniques, as Bill Gates (co-founder of Microsoft) stated that "a new generation of 'mind-mapping' software can also be used as a digital 'blank slate' to help connect and synthesize ideas and data – and ultimately create new knowledge" (Gates, 2006). The basics of developing a mind map will now be discussed.

Mind mapping

A *mind map* is a graphical representation linking related ideas and concepts. It is created with a central image, word or concept that is under consideration. Around this central concept, up to 10 main ideas relating to this concept can be drawn or placed. Using Kipling's six question words is a simple method of beginning: What? Why? When? How? Where? and Who? Each of these images, words or concepts are considered and up to a further 10 main ideas or concepts which relate to each of them are generated. Without too much effort, a large number of related ideas can be produced. With this simple technique, it is possible to get a clear understanding of most problems.

As an example of what is meant, a mind map that illustrates the concept of mind maps has been created using Free Mind software (2006). This is shown in Figure 7.2. Despite the fact that mind maps follow the way the brain is believed to work, mind mapping may initially seem to be an unnatural way to record or process information. Experience has shown that a little persistence pays off and people who use mind mapping find it a valuable tool that can be employed to assist in communication. These techniques can be an instrument in developing a systems approach to problem solving and most web pages store their information in a similar way, with links to relevant information being the cornerstone of the world wide web.

One of the key reasons that mind mapping is so successful is that it forces the listener to not just record the points a speaker or article is making, but to engage with the speaker and to consider not just the importance of each point but also its relationship to the general theme or argument. This results in a deliberate sense of active listening or thinking that works the brain harder tending to produce greater understanding. Lazslo (2001) also noted that "The full potential of human communication unfolds only when the communicators understand the strands of connection through which they communicate". To communicate at a high level requires people to make use of the different paths of communication that are available.

Preparation for communication is not purely the task of the sender. The receiver also has preparatory work to do in order to ensure good reception. Even when attending a meeting purely to listen and to obtain information, it is worthwhile planning for the event by thinking about what can be achieved and what information can be obtained. It is rare that the attendees at a meeting all have the same aims and agendas. With minimal preparation, it is easier to question speakers to obtain clarification and to ensure they provide the required information.

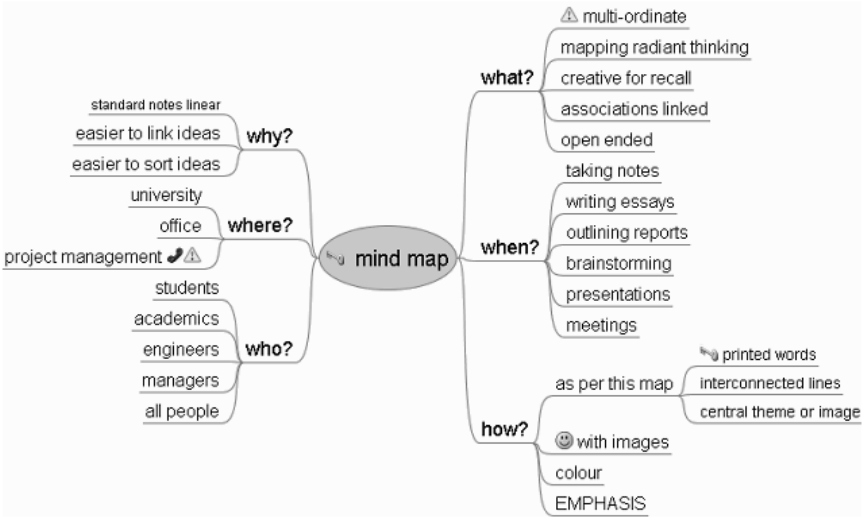


Figure 7.2 A mind map of the concept of mind maps.

Preparation is worthwhile prior to all communication. Probably the most important thing is to take sufficient time to organise yourself properly and develop techniques for storing information so that it can easily be found and retrieved. Being systematic is something that is well worth learning and developing.

7.3 ACTIVE LISTENING

People who are good listeners have a high success in getting the information they want from others. It has been suggested that they follow a process called *active listening* (Steinmetz, 1979). *Active listening* includes being encouraging, being reflective, probing for information, and using a summary technique.

One way of being encouraging is to use open rather than closed questions. A closed question is one where the answer is either a single word or a short phrase. Closed question words include: Is, have, has, does, could, can, will, are and shall. For example: “How old are you?” and “Where do you live?” are closed questions. *Closed questions* have the following characteristics:

- they request factual information;

- they are quick and easy to answer; and
- they enable the questioner to control the conversation.

On the other hand, an open question is one that generally receives an extended answer. Open question words include: describe; what; why; and how. Thus “What do you think you will be doing in five years?” is an open question. Open questions have the following characteristics:

- they ask the respondent to think and reflect;
- they try to elicit opinions, knowledge and feelings; and
- they deliver control of the conversation to the respondent.

The *open question* gets the respondent thinking and giving useful information about himself or herself or the subject being discussed. Note: some question words can elicit either a short answer or a long answer depending on how they are phrased. Some examples of closed and open questions are given in Table 7.1.

Table 7.1 Samples of closed and open questions.

Closed question	Open question
Are you happy with the job?	What do you like or dislike about the job?
Are you sad about the decision?	Describe your feelings about the decision?
Do you like your manager?	How do you feel about your manager?

Once the open-ended question has been asked, an active listener will give free rein to the person answering and so encourage an uninhibited response. The aim is to get the responder to talk freely, and part of this freedom to talk involves the listener not talking. Silence can promote talking but so can non-verbal communication. Appropriate body motion should not to be distracting, artificial or forced. Movement, through the head, eyes, hands or any other body part which signals interest and attention, improves communication. Awareness of other people’s non-verbal communication is paramount for clear understanding.

Reflective skills tend to help the listener keep track of the message. Reflecting skills improve the dialogue, by providing, as well as asking for, information. The most effective reflecting device is ‘paraphrasing’. It serves to confirm that the person has been listening, but also requests more information. Paraphrasing helps to clarify a message or attempts to reflect the feelings of the speaker by summarising what has been said in fewer words.

Applying active listening skills takes practice and people should beware of jumping to an incomplete picture of what others are talking about and then offering a master plan to solve the problems, without recognition of the thoughts, feelings and emotions of the person they are supposedly listening to. Rogers and Farson (1979) stress the importance of listening as “an important way to bring about changes in people”. They suggest it involves three activities: listening for the full meaning including both the content and the underlying emotions; responding to feelings (if the real message is an emotional one then the response should address that) and noting all the cues by observing the non-verbal messages.

7.4 NON-VERBAL COMMUNICATION

Non-verbal communication can be thought of as all the information transmitted other than the message contained in the actual words used. It includes both the visual element and the vocal tones (Samson and Daft, 2003). The face is used a great deal to transmit information and this is usually the first visual element noticed. People often notice the hand signals, shrugs, head movements and body movements but do not necessarily interpret the whole message. Non-verbal communication can be used for:

- expressing emotion (e.g. smiling to show happiness);
- conveying attitudes (e.g. staring, glaring to show aggression); and
- demonstrating personality traits (e.g. open palms to show accepting qualities).

Non-verbal behaviour varies across cultures, although the six emotions of anger, fear, disgust, sadness, happiness and surprise seem to prevail around the world. The ability to understand and use non-verbal communication is valuable for any workplace environment. Some non-verbal communication strategies are easier to use than others. For example, when going to a meeting or interview, the way a person dresses can portray a certain image. Other non-verbal communication strategies such as being punctual, animated and demonstrating friendliness by a smile are able to communicate an image about a person in a much more effective way than the spoken word.

Facial expressions, eye movements, and head movements constitute a significant percentage of recognised non-verbal communication. This is because people tend to look at a person's face and eyes when speaking with them. It is estimated that the human face can show more than 50,000 different expressions (Hamilton and Parker, 1990), and most people are able to interpret the meanings or feelings associated with many of the common ones, such as surprise, joy, suspicion, anger or menace. The skill to read small facial changes and eye movements requires greater observation and awareness but is useful to ensure that one obtains all the information being transmitted in a conversation.

Being able to interpret body language and gestures is vital in establishing and maintaining good working relations with work colleagues and clients. If clients exhibit discomfort or unwillingness to participate, then it is possible to alter the environment and the communication strategy to help them feel more comfortable. It is doubly important to be aware what messages your own body language is sending, to avoid sending an incorrect or ambivalent message, and, more importantly, to ensure that you send the correct message. A useful strategy is to watch for reactions when speaking and then to try and spot contradictory messages when, for example, the voice says one thing and the body says another. This can indicate deception but one should always check this out by follow-up questions in the context of the conversation. A simple test can be performed to illustrate the importance of non-verbal communication: listen to someone with your eyes closed, then listen and watch. It should be evident that there are differences and that it is much easier to understand someone when you can see them.

Non-verbal behaviour patterns for communication

To deliver a greater impact when a person delivers a message, the use of non-verbal behaviours will raise the level of interpersonal communication. The following material has been compiled from a variety of sources (Eunson, 1994; Hunt, 1979; James, 1995; Quilliam, Susan, 1995, Body Language, Carlton Books. Tidwell, 2005; Pease and Pease, 1999).

One of the most important behaviour patterns involves eye contact. Eye contact signals interest in others and increases the speaker's credibility and helps to regulate communication flow. However, people should be aware of cultural differences. In Western countries people who make eye contact open the flow of communication and this shows interest and credibility. On the other hand, people from Japan, China, Africa, Latin American, and the Caribbean tend to avoid eye contact to show respect. Western cultures see direct eye-to-eye contact as positive but there are peculiarities even within a country. For example, in the United States African-Americans use more eye contact when talking and less when listening, with the reverse being true for Anglo-Americans. While a prolonged gaze is often interpreted as a sign of sexual interest, Arabic cultures use prolonged eye contact because they believe it indicates interest, and hence assists them in gauging the truthfulness of others. A person who does not reciprocate is seen as untrustworthy.

Facial expressions, such as smiling, transmit friendliness, happiness, warmth, and a connotation of welcoming. People who smile frequently are perceived as more likable, friendly, warm and approachable. Smiling can be contagious and people tend to mirror smiling to gain rapport. They will be more comfortable and will want to listen more. It appears that facial expressions have similar meanings worldwide with respect to smiling, crying, or showing anger, sorrow, or disgust. Many Asian cultures suppress facial expression as much as possible. Many Mediterranean (Latino/Arabic) cultures openly express grief, whereas it has been a trait of men of English speaking cultures to hide grief or sorrow. Overall, there is a tendency for women to smile more than men, but in some groups such as African-Americans there is little difference in the degree of smiling between genders.

Gestures are a natural part of human communications and failing to make gestures while speaking may lead to a perception of the speaker being uninterested or boring. Congruent gestures with speaking capture the listener's attention, and facilitate understanding. It is impossible to list all gestures, but it is important to remember that what is acceptable in one’s own culture may be offensive in another. Some cultures are restrained in gestures while others are animated, with the consequence that the restrained cultures feel the animated cultures lack manners. On the other hand, animated cultures often feel restrained cultures lack emotion. The use of hands for pointing or counting differs from country to country. For example, in the United States, the United Kingdom and Australia pointing is done with the index finger, whereas in Germany the little finger is used. The Japanese prefer to use the entire hand (in fact, most Asian cultures consider pointing with the index finger to be rude).

Posture and body orientation can also be used for communication. Standing erect and leaning forward communicates to listeners that the speaker is approachable, receptive and friendly. Interpersonal closeness results when the speaker and listener face each other. Speaking with the back turned or looking at the floor or ceiling should be avoided as it communicates lack of interest. There are also cultural differences for a range of what might appear to be quite common actions. In Japan and Korea, bowing shows rank and is a sign of respect. Slouching is considered rude in most Northern European countries. Standing with hands in pockets is disrespectful in Turkey. Sitting with legs crossed is offensive in Ghana and Turkey, but traditional in Korea. Showing the soles of the feet is offensive in Thailand, Malaysia, and Saudi Arabia.

Proximity, the distance between speaker and listener, is a key area where problems can occur. Cultural norms dictate a comfortable distance for interaction with others and people should look for signals of discomfort caused by invading other people's space. Some of these signs are rocking, leg swinging, tapping, and gaze aversion. The context is also important: what is perfectly acceptable on a crowded train or bus is not acceptable in a normal business meeting. In Arab cultures, invasion of space and to stare closely into the eyes is considered normal.

Although speaking is the essence of verbal communication, it can also have non-verbal aspects to it. The way the words are spoken, through such aspects as tone, pitch, rhythm, volume and inflection can have an important effect on the message being transmitted. One of the major criticisms of many speakers is that they speak in a monotonic voice, thus appearing to be boring and dull to listeners.

Touch is generally culturally determined, with each culture having a clear concept of what parts of the body one may touch or not touch. Touch generally shows emotions or control and varies between genders and cultures. In Western cultures, particularly in the US, Australia, Canada and the UK, handshaking is common (even for strangers). People of Islamic and Hindu backgrounds avoid touching with the left hand because to do so is a social insult, as the left hand is used for toilet functions. It is the custom in India to break bread only using the right hand (a difficult task for non-Indians). Islamic cultures do not approve of touching between genders in public (even hand shakes), but consider such touching (including hand holding and hugs) between same-sex people to be appropriate. Many Asians avoid touching the head because the head houses the soul, and a touch puts it in jeopardy. Cultures such as the English, German, Scandinavian, Chinese, and Japanese which have high emotional restraint have little public contact, whereas those which encourage emotion, such as the Latino, Middle-Eastern and Jewish, have more frequent touching.

Judgements, based on looks and dress, seem to be a trait of all cultures. Europeans and Americans appear to place a high value on dress, grooming and personal attractiveness, partly induced by marketing and affluence. Around the globe there are differing cultural standards on what is appropriate, attractive in dress and what constitutes modesty. Companies have different dress codes which are used in the corporate world as a sign of status but dress codes can alter quickly.

It appears that as verbal codes are used within a particular cultural context, they evolve over time (Underwood, 2003). This also happens to non-verbal codes. Through travelling and working in different countries, an increased awareness of how specific gestures are different from one country to another is developed. For example, total confusion in a conversation can result from Indians shaking their head in response to matters being discussed. One needs to be very cautious and rephrase conversations to adjust to body language in this environment.

People tend to define a space bubble or proximity by their own culture. For example, when dealing with some Middle-Eastern and Asian cultures, one finds that space limits are vastly different from those in Western cultures. In some situations people stand so close that Westerners can feel quite uncomfortable with a constant desire to back away. Other cultures find Europeans cold and aloof because of this tendency to back off.

Checklist: Tips for improved non-verbal communication

► Make yourself at ease with the person you are communicating with. Avoid being too close or too far away. (Within 600-700 mm is a comfortable range for city dwellers of Anglo Saxon heritage.)
► Be attentive and try to relax, but avoid slouching or sitting rigidly. Show interest by leaning slightly toward the other person.
► Avoid staring or glaring, but try to maintain frequent eye contact.
► Respond with non-verbal communication while the other person is talking by simple nods for approval or agreement.
► All non-verbal gestures should be natural, smooth and unobtrusive. Do not allow gestures to dominate your words. Be aware of gestures that reveal negative emotions and frustration.
► Slow down your rate of speech to a little slower than normal, to avoid indicating impatience. Use the tone of your voice to give a feeling of warmth and acceptance.
► Do not mumble but maintain a clearly audible voice, not too loud nor too soft.
► Avoid using your limbs, hands and feet as barriers.
► Appropriate genuine smiling assists in gaining rapport.
► Closing eyes and yawning can block communication, so attempt to be alert when interacting with others.

Adapted from Messina (1999)

7.5 ORAL PRESENTATIONS

Woodrow Wilson, the 28th president of the United States said, "If I am to speak ten minutes, I need a week for preparation; if fifteen minutes, three days; if half an hour, two days; if an hour, I am ready now". This clearly indicates that good presentations do not just happen, and if someone makes it look easy then it is likely that much more effort and practice has gone into the talk than appearances would indicate.

Oral communication skills are required by all engineers, as a consultant to present a company's case to win work, to present results to senior management and clients, or to present details of projects to the community. Technical presentations can involve both oral and written communication. In all communications, an awareness of the audience is essential to ensure a consistent style, flow and clarity of detail. There is no easy way to ensure a presentation goes well. However, there are some key points that should be addressed well before the time of the presentation.

Checklist: Oral presentations

Preliminaries
► What is the objective of the talk?
► What do you want your audience to think or do as a result of this talk?
► Analyse the audience: what do you know about them, their backgrounds, ages, needs, motivation?
► Consider issues like the time of day, number in audience, and size of room.

Opening
► How will you hook the audience?
► What will your first words be?
► Ensure that you thank the chairperson and address the audience!.
► What is the purpose of the talk?
► What benefit will the talk be to the audience?

Signposts/Material Content
► Can you create a mind map showing the content of the talk?
► Point the way through your talk – use headings.
► Research your talk thoroughly. Preparation is the key to a good talk.

Notes
► Short memory aids only.
► Key words or ideographs or even a mind map on a small card.
► **DO NOT READ**

Visual aids
► Ensure visual aids are clear, easy and appropriate for ideas being expressed.
► Present them effectively - do not have colours that make the text unintelligible.
► Use a sans serif font (e.g. Arial) and a minimum font size of 18 pt.
► Have a maximum of 30 to 35 words per slide : fewer is better.
► Describing a picture or graph is better than a slide full of words.
► Reinforce, explain and illustrate your ideas.
► Aim for approximately one minute per slide (as a rough guide)

Delivery
► Transmit energy and enthusiasm – use positive gestures and eye contact.
► Ensure voice is clear and well modulated.
► Body behaviours have to be congruent – generally be yourself.

Closing
► What are the main points to be reinforced in the summary?
► What is the final impression I wish to make?
► What will be my final words?
► Summarise and finish with a well-prepared strong closing.
► Remember, it is the chairperson’s responsibility to handle question time.

A common failing with presentations is running short or running out of time. Speakers should consider their timing carefully and practise their talk to ensure that it will run to time. It is at the end of the talk where it is likely that the key points are to be made and where, if the ending is rushed, much impact will be lost.

Rather than just attempting to say everything more quickly, it is better to leave out a whole section so that time can be spent on giving a good summary of the rest of the work and leaving the audience with a positive feeling about the presentation. If there are people in the audience who are particularly interested in the details of the work, they can always seek the speaker out after the talk.

7.6 WRITTEN COMMUNICATION

Written communication, including reports, emails, letters, memoranda, sketches and drawings, is perhaps the form of communication where the most effort is expended. In talking and presenting material orally, instant feedback is generally received. An audience that is slowly drifting off to sleep is letting you know your presentation is not going well and that something needs to be done. A written report receives no such instant feedback, but can generate the same consequences, and there is no chance to make immediate amendments. During the normal course of events, engineers will be required to write resumés, business letters, memoranda and reports for many different reasons. In the following sections, suggestions are given to make writing as easy as possible and to ensure that it conforms to some basic standards. The idea that writing should follow a standard may seem somewhat constraining, but it should be remembered that one of the main reasons for communication is to present information clearly and to make the job of the reader as easy as possible. If this means a little extra effort for the writer, so be it.

Checklist: Writing

► Avoid jargon.
► Have a structure with an introduction, main content and a conclusion or recommendations.
► Ensure all written material is checked for spelling, grammar and punctuation.
► Read the whole piece from top to bottom.
► Have colleagues read at least key sections of the report.
► Leave sufficient time so that it is not rushed. Rushed sections are likely to be at the end and this is where the key results are found and impressions made.

Memoranda

The purpose of a memorandum is to give information to one or more people at once within an organisation. It is, at the same time, a useful and formal way of having your views or advice on a particular topic entered onto the official record. Although emails have largely replaced the written memorandum in the modern engineering office, they may still be used for more formal or more important transmissions, and for this reason it is important to be able to write and structure one in an appropriate format.

Memos are usually restricted in length and contain important information in a clear and easy to find format.

Checklist: Memorandum

To:	[Name of person to whom the memo is being sent]
Cc:	[Names of other people who need information]
From:	[Name of sender]
Re:	[What the memo is about]
Date:	[Date memo sent]

- § Make the message as short and clear as possible.
- § Use bold text to make the most important information clear.
- § Make sure that all memoranda are proofread and spell-checked.

The email

Email is a form of communication used every day in business and for private purposes. A few simple rules and a little care can eliminate the drudgery of having to respond to numerous emails every day. In many ways the email is similar to the memorandum. However, there is one important difference. The email has, by its nature, an urgency and an informality about it.

Checklist: emails

- ► Put enough details in the subject line so that meaning of the email is conveyed, such as “Remember Salinity Project meeting Tuesday 25 July”.
- ► To reduce the number of emails received, when appropriate insert “No Reply required” in the “Subject” line or the opening line of your e-mail.
- ► Avoid overuse of CAPITAL LETTERS, as this conveys anger and shouting; also avoid “exclamation marks” and bold, which can put too much stress in the message!!
- ► Cut and paste standard replies for frequently asked questions or requests.
- ► Although emails tend to be less formal, do a spell-check before sending.
- ► Only reply to the sender if necessary, rather than to all on the sender’s list.
- ► Only have one subject per email. People often respond to your first and last questions, but overlook or forget the others, so keep things simple.
- ► When replying to a series of questions, embed your answer after the question or other remarks at appropriate places.
- ► Any email you send could be forwarded to others so make sure you are aware of that and hence make the receiver aware of the fact if you do not want it to be forwarded on by marking it as “Confidential”.
- ► Use plain text, simple language, short paragraphs and keep messages under 25 lines long, if possible, as this is the length that can be read on one screen.
- ► Alert recipients when sending large attachments.
- ► When forwarding a message, put comments at the top rather than at the end.
- ► Always obtain permission from an author before sending on personal emails.
- ► Send information using a text format, if at all possible, rather than HTML.

A former NASA engineer who now runs time management seminars, asked a group of time management participants for ideas on improving time used on writing and sending emails. The above email checklist was developed from some of those responses (Turla, 2005) and combined with those in McShane and Travaglione (2003).

Emails can be quick to send and almost demand a quick response. Engineers should ensure that their responses are well considered because, as indicated previously, the reply is on the public record.

Business letters

When reports are written, they will generally be sent with a covering letter. That letter then is the first thing that the client will read and therefore is important in creating the first impression, similar to the cover of a report. Most companies have quite strict standards for letters and it is better to know what they are, rather than risk the task of many re-drafts, or worse, upsetting a client or boss. A typical letter is shown to bring out some of the key elements that are generally required.

XYZ Environmental Solutions
PO Box 2222
ADELAIDE 5001
1 January 2016

Ms Caroline Chong
Manager
Engineering Constructions
PO Box 888
SYDNEY 2003

Dear Caroline,

RE: Contract for Sediment Ponds AW2005/89

The body of the letter should be clear, concise and courteous. The way you organise the body of your letter will depend on the reason for writing it, but many of the other features will be common to all formal business letters.

Yours sincerely,

William Smith

The company letterhead will generally give contact details of the company. The date is important both for sequencing of communications and filing, but it also can have legal ramifications in situations where the time advice given or received is crucial to subsequent actions. In a business letter, the date is located under the company letterhead or the address. Formality generally applies to the details of the person to whom the letter is written, including his or her title and company details. As an aid to understanding, it is helpful to state clearly what the contents of the letter are in regard to (RE:). Once again, this is to make the receiver's task as easy as possible. A business letter should be structured so that it has a beginning, a middle and an end. In the main body of the letter, the main information is presented. A key aim is to ensure that the letter is not misunderstood and for this reason it is essential that the language is clear and simple to avoid misunderstandings. There are no awards for showing a large vocabulary in a business letter. The most important thing is to ensure that you communicate your message as effectively as possible.

The final words should tell the person what actions are required or perhaps give thanks for any help given. It is customary to sign off with; "Yours faithfully" in a formal business letter and when the name of the person is unknown; and with "Yours sincerely" if the person is known personally. Nowadays it is acceptable to use less formal closings such as "Kind regards" and "Best regards" if there is less formality in your letter. The closing signature can be located on either side of the letter. Most companies have a template for writing business letters in their own style and therefore this should be used.

Report Writing

Writing reports takes more time and is harder work than most people think. A poorly presented report, where the author is unaware or uninterested in the most basic rules, will be read with this in mind and will most likely lead to some information or recommendations being discounted or totally ignored. The engineering report conforms to quite a standard format. The aim is to ensure that the readers, no matter what their interest in the work, will be able to get the appropriate level of detail that they require: that level of detail is determined by the reader, and not the writer. For this reason, it is quite common for the same material to appear a number of times in the same report so that it can be found in an efficient manner.

The checklist for *formal reports* shows the structure that should be developed for technical reports.

The title page should include the title of the report, the names and affiliations of the authors, the date (month and year) of publication and perhaps some company identifiers.

The abstract or executive summary should take only a few pages at most and should summarise the whole report in a highly condensed form. It should allow anyone who just wants to know what the report is about to find what work or research was undertaken and what conclusions were reached, without having to look at the body of the report at all. It should not read as a teaser to the main report and generally does not include references, or extended descriptions.

Word processing software will generate tables of contents, lists of figures and lists of tables automatically. The information required is taken from the actual report so any changes can be reflected in an updated table of contents easily.

Checklist: A formal report

► Title Page includes title, authors and date

► Executive Summary includes:
reason for study;
summary of what work was carried out; and
summary of key findings and/or recommendations of study.

► Table of contents (may be automatically generated)

► Introduction
starts on page 1;
sets out reason for report / study; and
refers to copy of design brief (if available) in Appendix.

► Other Chapters
develop an argument that leads logically to the conclusion;
lead to a solution or outcome;
have sections and sub-sections that are numbered consistently; and
ensure all ‘facts’ from literature are referenced.

► Illustrations and Tables
ensure all have labels and captions in the proper place (below or above);
captions allow the figure or table to be understood independently of text;
all appear after first reference in text; and
ensure all figures are referenced in text.

► References
listed in alphabetical order in reference section; and
listed in a standard format (e.g., Harvard system, Snooks & Co., 2002).

► Appendices should only contain information that distracts from the flow of the argument in the main report and the reader of the report is not likely to want to see.

► Miscellaneous
Spell-check entire document;
write in the third person, i.e., “This was carried out” rather than “we did this”;
search for and eliminate 1st person text (I, me, my, we, us, our);
read the whole report before submission;
one-sentence paragraphs should be avoided;
single-paragraph sections or subsections should be avoided;
avoid vague terms such as ‘very large’, ‘really expensive’; and
avoid use of capital letters in middle of sentences unless for proper names.

Most reports start with a chapter entitled ‘Introduction’. It contains a brief description of the need for the report, gives general background information and may even give a brief summary of the contents. A report should be written in the form of an argument that takes the reader from the problem to a solution and a clear introduction is necessary to show how this argument will be made.

The body of the report may contain a number of different chapters that set out in logical order (rather than the order in which the work may have been carried out) the argument the author or authors are making.

The final chapters of a report tend to be a summary and set of conclusions and recommendations. In this chapter or chapters, the authors summarise and integrate any conclusions that have been made in report. Recommendations may be to do with how information should be used or what further work needs to be carried out. Following the report, there will generally be a need to list the sources of information which were referred to in the report. There are a number of standard ways of presenting this information. Most will contain, as a bare minimum, the author(s), the date of publication, the title of the work, the journal or report or book or conference proceedings that it appeared in and/or the publisher. Also given are page numbers if the work was in a journal, or the total number of pages if the work is a book. The general rules also apply to web references. Here one additional piece of information is generally required: the download date. This is because web references change quickly and it is useful to know when the information was sourced.

The last element of a report may be a single appendix or a series of appendices. Information and results which are useful background material but are not needed for understanding the main arguments in the report should be included in an appendix. Bulky test results or mathematical derivations can also be included, but note that not all test results need to be included. Generally within an engineering report there will be many figures and drawings to enhance the communication and to enable the reader to fully grasp the subject material. In design, the main form of communication is through engineering drawings.

Drawings

Much engineering communication is completed through drawings. The ideas evolve in sketches and doing calculations to formalise a design. The design of a system, whether it is an engine, a building, a water treatment plant or a space shuttle will be communicated through the various phases of a project from preliminary investigation to the construction phase of a project, as described in Chapter 3, by a very large number of drawings. With computer-aided drafting techniques, many different forms of drawings have evolved, including 3D details, animations and traditional plan and cross-sectional views. An important part of any drawing is the Title box because this has boxes for various signatories such as project manager, designer, drawer and checker. Signing off on these indicates the people responsible for the design.

7.7 COMMUNICATION IN GROUPS

The success or failure of large engineering projects inevitably depends on communication within and between groups. Engineers must therefore know the basics of communicating within groups and facilitating groups and group meetings.

Communication is an integral part of developing synergy in a group to develop trust and cooperation. Covey (1989) describes the synergy that comes from good communication as an understanding that the whole is greater than the sum of the parts. It comes from respecting differences between individuals and relies on building on strengths to compensate for weaknesses. Most importantly, it requires both trust and cooperation, as illustrated in Figure 7.3. Note that if there is high trust and high cooperation, then a win/win situation can result.

Many of the activities for good group performance were outlined in Chapter 6 with the leader of a group having to be aware of the backgrounds and styles of the members of the group and the need to ensure participation from all. Simple processes for conducting meetings were also outlined to ensure effective meetings.

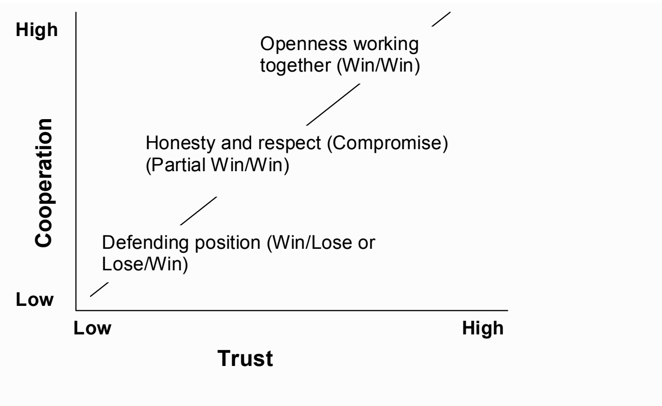


Figure 7.3 Levels of communication (adapted from Covey, 1989).

7.8 PLAGIARISM

Plagiarism is the unacknowledged use of the material written by others with or without the intention of passing it off as your own work. With the development of electronic information systems, plagiarism is on the increase, as it is becoming much easier to source material from a range of different areas.

Plagiarism is wrong and dangerous for a number of reasons. If you have read something, it is actually in your interest to cite the source and to quote directly if it is something that is going to help your argument. That is what is referred to as scholarship. Plagiarised material is often quite easy to detect, either by simply reading the material and reflecting on the changes in language, or by using one of the many electronic plagiarism detectors that are available nowadays that can check submissions against a vast range of publicly available material from around the world. It is essential, therefore, to reference all the sources of information which are used in any report or design. When using material, ensure that the source is credited somewhere in the material, e.g., according to “Smithers and Smith (1999).” When exact phrases from the original source are used, be sure to use quotation marks to identify any exact text. Paraphrasing can also be deemed to be plagiarism when the same ideas are used, without acknowledgement as in the original source, so make sure all ideas are referenced.

Contrasting Styles of Communication

Bennis (1997), the author of 26 books on leadership, has a favourite example of two contrasting styles of communication, those of the British Prime Ministers Gladstone and Disraeli.

It was said of Gladstone, a 19th century Liberal Prime Minister, that if you had dinner with him, you came away believing that he was the world’s brightest, wittiest, most charming man. However, if you had dinner with Disraeli, a peer who became a Conservative Prime Minister, you came away believing that you were the brightest, wittiest, most charming person in the country.

7.9 SUMMARY

All engineers require good communication skills, whether in managerial positions or technical specialist areas. Communication involves more than simply sending a message; it necessitates that the correct message is received by using feedback and active listening. Managing personnel in project teams also requires an understanding of how the individuals within the team can facilitate or block communication between members. In all communications, an awareness of who is receiving the message is essential to ensure a consistent style, flow and clarity of detail. There is no easy way to ensure an oral presentation or a written report is well received other than by putting a lot of effort into the structure and flow of the material, ensuring that it is relevant to the topic being discussed, and looking for feedback.

In larger organisations the manager of groups must be aware that there are some people who listen and others who do not. A strategy that assumes everyone understands what has been said can lead to disaster, therefore miscommunication can be avoided by asking listeners and readers if clarification of details is required. A good rule is to attempt to keep communication channels open and be receptive to communication at all levels. It is imperative that other people’s body language is observed when involved in discussions and when there are conflicts. It is important that as a manager and leader of engineering teams you have consistent communications, body language and actions, thus imparting a feeling of trust amongst the team. This example to fellow team members will then be the catalyst to open communication within the group and across the boundaries of other groups within the organisation.

Use simple language to convey your message in all reports, memoranda and letters. Remember to read and check the spelling and grammar of all outgoing correspondence. Always reference all sources of ideas and information.

PROBLEMS

7.1 What processes are important in giving a good presentation? What checks of the room where you are going to present would you make?

7.2 Give guidelines for reading body language by the following visual clues:
- (a) facial expression;
- (b) posture;
- (c) gestures;
- (d) clothing, grooming and environment.

7.3 Give guidelines for reading body language by the following auditory clues:
- (a) specific words that are spoken;
- (b) sound of the voice;
- (c) rapidity of speech, frequency and length of pauses.

7.4 Everyone will benefit from improving listening skills. Search the Internet to find information on becoming a better listener. Select one or two areas that you want to improve on and try those over the next 48 hours. Write a one-paragraph report on why you think you have improved or not.

7.5 Write an essay of no more than 1000 words explaining the importance of good communication skills for engineers. Use a mind map to prepare the outline of the essay and which can also be used for an oral presentation.

7.6 Draft a letter to seek employment from a company that you have seen is doing interesting work in an area of engineering that interests you.

REFERENCES

Bennis, W., 1997, *Organizing Genius, The Secrets of Creative Collaboration,* Transcript of Interview, March 26, http://www.pbs.org/newshour/gergen/march97/ bennis_3-26.html, Retrieved Feb 2006.

Boisjoly, Roger., 2006, "Ethical Decisions - Morton Thiokol and the Space Shuttle *Challenger* Disaster", onlineethics.org. Retrieved February 24, 2017.

Buzan, T., 1988, *Make the Most of Your Mind*. Pan.

Checkland, P., 1981, *Systems Thinking, Systems Practice*. John Wiley & Sons, Chichester.

Covey, S.R., 1989, *The Seven Habits of Highly Effective People*, Simon & Schuster, NY.

Eunson, B., 1994, *Communicating for Team Building*, Jacaranda Wiley Ltd.

Free Mind 2006, http://freemind.sourceforge.net/wiki/index.php/Main_Page, Retrieved Feb 2006.

Gates, Bill 2006, The Road Ahead - How 'intelligent agents' and mind-mappers are taking our information democracy to the next stage. *Newsweek,* Jan. 25, 2006 http://msnbc.msn.com/id/11020787/site/newsweek/, Retrieved Mar. 2006.

Haith-Cooper, M., 2003, An Exploration of Tutors' Experiences of Facilitating Problem-Based Learning. Part 1. An Educational Research Methodology Combining Innovation and Philosophical Tradition. *Nurse Education Today*, 23, 58-64.

Hamilton, C., Parker, C., 1990, *Communicating for Results: a Guide for Business and the Professions*, Wadsworth Publishing Company, Belmont, California.

Horan, P., 2002, A new and flexible graphic organizer for IS learning: The rich picture, in E and A Zaliwska (eds), *IS 2002 Proceedings of Informing Science and IT Education Conference,* Informing Science: Informing Science Institute: http://ecommerce.lebow.drexel.edu/eli/2002Proceedings/2002. Retrieved Feb. 2006.

Hunt, J., 1979, *Managing People at Work - A manager's guide to behaviour in organizations*, McGraw-Hill, UK.

James, D.L., 1995, *The Executive Guide to Asia-Pacific Communications: Doing Business Across the Pacific*, Kodansha Press, New York.

Laszlo, E., 2001, Human Evolution in the Third Millenium. *Futures*, 33, 649-658.

Manage Effective Workplace Communication, 2004, http://ntcci.harvestroad.com.au/wppuser/ntcci/training/courses2004/work_communication.html. Retrieved Feb. 2006.

McShane, S. and Travaglione T., 2003, *Organisational Behaviour on the Pacific Rim*, McGraw-Hill. www.mhhe.com/au/mcshane

Messina, James, J., 1999, *Tools for Communication: A Model of Effective Communication,* http://www.coping.org/communi/model.htm, Retrieved Aug. 2005.

Myers, David G., 2004, *Psychology*, 7th ed., Worth Publishers, Inc.

Pease, B., Pease, A., 1999, Why men don't listen & women can't read maps : how we're different and what to do about it, Pease Training International, Mona Vale, NSW, Australia.

Rogers, C., and Farson, R., 1979, Active Listening, in Kolb, D., Rubin, I., and MacIntyre, J., *Organizational Psychology* (3rd ed.), New Jersey: Prentice Hall.

Samson, D., and Daft, R., 2003*, Management,* Pacific Rim Edition, Thomson.

Shannon, C.E. and Weaver, W., 1949, *The Mathematical Theory of Communication.* The University of Illinois Press, Urbana, 91pp.

Snooks, & Co., 2002, *Style Manual for Authors, Editors and Printers*, 6th edn, rev, AGPS, Canberra.

Steinmetz, Lawrence L., 1979, *Human Relations: People and Work*, Harper & Row.

Tidwell, C., 2005, *Intercultural Business Relations Course*, Class Notes 18, Non-verbal Communication modes, http://www2.andrews.edu/~tidwell/bsad560/NonVerbal.html, Retrieved Aug, 2005.

Turla, Peter, undated, Time Management Tips for Getting Results by Peter "The Time Man " Turla, www.TimeMan.com, Retrieved Aug. 2005.

Underwood, M., 2003, *Communication, Cultural and Media Studies Info-base*, http://www.cultsock.ndirect.co.uk/MUHome/cshtml/index.html, Retrieved 30 August 2005.

CHAPTER EIGHT

Economic Evaluation

This chapter deals with the economic evaluation of engineering projects. The purpose of economic evaluation is to identify the benefits and costs of a project and hence to determine whether it is justified on economic grounds. Alternative solutions to an engineering problem can also be compared based on their respective costs and benefits. It is usually necessary to discount the future benefits and costs of a project in order to make comparisons in terms of present value. Various economic criteria have been proposed for the comparison of projects. Each of these has its merits, although the best on theoretical grounds is net present value.

8.1 INTRODUCTION

An engineering project involves the transformation of limited resources into valuable final products or outputs. For example, the construction of a dam involves the use of resources such as concrete, steel, human effort, and machine time. The dam is used to produce outputs such as water for domestic and industrial purposes, flood control, and recreation.

The engineer must be concerned not only with the technical feasibility of the project (i.e. will it work?) but also the economic feasibility. *Economic evaluation* is aimed at assessing whether the value of the final products exceeds the value of the resources used by the project. The values of the outputs when measured in economic terms are called the *benefits* of the project. Similarly the values of the resources used in its construction and maintenance when measured in economic terms are called the *costs* of the project.

In the private sector, benefits and costs are measured by cash flows into and out of the firm, respectively. That is, a conceptual boundary is drawn around the firm and benefits and costs are represented by cash flows across this boundary.

For public sector projects, benefits and costs must be considered for society as a whole. In this case, benefits and costs are not necessarily associated with cash flows. For example, the benefits of a public transport system are not necessarily measured by the revenue which it generates. The government may choose to run the system at a loss for social or environmental reasons. In this case the benefits to the users of the system would probably exceed the revenue generated by it. If, for example, public transport were offered free of charge, few people would argue that the benefits to society were zero.

Sometimes there is a different perception of benefits and costs at various levels of government. For example, the federal government may provide a subsidy of 50% of the capital costs of new sewage treatment plants. When a state government is evaluating a particular plant it would consider this subsidy as a reduction in the cost of the project. On the other hand, the federal government would consider the subsidy as merely a transfer payment from one branch of government to another and therefore not a true benefit or cost to society as a whole. In fact, the federal government may require the state to provide economic justification for the project in its (i.e. the federal government's) terms without including the subsidy effect.

8.2 THE TIME VALUE OF MONEY

The costs and benefits (or revenue) of an engineering project usually occur over a long time period. For example, consider the construction of a new freeway. A typical time stream of benefits and costs is illustrated in Figure 8.1. Costs will be high during the construction phase which may last for three or four years. These costs could include disruption and inconvenience to users of the existing and adjoining roads. Annual maintenance and repair costs will be low initially, but will increase due to ageing of the pavement, bridges, and other components. The benefits to road users will be primarily due to savings in vehicle operating costs, savings in travel time and a reduction in the number of accidents in relation to the pre-existing road network. These benefits would normally increase with time due to increasing volumes of traffic using the freeway.

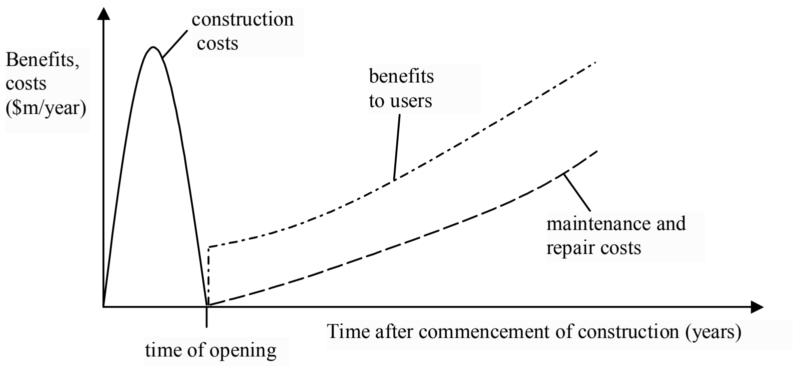


Figure 8.1 The time stream of benefits and costs for a freeway.

In carrying out an economic evaluation of such a project, it must be recognized that benefits or costs incurred in ten years cannot be directly compared with those incurred in the current year. Given the choice between $1000 now and $1000 in 10 years, very few people would choose the latter. In the first place, the effects of inflation are such that fewer goods and services could be purchased with the future sum than at present. However, even in the absence of inflation, human nature is such that there is a preference to consume goods now rather than later, and to postpone costs if possible. This is clearly evidenced by the fact that many individuals are willing to borrow money at an interest rate which exceeds the rate of inflation. Therefore, in carrying out economic evaluation, it is necessary to discount future benefits and costs in order to make them directly comparable with benefits and costs incurred now. This is carried out using discounting formulae derived from considerations of compound interest.

8.3 DISCOUNTING FORMULAE

In all the economic calculations that follow, there are two basic assumptions:

- there is a single rate of interest (the *discount rate*) that applies into the future for both borrowing and lending;
- interest is paid in a compound fashion; that is, if interest is earned, then it is added to the capital and taken into account for future calculations of interest.

Figure 8.2 Historical values of the standard variable home loan interest rate in Australia. (Source: Loansense, 2016)

If one looks at home loan interest rates and how they have varied over time in Australia (see Figure 8.2), it may be thought that the first assumption is questionable. However, despite the fact that rates do vary, it is usual for calculations to be carried out assuming an unvarying rate. If the rate does change at a later time, adjustments are made, again on the assumption of a constant rate into the future.

Present worth - lump sum

If a sum of money P is invested for one year at an interest rate of i, then after one year the value will be:

$$F = P(1+i) \tag{8.1}$$

where F is the future sum and P the present sum. For example, \$100 invested at a interest rate of 5 percent per annum (i = 0.05) for one year would yield \$105 at the end of the year. If that money is invested for a second year, then the value becomes:

$$F = P(1+i)(1+i) = P(1+i)^2 \tag{8.2}$$

After two years the initial \$100 yields \$110.25. After a period of n years it can be shown that the basic formula to calculate the future worth of a sum of money can be written as:

$$F = P(1+i)^n \tag{8.3}$$

where F is the future sum, P the present sum, i the interest rate and n the number of years (or to be more precise, the number of time periods over which the interest is paid). Some results for a simple situation are shown in Table 8.1.

Table 8.1 Future sums after a variable number of years with an interest rate of i and an initial sum of \$1000.

Years	Future return from \$1,000 invested for n years at various interest rates			
	2%	5%	8%	10%
1	1,020	1,050	1,080	1,100
2	1,040	1,103	1,166	1,210
5	1,104	1,276	1,469	1,611
10	1,219	1,629	2,159	2,594
100	7,245	131,501	2,199,761	13,780,612

It is convenient to show the process on a timeline (Figure 8.3) where the horizontal axis represents time and the vertical axis the units of money. While this appears trivial for the present example, once the situation becomes more complicated, the advantages of the time plot will become apparent.

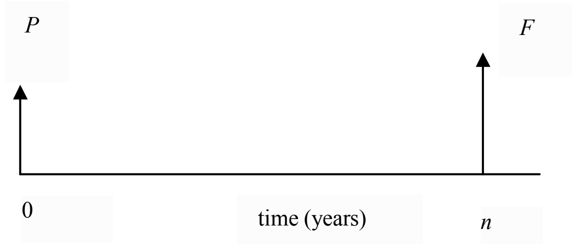


Figure 8.3 Timeline showing future value based on present value.

By rearranging Equation (8.3), it is possible to see how much must be deposited today to yield a certain sum of money after a given period.

$$P = \frac{F}{(1+i)^n} \tag{8.4}$$

where P is the present value of a lump sum of F, to be paid in n year.

Worked example

What is the present worth of $1000 that will be received in 10 years' time assuming a 10 % per annum (p.a.) compound interest rate?

Solution The timeline for the problem is shown in Figure 8.4. The present sum can be calculated from:

$$P = \frac{1000}{(1+0.10)^{10}} = \$385.54$$

The important point to grasp is that under the interest rate specified, $385.54 today and $1,000 in 10 years have equal worth. Alternatively, if $385.54 were placed in a bank account paying 10 % per annum. interest, the sum would accumulate to $1,000 in 10 years.

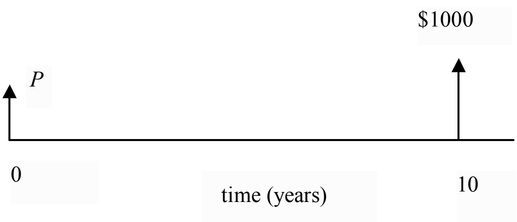


Figure 8.4 Timeline for the worked example.

Compound interest dates back to antiquity

The concept of compound interest (interest paid on interest) dates back to the ancient civilisations of Sumer and Babylon where it was usual to charge 20 % p.a. compound interest on loans of silver and 33 1/3 % p.a. compound interest on loans of barley.

Around 2400 BCE, the ruler of the Sumerian city of Lagash, Enmetena, engaged in a battle against the neighbouring city of Umma. Enmetena had a temple built to commemorate his victory and the details of the battle were recorded on the clay foundation stone of the temple. Also recorded on the foundation stone was the fact that Lagesh had loaned a large quantity of barley to Umma at 33 1/3 % interest p.a. The scribe had correctly calculated that the amount of barley to be repaid after 7 years was 7.5 times the original loan.

If the initial quantity borrowed was *X,* the future quantity owed can be calculated using Equation (8.3) as follows:

$$F = P\,(1+i)^n = X\,(1 + 0.33333)^7 = 7.492\,X$$

Source: Muroi, 2017

Present worth - uniform series

It is often necessary to find the present worth of a uniform series of annual payments. These may be regular mortgage repayments for a homeowner or the cost of leasing a particular piece of equipment for a manufacturer. The timeline for this is shown in Figure 8.5 where $\$A$ is paid at the end of each year for n years, and these payments are to be made equivalent to a single sum $\$P$ at the present time. Note that it is usual to assume that annual costs for maintenance or repair occur at the end of each year, whereas capital costs for new buildings or equipment occur at the start of the year.

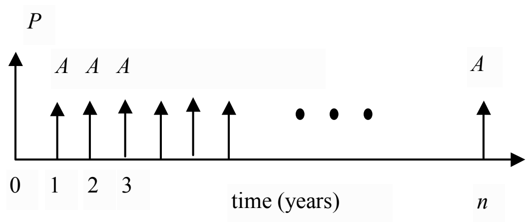


Figure 8.5 Timeline showing the equivalence between a present value, P, and a series of n annual payments of A.

Note that the first payment of $\$A$ is at the end of the first year. It is possible to discount the annual payments one by one. In this case:

$$P = \frac{A}{(1+i)} + \frac{A}{(1+i)^2} + \ldots \frac{A}{(1+i)^n} \tag{8.5}$$

However, by treating this as the sum of a geometric series, it is possible to write this in a more convenient way:

$$P = A\left[\frac{1-(1+i)^{-n}}{i}\right] \tag{8.6}$$

It is also possible to transpose the equation to determine what annual series is equivalent to have a present value of $\$P$:

$$A = P\left[\frac{i}{1-(1+i)^{-n}}\right] \tag{8.7}$$

where P is the present worth of a uniform series of payments of $\$A$ over a period of n years.

Worked example

A project is expected to yield \$3m a year in benefits for 30 years. What is the present value of the benefits assuming a discount rate of 8% ?

Solution Applying Equation (8.6):

$$P = 3\left[\frac{1-(1+0.08)^{-30}}{0.08}\right] = \$33.77 \text{ m}$$

Loan repayments and outstanding balances

When a loan is taken out, it is normal to repay the total debt over a number of years with repayments that are maintained at the same level over the whole period. This raises the problem of knowing how much is left to repay at any particular time. For example, in the case of a loan of \$100,000 taken out at a 7.5% annual interest rate over a 25 year period, how much is still owed at the end of 1 year or 10 years? In the following calculations it is assumed that repayments are made annually. A similar procedure may be applied to monthly or weekly repayments. The annual repayment on the loan can be calculated from Equation (8.7):

$$A = 100{,}000\left[\frac{0.075}{1-(1+0.075)^{-25}}\right] = \$8{,}971.07$$

To determine the outstanding debt after x years, it is necessary to consider the present value of the annual repayments that remain to be paid over a period of $(n - x)$ years. Therefore, the outstanding debt after 1 year (P_1) is a series of annual payments over 24 years. In this case:

$$P_1 = 8971.07 \frac{1-(1+0.075)^{-24}}{0.075} = \$98{,}528.96$$

Therefore, after a single annual repayment of $8,971.07 there is still $98,528.96 left to repay, a debt reduction of only $1,471.04. The other $7,500.34 has gone in interest repayments. As time goes by, the total repayment remains the same, but interest component gradually reduces while that going to reduce the outstanding debt gradually increases. This is shown in Figure 8.6 for the current example.

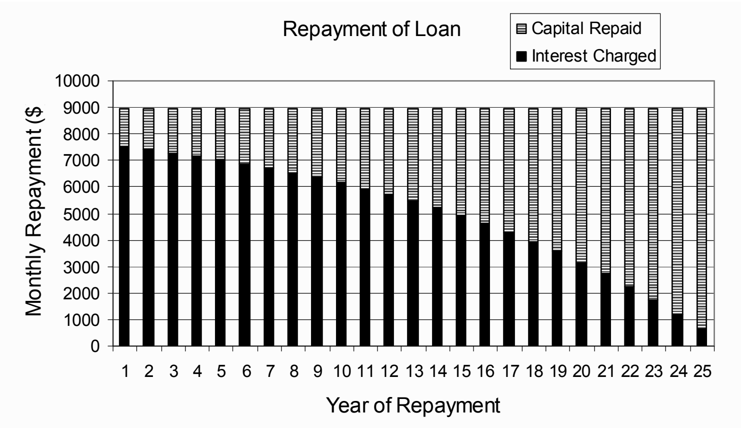


Figure 8.6 The split between interest (solid) and capital repayment (hatched) of each annual payment is shown. Note how the interest component gradually reduces over time.

After 10 years, 15 annual payments each of $8,971.07 remain to be paid, so the remaining balance of the loan (P_{10}) can be determined as follows:

$$P_{10} = 8971.07 \frac{1-(1+0.075)^{-15}}{0.075} = \$79{,}188.71$$

To illustrate the effect that more frequent payments have on the calculations the same problem is repeated for monthly repayments for the same annual interest rate. An annual interest rate of 7.5% equates to 0.625% per month and the monthly repayment (M) can be calculates as:

$$M = 100{,}000 \frac{0.00625}{1-(1+0.00625)^{-300}} = \$738.99$$

Over one year the total annual repayment is calculated from 12 monthly payments and is equal to $8867.89 (a slight increase on the previous figure). After one year there will be 288 repayments left to make leaving a remaining balance (P_{1m}) given by the following calculation:

$$P_{1m} = 738.99 \frac{1-(1+0.00625)^{-288}}{0.00625} = \$98{,}583.93$$

The balance still owed after one year is $98,583.93 which is slightly higher than for annual repayments.

The subject of different repayment intervals will be dealt more formally in the next section where it will be seen that it is possible to determine a relationship between the same interest rate compounded annually and monthly.

Effective interest rate

Interest rates are generally quoted as a percentage per annum (i.e per year), but the compounding period (the frequency with which the interest is charged or paid) is often less than a year. As we have just seen, this makes a difference to the calculations, and in order to allow a valid comparison the concepts of a nominal interest rate and an effective interest rate have been developed. If the nominal interest rate is r percent per annum and it is charged in m periods over a year then the interest rate applied at each period is (r/m). For example, an annual rate of 18% charged monthly means that the interest charged each month is 1.5%. Therefore if \$$P$ is invested at the start of the year the balance at the end of m periods can be calculated using Equation 8.3 as:

$$F = P(1+r/m)^m \tag{8.8}$$

If this monthly rate is to be quoted as an equivalent annual rate, i, then it must produce the same future sum at that equivalent annual rate:

$$F = P(1+r/m)^m = P(1+i) \tag{8.9}$$

so the equivalent annual rate can be calculated as:

$$i = (1+r/m)^m - 1 \tag{8.10}$$

Worked example

A credit agency claims to charge 18% p.a. but interest is calculated monthly. What is the effective rate of interest?

Solution By applying Equation (8.10) it is possible to determine the effective annual rate of interest.

$$i = (1+r/m)^m - 1 = (1+0.18/12)^{12} - 1 = 0.1956 = 19.56\% \text{ p.a.}$$

Hence, 18% per year charged monthly is equivalent to 19.56% per year charged annually. Note that in a number of countries, financial institutions must publish their interest rates as an equivalent annual rate so that consumers can make informed choices.

Interest and the natural logarithm

If the annual interest rate is 100% and $1 is invested for 1 year, then the way interest is charged (annually, monthly, etc.) can be seen to change the outcome considerably. If charged annually, the amount invested at the end of the year will be $2. If it is paid six monthly, the amount will be $2.25. These, and other possibilities, climaxing in continuous payment of interest are

Interest is paid ...	Final Sum
Annually	$2.00
Six Monthly	$2.25
Quarterly	$2.44
Monthly	$2.61
Weekly	$2.69
Daily	$2.714567
Hourly	$2.718127
Continuously	$2.71828182845 (*e*, the base of the natural logarithm)

The effect of inflation on economic evaluation

Inflation is represented by a rising general level of prices. It is important to note that during an inflationary period the prices of all goods and services are not necessarily rising nor are all prices necessarily rising at the same rate. The effect of inflation is that a dollar next year will buy less than a dollar today. There have also been periods in history when the general level of prices in some countries or regions decreased. This is called *deflation*.

Changes in the general level of prices may be measured by a price index. This is defined as 'the average price of a mixture of goods and services in a given year' divided by 'the average price of the same mixture of goods and services in a base year' (expressed as a percentage). A common price index is the consumer price index or CPI. The CPI measures changes in the prices of goods and services purchased by moderate-income families. It is based on the prices of a large number of representative items each weighted according to its relative importance in the typical family budget. Movements in the consumer price indexes for a number of countries are shown in Table 8.2.

Of particular relevance to engineering projects are the price indexes for construction and building costs. These indexes for a number of North American cities are available from the Engineering News Record website (Engineering News Record, 2017). In Australia, current and historical values of the building cost index can be obtained from Rawlinsons (2016).

It is important to distinguish clearly between the concepts of inflation and the time value of money. Even in the absence of inflation, very few people would be prepared to lend money at zero interest. As most people would prefer to consume goods and services now rather than later, there is clearly a real time value of money. In times of inflation, lenders expect a real rate of return on their money which exceeds the inflation rate. For example, anyone who receives 3 percent p.a. interest on an investment when the inflation rate is 5 percent p.a. is clearly losing money.

Table 8.2 Consumer price index for a selection of world countries. Base year (100.0) = 2010. Adapted from World Bank (2016)

Country	2013	2014	2015
Australia	107.7	110.4	112.0
Brazil	119.4	126.9	138.4
Canada	105.5	107.5	108.7
China	111.2	113.4	115.0
Germany	105.7	106.7	106.9
India	132.0	140.4	148.6
Indonesia	116.9	124.4	132.3
Japan	100.0	102.8	103.6
Korea	107.7	109.1	109.8
New Zealand	106.3	107.6	107.9
Russia	121.6	131.2	151.5
South Africa	117.5	124.7	130.3
Switzerland	99.3	99.3	98.2
United Kingdom	110.1	111.8	111.8
United States of America	106.8	108.6	108.7

Before we consider how inflation is taken into account in economic evaluation, some definitions are required.

Actual dollars (or *then-current dollars*) are dollars which are current in the particular years that the benefits and costs occur. For example, if a person bought a block of land in 1970 for \$50,000 (in then-current dollars) and sold it in 2000 for \$200,000 (in then-current dollars) have they made a profit on the investment? The answer is yes if \$200,000 in 2000 will buy more goods and services than \$50,000 in 1970. Otherwise, they have lost on the investment.

Constant-worth dollars are dollars which have the same purchasing power at a defined point in time, e.g., 2010 dollars. For example, a new power station is designed to provide 8000 GWh of electrical energy per year under certain operating conditions. The benefits to consumers of this energy are estimated to be \$600m in 2010 (the first year of operation). Thereafter, the station will continue to provide 8000 GWh of energy per year for its estimated operating life of 40 years. The annual benefits expressed in actual dollars will continue to increase throughout the life of the station due to inflationary increases in the price of electricity and the prices of goods produced with it. However, when expressed in constant-worth dollars (e.g., 2010 dollars) the annual benefits are likely to be constant over the life of the station if the same amount of energy is produced each year.

The *nominal interest rate* is the rate received on invested money (or paid on borrowed money) when calculated in terms of actual (or inflated) dollars. The *real* *rate of interest* on an investment is the rate received net of inflation. For example, if money is invested at a nominal interest rate of 8% p.a. when the annual inflation rate is 5% p.a., the real interest rate (i.e., above inflation) is approximately 3% p.a.

The economic evaluation of engineering projects can be taken into account in one of two equivalent ways:

- by using actual dollars and the nominal interest rate throughout;
- by using constant-worth dollars and the real rate of interest throughout.

The first method is commonly used in evaluating private sector investments in which the future cash flows are estimated in actual dollars and the nominal interest rate is usually known. The second method is more commonly used for evaluating public sector investments, as it is often easier to estimate future benefits and costs in constant-worth dollars. In this case there is no need to estimate an inflation rate. The interest rate or discount rate used is the real rate above inflation. As will be discussed in Section 8.10, the choice of the discount rate in the public sector is the source of some controversy. In this book, constant-worth dollars and real interest or discount rates will be used unless otherwise stated.

Relationship between nominal and real interest rates

Consider the annual maintenance costs of a bridge which requires a constant number of person-hours of maintenance each year during its life of n years. (This is a gross simplification, as it would be normal for the amount of maintenance to increase as the bridge deteriorates with age.) When expressed in constant-worth dollars, the annual maintenance cost of the bridge is a constant amount (say $\$A$) per year for the life of the bridge, as shown in Figure 8.7.

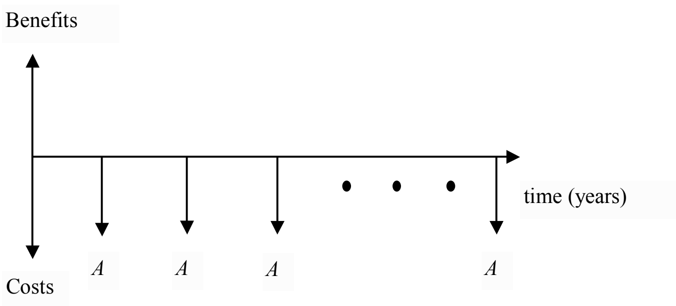


Figure 8.7 Bridge maintenance costs in constant-worth dollars.

If the anticipated inflation rate for the life of the project is f % p.a. (compound), then the maintenance costs in actual dollars are shown in Figure 8.8. The nominal rate of interest is assumed to be known and is denoted by i_n. The objective is to find a relationship between the real interest rate, i, the nominal interest rate, i_n, and the inflation rate, f. The present worth of the series shown in

Figure 8.8 is denoted by *P*. It can be determined by applying Equation (8.4) to each annual payment using actual dollars and the nominal interest rate:

$$P = \frac{A(1+f)}{(1+i_n)} + \frac{A(1+f)^2}{(1+i_n)^2} + \cdots \frac{A(1+f)^n}{(1+i_n)^n} \tag{8.11}$$

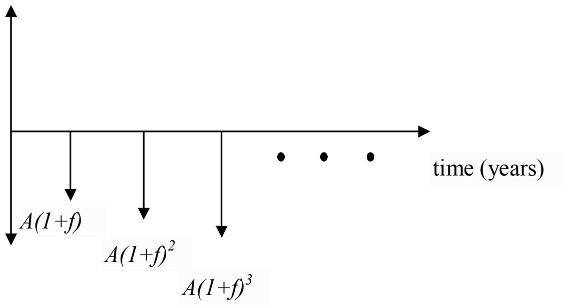


Figure 8.8 Bridge maintenance costs in actual dollars (*f* = annual inflation rate).

The right-hand side of Equation (8.11) can be simplified, as it is the sum of a geometric series:

$$P = \frac{A(1+f)}{(1+i_n)} \left\{ \frac{1-(1+f)^n/(1+i_n)^n}{1-(1+f)/(1+i_n)} \right\} \tag{8.12}$$

Now, the present worth of the series shown in Figure 8.7 is found using constant-worth dollars and the real interest rate, *i*. Rewriting Equation (8.5):

$$P = \frac{A}{(1+i)} + \frac{A}{(1+i)^2} + \cdots \frac{A}{(1+i)^n} \tag{8.13}$$

It should be apparent that the two equations (Equations (8.11) and (8.13)) are equivalent if:

$$(1+i) = \frac{(1+i_n)}{(1+f)} \tag{8.14}$$

Therefore, the real interest rate can be calculated using the following equation:

$$i = \frac{(1+i_n)}{(1+f)} - 1 \tag{8.15}$$

Worked example

In Kenya in 1998 the nominal rate of interest for borrowing was 21% and the annual inflation rate 15% (Adeoti et al., 2000). What was the real rate of interest under these conditions?

Solution The real rate of interest, i, can be found by applying Equation (8.15):

$$i = \frac{(1+i_n)}{(1+f)} - 1 = \frac{1+0.21}{1+0.15} - 1 = 0.052 \text{ or } 5.2\% \text{ p.a.}$$

8.4 EVALUATION CRITERIA

Various economic criteria have been proposed for use in comparing engineering projects. The following criteria are described briefly and compared:

- payback period;
- net present value;
- equivalent annual worth;
- benefit-cost ratio; and
- internal rate of return.

Payback Period

The payback period is defined as the time it takes for the project to generate sufficient net benefits to cover its initial cost of construction and implementation. For example, if a project has an initial capital cost of $10m, annual benefits of $3m and annual operating costs of $1m, the payback period can be calculated as:

Net annual benefits = $3 m - $1 m = $2 m
Payback period = $10 m / $2 m = 5 years

Net Present Value

The net present value (NPV) of a project is defined as the present value of all benefits minus the present value of all costs. It can be written as:

$$NPV = PVB - PVC \qquad (8.16)$$

where *PVB* is the Present Value of Benefits and *PVC* is the Present Value of Costs. By using the formulae for *PVB* and *PVC* we can write:

$$NPV = \sum_{t=0}^{n} \frac{B_t}{(1+i)^t} - \sum_{t=0}^{n} \frac{C_t}{(1+i)^t} \qquad (8.17)$$

where B_t is the benefit in year t, C_t is the cost in year t, i is the discount (or interest) rate and n is the life of the project (years). The calculation of NPV therefore involves discounting all costs and benefits back to the present time to allow them to be compared at the same point in time. It is important to note that while Equation (8.17) is mathematically correct, it is unlikely that it would be the way the actual calculation will be carried out. For example, if the costs are a series of annual payments, Equation (8.6) would most likely be used to discount them back to the present time.

Payback period case studies

Best (2000) quotes a payback period of 4 years for the lighting and heating of a building for Lockheed in the 1980s where natural light was used in place of electric lights, and where this also reduced the need for air conditioning. It was further stated that productivity gains due to the healthier environment in the building gave the company a competitive advantage in its bid for a large contract, the profit from which would have paid for the entire building.

Morgan and Elliot (2002) used payback period to justify the implementation of a mechanical mixing system for a water supply reservoir. The motors used only 8.5% of the power required for a traditional aerator plant and had an estimated payback period of 4 years based on direct savings in energy costs.

The price of solar cells has been reported (*The Australian* - 8/8/01) as falling to 20% of that of 25 years ago. Rooftop systems that can meet half a home's electricity needs for more than 20 years now cost as little as US$10,000. This was quoted as having a five- to six-year payback period in California compared with 20 years a few years ago.

Equivalent annual worth

The NPV brings all benefits and costs to the present time and gives the answer as a single present day sum of money. There may be an argument for representing the project in terms of its net annual earnings over the period of the project. In this case the Equivalent Annual Worth (EAW) may be more appropriate. The EAW is found by converting all project benefits into a series of constant annual benefits for n years, where n is the life of the project. Similarly all project costs are converted into a series of constant annual costs for n years. The equivalent annual worth is then equal to the equivalent annual benefit minus the equivalent annual cost. A reliable method for calculating EAW is to determine it from the NPV applying Equation (8.7) which gives:

$$EAW = NPV\left[\frac{i}{1-(1+i)^{-n}}\right] \tag{8.18}$$

Benefit-cost ratio

The benefit-cost ratio (B/C) is widely used for evaluating projects in the public sector. As normally defined, it is the present value of all benefits divided by the present value of all costs:

$$B/C = \frac{PVB}{PVC} \tag{8.19}$$

where PVB is the Present Value of Benefits and PVC is the Present Value of Costs. It can be calculated as:

$$B/C = \frac{\sum_{t=0}^{n} \frac{B_t}{(1+i)^t}}{\sum_{t=0}^{n} \frac{C_t}{(1+i)^t}} \tag{8.20}$$

where the terms are as defined previously.

Again, it is worth noting that while Equation (8.20) is mathematically correct, it is unlikely that it would be the way the actual calculation would be carried out. It is much more likely that the benefits and costs would be discounted back to present worth using appropriate formulae for lump sums or uniform annual series.

Benefits and costs of natural and nature-based coastal defences

Narayan et al. (2016) studied the coastal protection benefits and costs of 52 projects that involved restoration of natural systems including coral reefs, oyster reefs, salt marshes and mangroves. The benefits were based on the reduction of wave heights and included savings in erosion damage costs, savings in damage costs from storms and savings in costs of adjacent coastal structures. They found that 41% of the mangrove projects and 6% of the salt marsh projects had a benefit-cost ratio greater than one.

Furthermore, Narayan et al. (2016) compared the cost of these restoration projects with the cost of constructing submerged reefs at the corresponding locations. The dimensions of the reefs were chosen to give the same reduction in wave heights. They estimated that the ratio of the cost of a submerged reef to the cost of nature-based defences averaged 5 for the mangrove projects and 2 for the salt marshes.

Internal rate of return

The economic criteria discussed so far (with the exception of the payback period) involve the use of a discount rate to convert all future benefits and costs to present value. The internal rate of return (also called "yield") is a different concept in that it is the implied interest rate of the investment. The internal rate of return is the discount rate at which the present value of benefits just equals the present value of costs over the life of the project. The internal rate of return (IRR) of a project, *r*, can be determined by finding the discount rate at which the present value of the benefits equals the present value of the costs:

$$PVB = PVC \tag{8.21}$$

This can also be written as:

$$\sum_{t=0}^{n} \frac{B_t}{(1+r)^t} = \sum_{t=0}^{n} \frac{C_t}{(1+r)^t} \tag{8.22}$$

where the terms are as defined previously. In general this will involve the solution of a polynomial in order to determine *r*.

Internal rate of return case studies

Campus Review (2002) reported results from an OECD evaluation of the benefits of a university education. It was found that the private rate of return to the individual was about 11% p.a. and the return to society ranged between 6% and 15% p.a. This was based on higher average earnings and lower risks of unemployment. It was suggested that the rates were higher than conventional investments and therefore money spent on a university education was money well spent.

8.5 A COMPARISON OF THE EVALUATION CRITERIA

There are three basic types of decisions requiring economic evaluation. They are:

1. go or no-go decisions, for example, whether or not to build a new airport;
2. the choice of a single project from a list of mutually exclusive projects, for example, the choice between a bridge and a tunnel for crossing a major waterway; and
3. the choice of a number of projects when the total funds available are limited, for example, how to allocate an annual capital works budget between feasible projects.

Go or no-go decisions (type 1 decisions)

In deciding whether or not an individual project is justified on economic grounds, any one of the following criteria may be applied:

$$NPV > 0 \quad \text{or} \quad EAW > 0 \quad \text{or} \quad B/C > 1 \quad \text{or} \quad IRR > i$$

Only one of these needs to be looked at since they will all lead to the same decision. For example, if the present value of benefits exceeds the present value of costs, then *NPV* will be positive, *EAW* will be positive, *B/C* will exceed 1 and the *IRR* will exceed the discount rate.

Choice from a list of mutually exclusive options (type 2 decisions)

Often economic decisions are used to choose from a number of different options where only one will be chosen. For example, what sort of car to buy, or which laptop computer to purchase. In each case, one will be chosen in preference to the others. For example, consider two air-conditioning systems that are identical in performance but differ in their costs. A summary of relevant information is given in Table 8.3. For this example, an equivalent annual benefit has been added to make the comparison more interesting. A discount rate of 12% p.a. will be used in the comparison. The question is: which is the better system on economic grounds?

Table 8.3 Costings for two air-conditioning systems. (Adapted from Rawlinsons, 2004). Note: salvage value is not available at end of project.

Characteristic	System X	System Y
Capital cost	\$115,600	\$158,800
Life of plant	10 years	15 years
Annual costs	\$37,800	\$28,200
Annual benefits	\$80,000	\$80,000
Salvage value	\$3,000	\$7,000

As the two systems have different lives, it is important to undertake the economic evaluation over an appropriate time period. If 10 years is used, System X will be due for salvage at the end of the period, whereas System Y will have 5 years life remaining. If, on the other hand a period of 15 years is used, System Y will be due for salvage at the end of the period, whereas System X will need to be replaced after 10 years and the replacement system will have 5 years left at the end of the period of analysis, The basic principle is to use a period of analysis that is the lowest common multiple of the lives of the various alternatives. In this case this is 30 years, as this allows for the initial purchase plus exactly two replacements of System X and the initial purchase plus exactly one replacement of System Y. The exception to this would be if the company plans to close operations after a certain period (e.g., a mine because it is expected to run out of ore). In this case the period of operations should be used for the comparison.

Having chosen 30 years as the period of analysis, a timeline is drawn for the whole 30-year period for System X (see Figure 8.9). Note that since the plant only lasts 10 years, it will have to be replaced twice. If System X is the better choice initially, it will also be the preferred choice after 10 years and 20 years. In the absence of better information, it is assumed that the original prices are the best estimate for the values to be used in subsequent purchases.

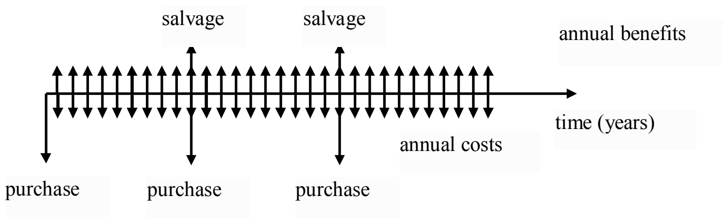


Figure 8.9 The timeline for System X showing the purchase and salvage of three units, each lasting 10 years together with the annual costs.

The Present Value of the Costs can be calculated as:

$$PVC = 115600 + \frac{115600}{1.12^{10}} + \frac{115600}{1.12^{20}} + 37800\frac{1-1.12^{-30}}{0.12} = \$469{,}289.94$$

The Present Value of the Benefits can be calculated as:

$$PVB = \frac{3000}{1.12^{10}} + \frac{3000}{1.12^{20}} + 80000\frac{1-1.12^{-30}}{0.12} = \$645{,}691.64$$

Therefore, $NPV = PVB - PVC = \$176{,}401.70$

The Internal Rate of Return is calculated by solving for the rate, r, such that:

$$115600 + \frac{115600}{(1+r)^{10}} + \frac{115600}{(1+r)^{20}} + 37800\frac{1-(1+r)^{-30}}{r} =$$
$$\frac{3000}{(1+r)^{10}} + \frac{3000}{(1+r)^{20}} + 80000\frac{1-(1+r)^{-30}}{r}$$

The solution is $r = 0.347$, that is 34.7%.

The same calculations are now carried out for System Y over the whole 30-year investment period (see Figure 8.10). Since System Y has a longer life, it will only have to be replaced once as is evident from Figure 8.10.

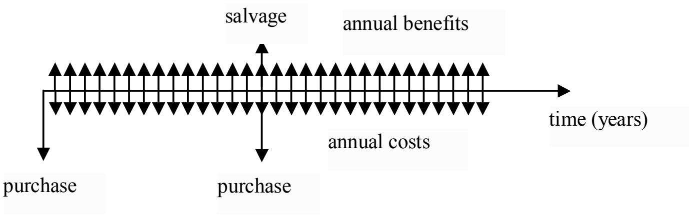


Figure 8.10 The timeline for System Y showing the purchase of two units, each lasting 15 years together with the annual costs and salvage benefits.

$$PVC = 158800 + \frac{158800}{1.12^{15}} + 28200\frac{1-1.12^{-30}}{0.12} = \$414{,}968.35$$

$$PVB = \frac{3000}{1.12^{15}} + 80000\frac{1-1.12^{-30}}{0.12} = \$644{,}962.81$$

$$NPV = \$229{,}994.45$$

The Internal Rate of Return is solved in a similar fashion to the previous value. The solution is $r = 0.321$ or 32.1%. Based on the calculations it is possible to list the various economic indicators as shown in Table 8.4 for the two systems.

Table 8.4 Comparison between System X and System Y.

Economic Criterion	System X	System Y
NPV	$176,401	**$229,994**
B/C	1.38	**1.55**
IRR (%)	**34.7**	32.1

It is evident that not all indicators agree on the better system, but, economists support the use of the NPV criterion for this type of decision. Hence System Y should be selected despite having an inferior Internal Rate of Return.

A variation on the above decision-making situation is when one design must be chosen from among a number of alternatives, each of which performs substantially the same function. For example, the choice between two designs for a road bridge, one of which has prestressed concrete girders and the other steel girders. In both cases, the girders act compositely with a reinforced concrete deck. The benefits of the bridge may be assumed to be the same for both designs and may therefore be ignored in the decision. In this case the NPV criterion is equivalent to finding the design which minimises the present value of costs (assuming that the present value of benefits exceeds the present value of costs for the chosen design). Alternatively, the choice may be made by minimising the equivalent annual cost. The B/C and IRR criteria cannot be used in this situation.

Capital budgeting problems (type 3 decisions)

It will be the case for most organisations that the cost of the full list of desirable projects exceeds the available funds and a choice has to be made regarding which projects will proceed in the current budget round and which will not. In the previous section, it was found that the different methods of economic comparison (B/C ratio, NPV and IRR) can give different results when used to choose one project from among a list of options. This problem will also occur in capital budgeting.

Capital budgeting using B/C or IRR

When making a capital budgeting decision based on the use of the benefit cost ratio or the internal rate of return, the general procedure is to rank the projects from best to worst in terms of the criterion. Projects are then chosen from the top down until the budget is exhausted. The method is demonstrated with an example.

Worked example

A city council is attempting to plan its capital works for the coming year and must decide which projects should be selected using B/C ratio. The available budget is \$35 m. A list of possible projects is given in Table 8.5. It is assumed that all of the costs are capital costs for this example.

Table 8.5 Potential projects, their benefits and costs.

Project	*PVB* (\$m)	*PVC* (\$m)	*NPV* (\$m)	*B/C*
A	40	15	25	2.67
B	18	10	8	1.80
C	18	12	6	1.50
D	50	20	30	2.50
E	28	10	18	2.80
F	26	20	6	1.30

Solution Using B/C, the projects are ranked in order as shown in Table 8.6 and selected off the top until the budget is exhausted. Selecting from the top gives Project E with a cost of \$10 m. This leaves \$25 m to spend. Therefore, Project A can also be selected as it has the next highest B/C ratio and a cost of \$15 m. This leaves \$10 m in the budget. Project D is next on the list in terms of B/C, but its cost of \$20 m exceeds the \$10 m that remains in the budget, so the search continues. Project B is next and costs \$10 m which exactly exhausts the budget. Therefore, the projects selected are E, A and B with a total PVC of \$35 m and a NPV of \$51 m.

Table 8.6 Potential projects, their benefits and costs sorted in order of B/C ratio.

Project	PVB($m)	PVC($m)	NPV($m)	B/C
E	28	10	18	2.80
A	40	15	25	2.67
D	50	20	30	2.50
B	18	10	8	1.80
C	18	12	6	1.50
F	26	20	6	1.30

Capital budgeting using NPV

When using NPV, the procedure is a little more complicated in that a set of projects must be chosen which maximise NPV, subject to the available capital budget. Therefore, it is necessary to set up an optimisation model which is, in fact, an integer linear programming problem. This will be demonstrated using the same example above.

To set up the problem in a standard format, the first step is to define variables X_i where i = A, B, C, D, E, F such that if $X_i = 1$ the project is selected, and $X_i = 0$ if not. The aim is to maximise the NPV so a value Z is defined which will be the total NPV over all the projects:

$$\text{Maximise } Z = 25X_A + 8X_B + 6X_C + 30X_D + 18X_E + 6X_F$$

The cost of all projects selected must not exceed the available budget so the following constraint applies:

$$15X_A + 10X_B + 12X_C + 20X_D + 10X_E + 20X_F \leq 35$$

and $X_i = 1$ or 0 (binary variables) for i = A, B, C, D, E, F.

The solution of integer linear programming problems is beyond the scope of this book. The solution of linear programming problems with continuous variables is discussed in Chapter 13. For simple problems, the solution can be found by inspection or a limited trial and error procedure as:

$$X_D = X_A = 1 \text{ with all other } X_i = 0.$$

That is, Projects A and D should be undertaken, but not the others. In this case the adopted projects have a PVC of $35m and NPV of $55m. This NPV is $4m higher than for the projects chosen using B/C. In general it can be shown that NPV is a better criterion than B/C for capital budgeting projects. It can also be shown that IRR is inferior to NPV in capital budgeting.

World Bank programs and projects

According to Talvitie (2000), when the World Bank lends money for infrastructure programs (e.g., road development) in the form of individual projects in developing countries, it does so with the aim of reducing poverty, but is also mindful of ensuring that projects are well engineered and technically feasible. For major road projects, steps are taken to safeguard the environment, cultural heritage, indigenous people, and endangered species, and in fact approximately 50% of project preparation costs are spent on these issues while the design and traffic engineering component may only be 20–25%.

In terms of selecting actual projects, Talvitie states that the project priority setting and program development and evaluation steps occur together rather than simply picking projects off the top of the list until the funds are exhausted; an approach which may lead to a less-than-optimal mix of projects.

8.6 ADVANTAGES AND LIMITATIONS OF EACH CRITERION

From the previous section it should be clear that each economic criterion may lead to a different choice of project(s) when choosing one or more from a list of feasible alternatives. It is important therefore to consider the advantages and limitations of each criterion in turn.

Payback period

Although the payback period is often quoted in the press, there are various problems with the method. Firstly, the criterion does not deal with the time value of benefits and costs. Secondly, and equally importantly, the method may ignore significant benefits or costs that occur after the payback period. On the other hand, the advantage of the method is that it is an intuitively appealing concept and can make a compelling argument for implementing a project.

Payback period has also been used in other aspects of economic decision making. For example, in an assessment of the economic viability of a biogas project for Nigeria, Adeoti et al. (2000) promoted the payback period as a measure of project riskiness.

However, because of its limitations, payback period is not recommended for use in the evaluation of major engineering projects.

The internal rate of return

This criterion has the advantage of not requiring a prescribed value of discount rate for its determination. As the specification of a discount rate may be the source of considerable controversy, particularly in the public sector, this is an advantage in project selection. However, this advantage is outweighed by the disadvantages of the criterion, namely:

- to determine the IRR involves the solution of a polynomial equation;
- sometimes more than one value of IRR may be obtained for a particular project.

Worked example

Figure 8.11 shows the time stream of expected revenue and costs for a proposed open-cut mining development. There is an initial cost to establish the mine and associated works and a high cost at the end of the project due to clean-up and re-vegetation activities. What is the IRR for the project?

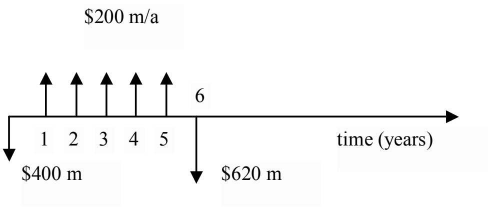


Figure 8.11 Timeline showing costs and benefits for open cut mining development.

Solution Let i = the discount rate of the project. Then the Net Present Value can be calculated as:

$$NPV = 200\frac{1-(1+i)^{-5}}{i} - \frac{620}{(1+i)^6} - 400$$

Figure 8.12 shows a plot of NPV as a function of i.

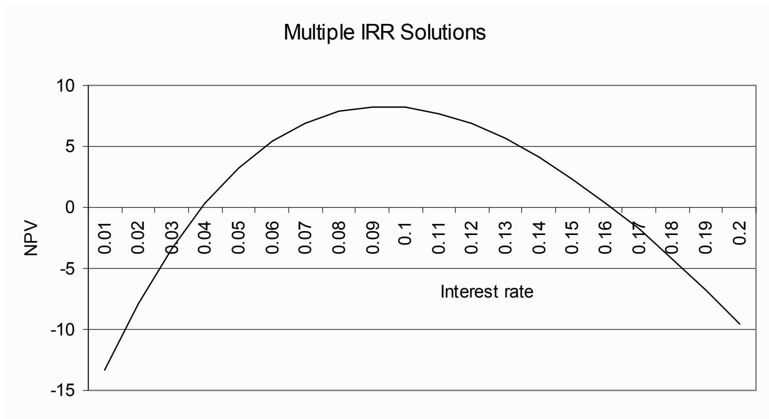


Figure 8.12 Plot of NPV versus interest or discount rate. The IRR is where NPV = 0.

It can be seen that there are two values of *i* for which the NPV is zero, namely 3.8% and 16%. Both of these values may be interpreted as the internal rate of return of the project. If the project were being ranked in a capital budgeting exercise it could receive a high or low place depending on the value of IRR used. Although this is a somewhat unusual case, it does demonstrate one difficulty with using the IRR criterion, i.e., it is the solution of a polynomial equation for which more than one root may exist.

Equivalent annual worth

To determine the EAW of a general time series of benefits and costs, it is necessary first to compute the NPV of the series and convert the NPV into a uniform annual payment throughout the life of the project. Considering that EAW gives the same ranking as NPV for projects with equal lives, a strong case can be made for using NPV directly.

Benefit-cost ratio

The benefit-cost ratio has been used extensively in a number of countries for comparison of projects of various sizes. It has some degree of appeal when used in the context of allocating scarce capital funds, i.e., capital budgeting. However, there is some ambiguity in its definition for projects with high recurrent costs (e.g., operating, maintenance, and repair costs). The basic issue here is whether the present value of recurrent costs should be treated as negative benefits and deducted from the numerator of the B/C ratio, or treated as costs and added to the denominator. A different value of B/C ratio will result in each case. For example, consider the economic comparison of two projects with the relevant economic data summarized in Table 8.7.

Table 8.7 Potential projects, their benefits and costs.

Economic Parameter	Project 1	Project 2
PV of benefits B ($m)	1.5	4.0
Construction costs C ($m)	1.0	1.0
PV of OMR costs K ($m)	0.0	2.0
B/C No. 1 = B/(C + K)	1.50	1.33
B/C No. 2 = (B - K)/C	1.50	2.00
NPV = B - K - C	0.5	1.0
OMR cost = operating, maintenance and repair costs		

For B/C Ratio No.1 in Table 8.7, the present values of all costs are included in the denominator. For B/C Ratio No.2, the present values of recurrent costs are subtracted from the present value of benefits in the numerator. The rationale for this latter definition is that in many cases it is really the capital costs which are limited; the recurrent costs may be met from the annual benefits.

If the projects are compared using the first definition of B/C ratio, Project 1 is preferable. Conversely, if the second definition of B/C ratio is used, Project 2 is preferable. On the other hand, NPV always gives an unambiguous ranking of projects, as it does not matter whether OMR costs are subtracted from the benefits or added to the construction costs. In this case, the NPV criterion favours Project 2.

It might be argued that this is not a major problem; why not just define the B/C ratio in one way or the other and proceed? However, in many situations the project benefits are composed entirely of savings in cost. For example, the benefits of most road projects consist of savings in travel time, savings in operating cost, and savings in cost due to a reduction in accidents. In such cases the B/C ratio obtained depends on the arbitrary definitions of benefits and costs, whereas NPV is unambiguous.

Net present value

Most economists advocate NPV as the most appropriate criterion for comparing projects. It measures the net gain to society (or the company) of the proposed project. It can be used in any of the three decision-making contexts discussed above. In the case of a go or no-go decision, the choice is independent of the criterion used. However, in choosing a single project from a list of mutually exclusive projects, NPV appears to favour large projects.

Worked example

An engineering firm is considering the purchase of a multi-core computer since it currently pays for the use of a super computer on a time-sharing basis. Two models of multi-core computer are being considered. The relevant economic data are given in Table 8.8. Which option should be chosen?

Table 8.8 Potential projects, their benefits and costs.

Economic data	System A	System B
Purchase Price ($)	10,000	20,000
Annual benefits	4,000	8,000
Life (years)	5	5
PV of benefits ($)	18,914	37,828
NPV ($)	8,914	17,828
EAW ($/year)	1,226	2,452
B/C	1.89	1.89
IRR (%)	28.65	28.65

Solution The benefits represent the estimated savings in computing cost due to reduced usage of the super computer. From Table 8.8 it can be seen that System B has the same life as System A, but twice the cost and twice the benefits. Its NPV and EAW are therefore twice those of System A, while its B/C ratio and IRR are the same. The tendency for NPV (and EAW) to favour large projects is apparent. If there is no constraint on the amount of money that can be spent, the increased NPV of System B is a net gain to the firm and therefore it is the better option. If there are limited funds available for investment due to other equipment needs, the problem should be treated as part of a capital budgeting exercise.

8.7 ECONOMIC BENEFITS

One important issue which needs to be addressed in economic evaluation is the estimation of the benefits and costs of any particular project. In order to understand the basis for this estimation some concepts from microeconomic theory can be drawn upon. The theory of price determination in perfectly competitive markets is presented in Appendix 8A in this chapter.

A fundamental distinction should be drawn between the estimation of benefits in the private and public sectors. In the private sector, the benefits (or revenue) of a particular project can be identified as *cash flows* into the firm. Benefits may be determined by the additional output of goods and services multiplied by the market prices for these commodities. In a competitive market situation, each individual firm is obliged to sell its output at the ruling market price due to pressure from its competitors. In this situation, the output of any one firm is so small that it has a negligible impact on the market price. Furthermore, the market price is a measure of the willingness-to-pay of the marginal consumer of that product. Therefore, the market price is a measure of the marginal benefit to society of providing additional units of the commodity.

In the public sector, the basic economic objective is to maximize benefits minus costs for society as a whole. In many cases benefits (and costs) will not equal the cash flows which result from the project. For example, public transport could be heavily subsidised (or even offered at no charge to some members of the community). Few would argue that the price paid for a public transport trip is a true measure of the benefit of that trip.

The theoretically correct measure of public sector benefits is the *willingness-to-pay* by consumers for the outputs of the project. The amount that a consumer is willing to pay for a commodity is a measure of the value of the commodity to that consumer. This concept is illustrated in Figure 8.13 for the provision of a new good (e.g. a new type of smart phone) to a community. It is assumed that there is a maximum price, P_1, that any person is willing to pay for the new good. The market price, as determined by the intersection of the demand and supply curves for the good, is P_0. However, as indicated by the demand curve, one consumer is willing to pay $\$P_1$ for the good. Thus $\$P_1$ is a measure of the value of the good to that consumer. Someone is willing to pay $\$P_2$ ($< \$P_1$) for the second unit and so on. Finally, someone is only willing to pay $\$P_0$ for the good. Anyone not willing to pay

$\$P_0$ will, of course, not purchase the good at the current market price. Therefore, the total willingness-to-pay for the good is represented by the total shaded areas ($A + C$) in Figure 8.13. The amount actually paid by the consumers will equal the revenue P_0Q_0 (area C). The difference between the willingness-to-pay and the revenue is called the *consumers' surplus* (i.e., the area A). It represents the *net* benefits to the consumers of the good. The revenue is a benefit to the producers.

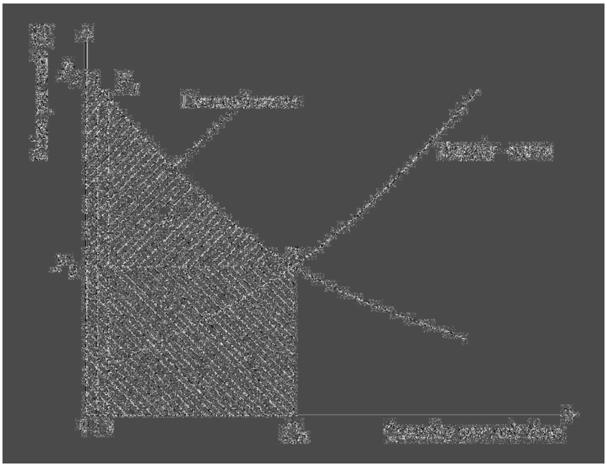


Figure 8.13 The willingness-to-pay concept.

In many cases the benefits of a public project can be assessed by the increase in consumers' surplus resulting from the project. For example, a state highway authority proposes to duplicate an existing two-lane highway which is currently heavily congested. The benefits of this project can be assessed using the increase in consumers' surplus. Figure 8.14 shows the demand-supply situation for the highway. The price per trip is represented by the sum of the operating cost plus the value of travel time per trip. The supply curve for the existing highway represents the price per trip as a function of the two-way volume.

The equilibrium volume on the existing highway is Q_1 vehicles per hour at a cost per trip of P_1. The new equilibrium volume on the duplicated highway will be Q_2 at a cost per trip of P_2. Note that an increase in volume is expected because an additional number of motorists now find it attractive to travel on the highway at the reduced price of P_2 per trip.

The benefit to all motorists is equal to the increase in consumers' surplus which is the shaded area in Figure 8.14. This consists of two parts:

- a reduction in cost to the existing users of the road (the rectangular area A)
- the consumers' surplus to new users (the approximately triangular area B)

If the demand curve is linear, the benefits are given by:

$$\text{Increase in consumers' surplus} = (P_1 - P_2) \times (Q_1 + Q_2)/2 \qquad (8.23)$$

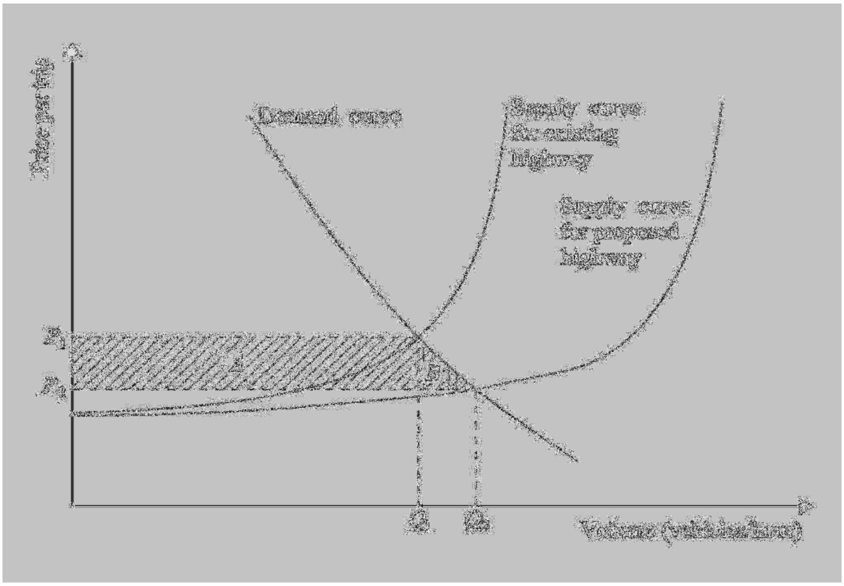


Figure 8.14 Increase in consumers' surplus of a duplicated highway.

Note that the concept of revenue is not relevant in this problem, as the price is not paid to a producer but represents a resource cost to society.

As a second example, the demand-supply situation for the provision of water to a city by a water authority is shown in Figure 8.15. The authority is considering constructing a new dam which would cause the supply curve to move to the right as shown, as the dam will increase the supply of water at any particular price. Assume that the authority sets the price of water equal to that which would apply in a competitive market situation.

Under the existing situation, Q_1 units of water are supplied per year at a price of P_1. With the new dam in place, the price would fall to P_2 and the volume supplied would increase to Q_2.

The benefits of the new dam will consist of the following:

- the benefits to the consumers as represented by the increase in consumers' surplus. This is area (A + B) in Figure 8.15
- the increase in revenue to the water authority. This is given by (P_2 Q_2 – P_1 Q_1) in Figure 8.15 or area (C + D) minus area (A + C)

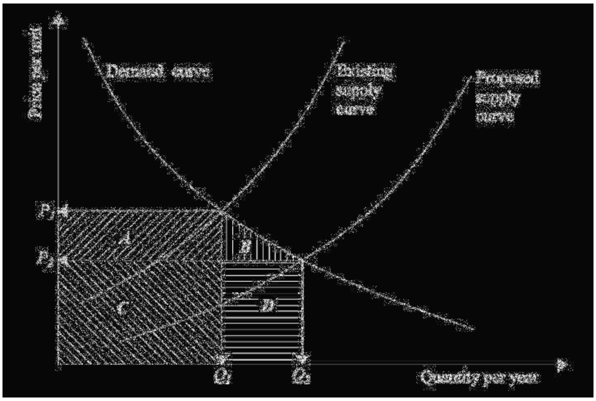


Figure 8.15 Benefits of an increase in urban water supply under market pricing.

The total benefits will be the sum of these, i.e., total benefits = increase in consumers' surplus plus increase in revenue

= $(A+B) + [(C+D) - (A+C)]$

= $(B+D)$

As shown in Figure 8.15 this is simply the increase in willingness-to-pay of the consumers. Assuming a linear demand curve it is given by

$$\text{total benefits} = (Q_2 - Q_1) \times (P_1 + P_2)/2 \qquad (8.24)$$

If the increase in Q is small, $P_1 = P_2$ and Equation (8.24) becomes

$$\text{total benefits} = (Q_2 - Q_1) \times P_1 \qquad (8.25)$$

i.e., total benefits = increase in output x market price

Note that in the case of a large public authority (or a private monopoly) an increase in supply could be large enough to cause a price change.

Now suppose the water authority does not charge the market price for water, but instead charges a price of P_3 as set by the government (see Figure 8.16). If the initial quantity supplied is Q_1 the authority would have to ration the water between consumers as the total demand Q_3, at a price P_3, exceeds the available supply Q_1. This is common practice in the supply of water for irrigation purposes in which each irrigator has an annual allocation that cannot be exceeded under normal circumstances.

With the existing supply curve and a rationed quantity Q_1 the total willingness-to-pay is given by areas $(A' + C')$. This is composed of the authority's revenue (area C') and the consumers' surplus (area A'). The quantity Q_1 $(P_1 - P_3)$ is the difference between the revenue for the water under a competitive market system and the actual revenue. It represents a transfer payment from the authority to the consumers as a result of the government pricing policy.

Desalination benefits

In 2006 the Israeli government was planning the construction of desalination plants to supplement the water supply for its people. While the costs were relatively easy to quantify, based on the cost of new plant, running costs due to electricity and the replacement of reverse-osmosis membranes, the benefits were more nebulous. However, by considering effects such as reduced scaling in pipes, extended lifetimes of electric and solar heaters, savings in soap in washing clothes and dishes, it was possible to determine a benefit of approximately US\$0.11 /m3. When savings in pumping costs were taken into account, the total benefit rose to approximately US\$0.15 /m^3.

Source: Dreizin, 2006

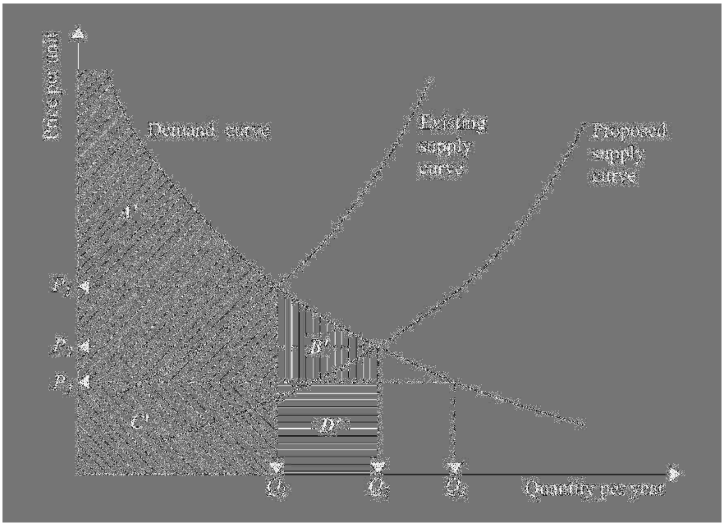


Figure 8.16 Benefits of an increase in urban water supply under non-market pricing.

If the supply curve shifts as a result of building a new dam and quantity Q_2 is supplied at price P_3, the increase in benefits will be the increase in consumers' surplus plus the increase in revenue, i.e. area (B' + D') in Figure 8.16. This is identical to the increase in willingness-to-pay given by (B + D) in Figure 8.15. Therefore, the government pricing policy does not affect the total benefits of the scheme, only the way in which these benefits are distributed between consumers and the water authority.

Humans to Mars?

Ehlmann et al. (2005), in a paper promoting a mission to land humans on Mars, estimate the costs of such a venture at between US$20 billion and US$450 billion where the latter included using the moon as an intermediate staging post. With the costs firmly established, the challenge was to detail the benefits for such a mission. These, it was suggested, were to be found in the promotion of industry, engineering, technology and science. The authors put forward the US$85 billion satellite industry as a quantifiable benefit of previous space exploration and suggested that future work would contribute in similar ways. They also argued that the work would create new markets, allow for more efficient use of resources and create high-wage jobs.

8.8 ECONOMIC COSTS

The concept of cost is a subtle one which may cause confusion in some engineering studies. Essentially, the same concept of cost is applicable to public sector and private sector projects, although the question of cost to whom must be addressed. For the private sector, only costs incurred by the firm are of relevance, whereas for public sector projects the total cost to all members of society must be evaluated. The correct concept of cost is the incremental opportunity cost of resources used plus any *externalities* (for example pollution, social disruption) as a direct result of the project.

The opportunity cost of a resource is its value when used in the best available alternative use. This is a measure of the value which must be forgone by using the resource in the project under consideration. The opportunity cost may be quite different from what was actually paid for the resource.

For example, a company has a vacant block of land which is valued at $500,000 on the open market. The company paid only $100,000 for the land some 10 years ago. If the company is considering building an office block on the land, the opportunity cost of the land is at least $500,000 because the land could be sold at this price. This is the cost which should be used in an economic evaluation of the proposed office block. The opportunity cost may be higher if a better alternative use is available. The actual price paid for the land is not relevant to this decision.

As a further example of this concept, consider a contractor who has a crane which costs $500 per day to operate and which can be hired out for $1,000 per day. The opportunity cost to the contractor of using the crane on a particular construction project is $1,000 per day, as this is the benefit which he must forgo in order to use the crane.

The opportunity cost concept has particular significance in public sector evaluation. If, for example, a project in the public sector employs labour which was previously unemployed, the opportunity cost of doing so is considerably less than the actual wage rate. As the alternative use of the labour is, in fact, unemployment, the opportunity cost is the value which the workers place on their additional recreation time when unemployed. The question of unemployment benefits paid to the individual by the government does not enter the evaluation, as these are a transfer payment between one section of the community and another, i.e. the benefit of this payment to the unemployed equals the cost to the taxpayers, so its net effect on the total community is zero.

The incremental or marginal costs of a project must be carefully assessed. In project evaluation, the marginal costs are the additional costs of undertaking the project compared with not undertaking it. There is sometimes a tendency to confuse marginal cost with average cost. The distinction between the two is illustrated by the following example.

Worked example

An engineer purchases a new car for $20,000. She uses the vehicle to commute to work and for recreational travel on weekends. The estimated annual costs for the vehicle are divided into standing costs (loss of interest on capital, depreciation, registration and insurance) which are estimated to be $8,200 p.a. and running costs (maintenance and repairs, petrol and oil, tyres) which are estimated to be $1,800 p.a. If she travels 20,000 km per year, the average cost of travel is $0.50/km ($10,000/20,000 km). Now if her employer asks her to use her own vehicle to make a single trip of 50 km to a construction site, what is the cost of this trip? One possible answer would be to use the average cost of $0.50/km and thereby estimate a trip cost of $25. However, the extra trip is incremental to her private travel and therefore the incremental or marginal cost should be used. The registration and insurance costs and loss of interest on capital are unaffected by the extra trip and hence are irrelevant in this situation. Likewise depreciation is more strongly related to the ageing of the vehicle rather than distance travelled so it does not enter the calculation. The only incremental costs are petrol and oil, tyres, and maintenance and repairs (assuming these depend on distance travelled). Therefore, the marginal cost for a short trip is $0.09/km ($1,800/20,000 km). This gives a cost of $4.50 for the 50 km trip. If the engineer is reimbursed at the average cost of $0.50/km, she makes a nice profit on the trip.

Now consider a different situation in which the engineer uses her car for 15,000 km of business travel and 5,000 km of personal travel per year. It could be argued in this case that the vehicle is primarily for business purposes and the private travel is the marginal component. It would then be reasonable for the company to reimburse all standing costs of the vehicle as well as the running costs associated with the business travel. Alternatively, the company may provide a vehicle and ask the engineer to pay the running costs associated with private travel.

Sunk and recoverable costs

A further distinction needs to be made between sunk and recoverable costs. Sunk costs are costs which have been incurred in the past and are no longer recoverable. As such they are irrelevant to decisions about future actions.

Worked example

A firm has a printer which originally cost $12,000 to purchase and now costs $2,500 per year to operate and maintain. It can be sold for $5,000. An equivalent new printer costs $15,000, but it is estimated that it costs only $1,500 per year to run because of an attractive maintenance contract being offered by the manufacturer. What are the benefits and costs which should be considered in deciding whether or not to purchase the new printer? A time-frame of 5 years should be used in analysing the decision. At the end of this time the new printer is expected to be worth $4,000 and the old printer nothing.

Solution We need to consider the incremental benefits and costs of the new printer relative to the old. These are shown in Figure 8.17. The incremental costs of purchasing the new printer is $10,000 ($15,000 minus the $5,000 sale value of the old printer). The incremental benefit is the $1,000 per year savings in operating cost plus the $4,000 value of the new printer at the end of 5 years. Note that the original purchase price of the old printer (i.e. the $12,000) is a sunk cost and does not enter the decision. Clearly the decision would be no different if the old printer had cost $100,000 or had been obtained for nothing.

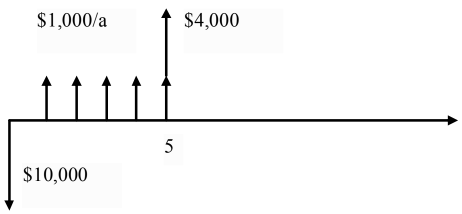


Figure 8.17 Benefits and costs of purchasing a new printer.

8.9 COST ESTIMATION

In the preparation of tenders, where engineering firms bid to undertake jobs, it is important to be able to estimate costs accurately. For example, the cost of a complete building, or the cost of installation of air conditioning, or in fact any of the many possible components that ultimately make up the total job. To assist in this process there are publications available that contain this type of information, taking into account the current year, the location of the proposed work and many other variables. These publications include:

- *The Building Construction Cost Data Book* 74th Edition by R.S. Means (2016) that provides building cost data for 930 locations in the US and Canada

- European construction costs available at www.constructioncosts.eu that has an online cost calculator as well as construction cost handbooks for all countries in Europe
- Rawlinsons (2016) *Australian Construction Handbook* that provides construction cost data for all capital cities in Australia

As an example of the types of information offered by these references, Rawlinsons (2016) contains estimates and detailed prices on the following:

- whole buildings (based on area or usage);
- electrical services (lighting, power, heating, closed circuit television installations, uninterruptable power supply);
- mechanical services (air-conditioners, hot water convectors, water boilers, heat exchangers, pipe work, duct work, natural ventilation); and
- civil services (excavation, laying concrete, installing structural steel).

The 2016 edition of Rawlinsons, book also contains useful environmental information on topics such as green design, environmental management and the embodied carbon of various materials.

The costs given are based on average values of actual contracts let in the preceding 12 months in the respective cities. It is difficult to summarise the full scope of the publication but some examples (current for 2016 prices) may assist in the process.

- The price of a 300 to 500 seat cinema in suburban Adelaide that includes air conditioning, ancillary facilities, seats and projectors, etc. is between $7290 and $7860 per seat. Hence a 300-seat theatre would cost around $2.2 m.
- To supply and install an evaporative air conditioner that meets a recommended 30 air changes per hour in a suburban house with 2.6 metre (nominal) ceiling height in Perth would cost between $41 and $53 per square metre of floor area.
- The supply and placing of an unreinforced strip footing in 25 MPa concrete in Melbourne costs $264 per cubic metre.
- To supply and trench 300 mm diameter ductile iron cement lined (DICL) water supply pipe (PN 35) with rubber ring joints in Adelaide costs $325 per metre of length.
- The installation of a set of traffic lights, including fully activated detectors, electrical services between signals, control box, pedestrian push buttons and indication signs to the intersection of an access road with a dual carriageway costs between $125,000 and $155,000 in Sydney.

8.10 SELECTION OF DISCOUNT RATE AND PROJECT LIFE

The choice of the discount rate and project life may have an important influence on the economic viability of a particular project. Many engineering projects, for example roads, water supply, electricity supply and communications networks, involve high initial costs and benefits which grow steadily with time over many years. The computed NPV of such projects falls with increasing values of the discount rate, and may become negative for high values. For example, a proposed new highway will cost \$10 m to construct. The benefits due to savings in operating costs and travel time are estimated to be \$1.2 m in the first year of operation but will grow at 2% p.a. (compound) due to growth in the volume of traffic using the road. Maintenance costs are expected to be constant at \$200,000 per year of operation. The NPV (\$m) of the road for various values of life and discount rate can be calculated as follows:

$$NPV = \frac{1.2}{(1+i)}\left\{\frac{1-(1.02)^n/(1+i)^n}{1-(1.02)/(1+i)}\right\} - \frac{0.2\left\{1-(1+i)^{-n}\right\}}{i} - 10$$

The results are shown in Figure 8.18. Clearly, for any given value of the discount rate, the NPV increases with increasing values of project life. Also, for any given value of project life, the NPV decreases with increasing values of the discount rate. It should be noted that the economic desirability of the project depends strongly on the discount rate. For example, if i = 5%, NPV is positive provided the project life exceeds 12 years. On the other hand, for a value of discount rate of 15% the NPV is negative for all values of project life. For high values of discount rate, the NPV is relatively insensitive to changes in project life. For example, with a discount rate of 10% the NPV is \$2.0 m if the project life is 35 years and \$2.76m if the project life is 50 years.

There has been considerable controversy associated with the selection of discount rate (and project life), particularly for public sector projects. For example, consider the case of a public authority which is required to justify all projects on an economic basis and only undertake projects which have a positive NPV. Clearly the list of economically viable projects will be much greater if the authority uses a low value of discount rate than if a high value is used. This fact has led to those who support large public works expenditure to argue for a low value of discount rate, whereas those who wish to reduce public works expenditure often argue for a high value. Before discussing this matter further, we shall consider the choice of discount rate for evaluating private sector projects

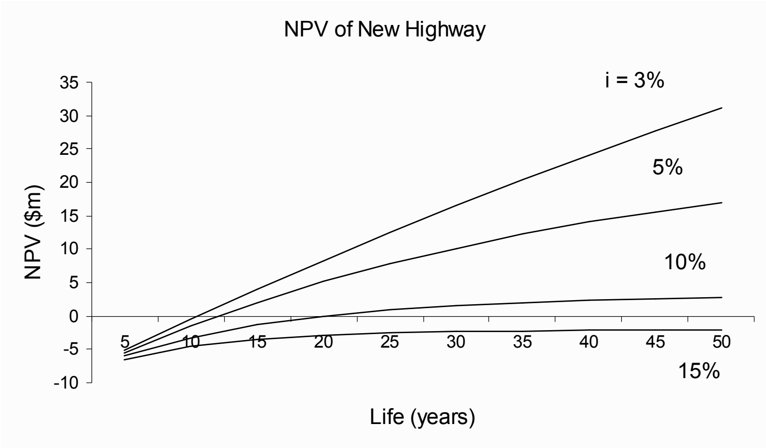


Figure 8.18 The effect of discount rate and life on the NPV of a road project.

The discount rate in the private sector

The correct concept for the discount rate is the ***opportunity cost of capital.*** The opportunity cost of capital for a particular company or individual at a particular point in time is the highest rate of return that it could obtain by having extra capital available. Obviously this depends on the investment opportunities at that time.

Most companies raise their capital by a combination of borrowing from financial institutions (called ***debt capital***) and reinvesting some of the company profits or retained earnings (called ***equity capital***). In this case the opportunity cost of capital is a weighted average of the interest rate paid on debt capital and the rate of return that the shareholders could obtain by investing their capital to the best advantage elsewhere. This usually results in a higher rate than the interest paid on debt capital.

The discount rate in the public sector

As noted above, the choice of a discount rate in the public sector is often controversial. The three most common theoretical bases for the choice of a discount rate are

- the average yield on long-term government bonds;
- the social opportunity cost; and
- the social time preference rate.

As there is no agreed method, we shall consider the arguments for and against each one in turn.

The average yield on long-term government bonds

It has been argued that the interest rate paid on long-term government bonds is the rate at which the government borrows money from the public. It is therefore appropriate that it should be used as the discount rate for evaluating public sector investments. In Australia, this rate is typically 1 to 2 % p.a. above the inflation rate. However, if we examine the federal budget as a whole, it is clear that most government revenue is raised by taxes, not by the sale of government bonds. In fact, bonds are used by the government to regulate the money supply rather than to raise revenue. The argument to link the discount rate to the government bond rate is therefore questionable.

Another argument which supports the use of the long-term government bond rate is that this represents the risk-free interest rate in the private sector. Most investments carry some element of risk, due to uncertainties in the future economic climate. Even banks and building societies have been known to default on interest payments during major depressions. The interest on government bonds is guaranteed by the government and so carries no risk. Any private investment should return at least this interest rate, otherwise investors would withdraw their money and put it into government bonds. It is argued that it is therefore appropriate to use it for public sector investments.

The social opportunity cost

It is argued that funds invested by the government in major capital works have basically been displaced from the private sector and should, therefore, earn a rate of return at least as high as that prevailing in the private sector (based on the opportunity cost of capital concept). However, not all revenue raised by government taxation would have been invested in the private sector. In fact a high percentage of government revenue is taken from expenditure that would have gone into consumption and would produce no investment return. This fact must be taken into account when assessing the social opportunity cost. A further difficulty is in determining the appropriate marginal rate of return in the private sector. A typical rate for the social opportunity cost is 10–15% above inflation.

The social time preference rate

It is argued that, by undertaking investments in the public sector, society is effectively forgoing consumption in the current time period in order to achieve increased consumption in the future. The choice of a discount rate is equivalent to placing a weight on benefits and costs incurred in future years when compared with the present. Some economists consider that society is free to choose whatever stream of future consumption that it likes, and accordingly it may choose any set of relative weights on benefits and costs in future years.

The discount rate corresponding to this chosen set of relative weights is the social time preference (STP) rate, which will in general differ from the interest rate prevailing in the private sector.

To illustrate the concept of a social time preference rate, suppose that society (or its elected representatives) decides that benefits and costs received in one year should have a weight W_1 relative to benefits received in the present. Similarly, benefits and costs received in two years have a weight W_2 relative to the present. In general, W_t is the weight placed on benefits and costs received at the end of year t relative to benefits and costs received in the present. (Clearly $W_0 = 1$).

Then the net community benefits (NCB) can be defined as:

$$NCB = \sum W_t \, (B_t - C_t) \qquad (8.26)$$

Based on the preference of most individuals to consume now rather than later, it is reasonable to assume that:

$W_0 > W_1 > W_2 > \ldots$, etc.

A somewhat stronger assumption is that the weight placed on the benefits and costs in any particular year is a constant fraction of the weight placed on the benefits and costs in the previous year, i.e.,

$$\begin{aligned} W_1 &= kW_0 = k \\ W_2 &= kW_1 = k^2 \\ &\ldots\ldots\ldots \\ W_t &= kW_{t-1} = k^t \end{aligned} \qquad (8.27)$$

where k is a chosen parameter between 0 and 1.

Suppose we define k as follows:

$$k = 1/(1 + r) \qquad (8.28)$$

where r is the **social time preference** (STP) rate.

Combining Equations 8.26, 8.27 and 8.28 we obtain:

$$NCB = \Sigma \, \{ B_t / (1 + r)^t \; - C_t / (1 + r)^t \} \qquad (8.29)$$

The similarity with Equation (8.17) should be noted. It should be clear that choosing the STP rate, r, is equivalent to choosing a set of weights on future benefits and costs for society as a whole.

The STP rate may be any rate chosen by society (even 0%). Proponents of the STP rate usually argue for a low rate (in the order of 2–5%) on the grounds that government investments provide basic community infrastructure and therefore benefits over a long time scale should be considered.

Summary: Discount rate for public sector projects

Ultimately, the choice of the discount rate for public sector projects tends to be a political decision. Often, sensitivity analyses are carried out in which the NPV for all projects is calculated using a low, medium, and high value of the discount rate. For example, the NPV of all projects can be determined using discount rates of 3, 6 and 10% p.a. and the results compared.

Project life

Many engineers consider the life of an engineering project to be equal to the physical life of the constructed facilities. The physical life of a system ends when it can no longer perform its intended function. For example, a building is at the end of its physical life when it is no longer fit for habitation. A road pavement is at the end of its physical life when it is suffering total break-up and can no longer be repaired. However, two other concepts of project life are relevant in economic evaluation. These are the economic life and the relevant period of analysis.

The economic life of an engineering system ends when the incremental benefits of keeping the system operating for one more year are less than the incremental costs of maintenance and repair for one more year. As most engineering systems can be kept operating almost indefinitely with suitable replacement parts, the economic life is usually less than the physical life.

The relevant period of analysis is the maximum period beyond which the system performance has virtually no effect of its present value of benefits and costs. For high values of discount rate, the NPV of a project does not significantly increase beyond a certain time. This time is the relevant period of analysis. For example, if a discount rate of 15% p.a. is being used, it is pointless arguing whether the life of a building will be 30 years or 100 years as it will make no difference whatsoever in its economic evaluation. For lower values of discount rate, more care must be taken in choosing the project life.

In general, the project life is the smallest of the physical life, the economic life and the relevant period of analysis. Some typical lives used for civil engineering projects are given in Table 8.9.

Table 8.9 Lives of some engineering systems.

Engineering system	Life
Vehicles	5–10 years
Roads	20–30 years
Bridges, buildings	30–50 years
Dams	50 years
Tunnels, cuttings, embankments	100 years

8.11 SUMMARY

The aim of economic evaluation is to assess if the benefits of a project exceed its costs and hence whether or not it is worth undertaking from an economic perspective. The benefits and costs of a project need to be carefully defined as they do not necessarily correspond to cash flows, particularly in the public sector where a number of goods and services are supplied to the community at a subsidised price.

The benefits and costs of engineering projects often occur over a long time period. Hence, the effects of the time value of money need to be taken into account in assessing the economic viability of a project. A discount rate is applied to future benefits and costs in recognition that they are less significant than benefits and costs incurred today.

There are a number of economic criteria for evaluating projects. These include the payback period, net present value, equivalent annual worth, the benefit-cost ratio and the internal rate of return. In general, economists favour net present value as it is not sensitive to the issue of whether savings in cost should be counted as positive benefits or negative costs and it produces an unambiguous ranking of projects.

The selection of the discount rate, particularly in the public sector, can be controversial and there are a number of bases for its selection. The social time preference rate is preferred as it is recognises that the choice involves social values that can be expressed through the political process.

PROBLEMS

8.1 \$2000 is invested at 5% p.a. for 20 years. What sum will be received at the end of the investment period?

8.2 What is the present value of \$2,000 that will be received in 5 years if the interest rate is 6% p.a.?

8.3 What is the present value of a series of annual payments of \$10,000 over 10 years if the interest rate is 8% p.a.?

8.4 What are the annual repayments for a loan of \$100,000 if the interest rate is 8% p.a. and the period of the loan is 45 years?

8.5 Following on from Question 8.4, how much is still owed after the first annual repayment on a \$100,000 loan taken out at an interest rate of 8% p.a. over a 45-year period?

8.6 What are the monthly repayments for a loan of \$100,000 if the annual interest rate is 6% p.a. (computed monthly) and the loan is over 50 years?

8.7 A project has initial costs of $100,000 and annual benefits of $10,000. What is the net present value of the project assuming an 8% p.a. discount rate over 25 years? Assume all costs and benefits occur at the end of the year.

8.8 If the benefit-cost ratio is defined as B/(O+K) where B are the benefits, K the initial costs and O the ongoing costs, what is the benefit-cost ratio in the case where initial costs are $20,000, ongoing costs $5,000 p.a. and benefits $10,000 p.a. for a discount rate of 6.8% p.a. and over a period of 20 years? Assume all costs and benefits occur at the end of the year.

8.9 If the benefit-cost ratio is defined as (B-O)/K where B are the benefits, K the initial costs and O the ongoing costs, what is the benefit-cost ratio in the case where initial costs are $20,000, ongoing costs $5,000 p.a. and benefits $10,000 p.a. for a discount rate of 6.8% p.a. and over a period of 20 years? Assume all costs and benefits occur at the end of the year.

8.10 As engineer for a company developing high efficiency solar panels, you have an annual research and development budget of $500,000. Each year you call for submissions from within the company for projects that could be done. As part of this, the proponents must estimate the cost of the proposal and what benefits would come from a successful outcome. This year you have eight proposals to consider, and one criterion for funding is economic benefits. The details of the proposals are given in Table 8.10. Any calculations that you do should assume a project life of 5 years and a discount rate of 20% p.a.

(a) Use the B/C ratio to determine which projects you would select. In this case the best proposals are chosen from a ranked list.

(b) Use NPV to determine which projects you would select. In this case the problem can be considered as one that could be solved using linear programming, but you may attempt to find a solution by inspection. As an additional exercise, set the problem up in EXCEL and use Solver to verify the solution obtained by hand.

Table 8.10 Summary of economic factors associated with the eight research proposals to be assessed.

Project	Initial Cost($)	Annual Benefit($)
A	100,000	55,000
B	200,000	100,000
C	50,000	24,000
D	10,000	8,000
E	20,000	20,000
F	345,000	120,000
G	100,000	34,000
H	250,000	160,000

8.11 A manufacturing company has space for one additional automated machine in its factory. It can purchase either Machine A which produces gadgets or Machine B which produces widgets. Economic data for the two machines are presented in Table 8.11.

Table 8.11 Costs and benefits for two machines.

	Machine A	Machine B
Initial Cost ($)	310,000	370,000
Revenue ($ p.a.)	70,000	120,000
Maintenance ($ p.a.)	15,000	20,000
Life (years)	10	5
Scrap value ($)	20,000	0

(a) Compute the net present value, equivalent annual worth, B/C ratio for each machine using a discount rate of 7% p.a. and a planning horizon of 10 years.

(b) Which machine should be chosen and why?

8.12 A car that was purchased 10 years ago for $20,000 is now thought to be worth $5,000 as a trade-in on a new model. Due to its age, the old car now costs $2,000 per year in repairs alone, which it is assumed a new car would not need. Based on economics, should a new car be purchased for $30,000 if it is intended to keep it and evaluate it over a 10-year planning horizon? Assume that at the end of the 10 years a car purchased now would be worth $5,000 and the existing car would be worthless.

REFERENCES

Adeoti, O., Ilori, M.O., Oyebisi, T.O. and Adekoya, L.O., 2000, Engineering Design and Economic Evaluation of a Family-Sized Biogas Project in Nigeria. *Technovation*, 20, 103–108.

Baumol, W. J. 1977. *Economic theory and operations analysis,* 4th edn. Englewood Cliffs, NJ: Prentice Hall.

Best, R., 2000, Low Energy Building Design: Good for the Environment and Good for Business. The Environmental Engineer. *Journal of the Environmental Engineering Society*, I.E.Aust., 1(1), 12–15.

Brookshire, D.S., Burness, H.S., Chermak, J.M. and Krause, K., 2002, Western Urban Water Demand, *Natural Resources Journal* 42, 873–898

Campus Review, 2002, https://www.campusreview.com.au

DeGarmo, E. P., Sullivan, W. G. and J. R. Canada 1984. *Engineering economy,* 7th edn. New York: Macmillan.

deNeufville, R. and Stafford, J. H. 1971. *Systems analysis for engineers and managers*. New York: McGraw-Hill.

Dreizin, Y., 2006, Ashkelon Seawater Desalination Project – Off-Taker's Self Costs, Supplied Water Costs, Total Costs and Benefits. *Desalination*, 190, 104–116.

Ehlmann, B.L., Chowhury, J., Marzullo, T.C., Collins, R.E., Litzenberger, J., Ibsen, S., Krauser, W.R., DeKock, B., Hannon, M., Kinnevan, J., Shepard, R. and Grant, F.D., 2005, Humans to Mars: a feasibility cost benefit analysis, *Acta Astronautica* 56 (9-12) 851-858

Engineering News Record, 2017, http://www.enr.com/economic

Forkenbrock, D.J. and Foster, N.S.J., 1990, Economic Benefits of a Corridor Highway Investment. *Transport Research Part A: General*, 24A(4), 303–312.

Hensher, D.A., 2008, Assessing systematic sources of variation in public transport elasticities: Some comparative warnings, *Transportation Research Part A* 42, 1031–1042

Jackson, J., Jorge, J.-D. and de Rus, G., 2004, Cost-Benefit Analysis of Investments in Airport Infrastructure: A Practical Approach. *Journal of Air Transport Management*, 10, 311–326.

McConnell, C.R., 1980, *Economics (Australian Edition)*. Sydney: McGraw-Hill.

Lipsey, R. G., Langley, P. C. and Mahoney, D. M. 1981. *Positive economics for Australian students*. London: Weidenfeld & Nicolson.

Loansense 2016, Information on the Reserve Bank of Australia – Australian Interest Rates for 1959 to 2016. http://www.loansense.com.au/historical -rates.html (accessed on March 28, 2017).

Means, R.S., 2016, *Building Construction Cost Data Book* 74th Edition (available at www.rsmeans.com/products/books/2016-cost-data-books/2016-building-construction-cost-data-book.aspx).

Meredith, D. D., Wong, K. W., Woodhead, R. W. and Wortman, R.H., 1985, *Design and planning of engineering systems,* 2nd edn. Englewood Cliffs, NJ: Prentice Hall.

Mishan, E. J., 1982, *Cost-benefit analysis,* 3rd edn. London: George Allen & Unwin.

Morgan, P. and Elliott, S.L., 2002, Mechanical Destratification for Reservoir Management: An Australian Innovation. Water, *Journal of the Australian Water Association*, 29(5), 30–35.

Muroi, K., 2017, The oldest example of compound interest in Sumer: seventh power of four-thirds, *Cornell University Library, Mathematics Archive*, https://arXiv.org/abs/1510.00330, (accessed on April 6, 2017).

Narayan, S., Beck, M.W., Reguero, B.G., Losada, I.J., van Wesenbeeck, B., Pontee N., Sanchirico, J.N., Carter Ingram, J., Lange, G. and Burks-Copes, K.A., 2016, The effectiveness, costs and coastal protection benefits of natural and nature-based defences, *PLoS ONE* 11(5): e0154735. doi:10.1371/journal.pone.0154735

Ossenbruggen, P. J., 1984, *Systems analysis for civil engineers*. New York: Wiley.

Riggs, J. L., 1977, *Engineering economics*. New York: McGraw-Hill.

Rawlinsons, 2016, *Australian Construction Handbook 34th Edition*. (Rawlhouse Publishing Pty. Ltd.), 1005pp.

Sage, A. P., 1982, *Economic systems analysis. Microeconomics for systems engineering, engineering management and project selection*. Amsterdam: Elsevier.

Scheierling, S. M., Loomis, J.B. and Young, R.A., 2006, Irrigation water demand: A meta-analysis of price elasticities, *Water Resources. Research,* 42, W01411, doi:10.1029/2005WR004009.

Smith, A. A., Hinton, E. and Lewis, R. W., 1983, *Civil engineering systems analysis and design*. London: Wiley.
Steiner, H. M., 1980, *Public and private investments.* New York: Wiley.
Talvitie, A., 2000 Evaluation of Road Projects and Programs in Developing Countries. *Transport Policy*, 7, 61–72.
The Australian, 2001, http://www.theaustralian.com.au
World Bank, 2016, www.data.worldbank.org/indicator/FP.CPI.TOTL (accessed on Nov 22, 2016).

APPENDIX 8A: PRICE DETERMINATION IN PERFECTLY COMPETITIVE MARKETS

The perfectly competitive market is a theoretical model used by economists to analyse market behaviour in much the same way that engineers use free-body diagrams to analyse the forces on a group of rigid bodies.

It is hypothesized that there is a market of buyers and sellers which determines the market price for any particular good or service (called a "commodity").

The assumptions inherent in the perfectly competitive market are as follows

- there are a large number of buyers and sellers of the commodity
- any particular commodity is identical for all sellers
- all buyers and sellers have perfect information about prices and production costs
- there is free entry and exit of firms into the market

The market supply curve

For any particular commodity there is a market supply curve which shows the total amount of the commodity which will be produced per unit time at any particular price. For example, Figure 8.19 shows the market supply curve for timber in a particular economy. (Note that it is traditional to plot price on the vertical axis.) As price increases, we would expect a greater production of timber as more producers come into the market and existing producers increase their output by working overtime, investing in new machinery, and so on.

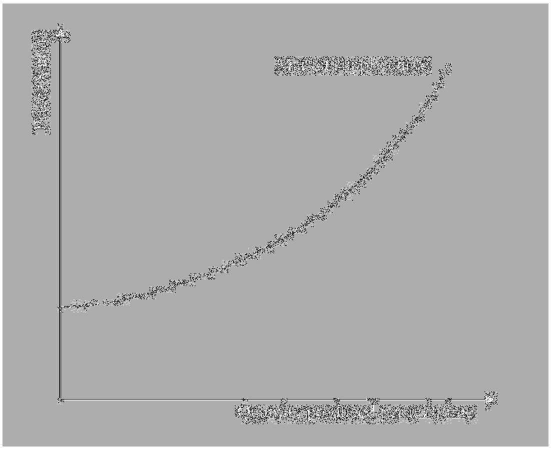


Figure 8.19 Market supply curve for timber.

Other factors apart from the price of the commodity can affect the quantity supplied. If these change, they will cause a shift in the position of the market supply curve in Figure 8.19. These factors are as follows:

- The price of inputs (for example, labour or materials) to the production process
- The price of competing outputs. For example, an oil refinery can be used to produce many products including petrol and fuel oil. If the price of petrol increases relative to the price of kerosene, the market supply curve for fuel oil would tend to shift to the left
- The state of technology. Technological advances can reduce production costs and cause the market supply curve to shift to the right
- The time period. The market supply curve for a short time period tends to be near vertical, as there is little time to respond to changes in price. In the longer term, the curve tends to be flatter, as there is more time for producers to respond to changes in price
- Expectations. The quantity supplied at any particular price will depend on whether producers expect prices to increase, decrease or remain the same in the next time period

The market demand curve

For any particular commodity there is also a market demand curve. This shows the quantity per unit time of the commodity which would be purchased at any particular price. Figure 8.20 shows the market demand curve for timber in a particular economy. As the price of timber increases, we would expect the demand to fall as consumers use substitute materials such as plastic, steel, and aluminium.

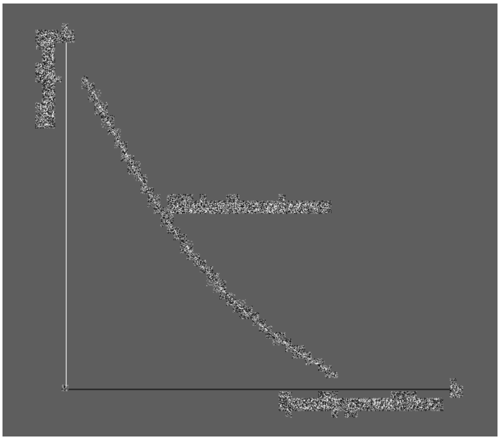


Figure 8.20 Market demand curve for timber.

Apart from price, a number of other factors affect the position of the market demand curve. These include the following:

- The prices of other commodities. Certain commodities are substitutes for one another. Consider, for example, gas and electricity. If the price of gas increases, the market demand curve for electricity tends to shift to the right. Other commodities such as concrete and steel reinforcing rods are complementary goods. If the price of concrete increases, there is a shift to the left of the demand curve for steel reinforcing rods.
- Income and wealth of consumers. If there is a general rise in income levels There is an increase in demand for all normal goods. This is reflected by a shift to the right of the demand curve.
- Number of consumers. An increase in the number of consumers causes the market demand curve to shift to the right.
- Tastes of consumers. Changes in taste and fashion may cause a shift in the demand curve for some commodities, e.g., clothing.
- The time period and expectations.

Market equilibrium

The market price for any particular commodity in a perfectly competitive market is determined by the interaction of demand and supply. At equilibrium the market price corresponds to the point of intersection of the market supply and market demand curves. This point is illustrated in Figure 8.21, where the equilibrium price is given by P^* and the equilibrium quantity by Q^*.

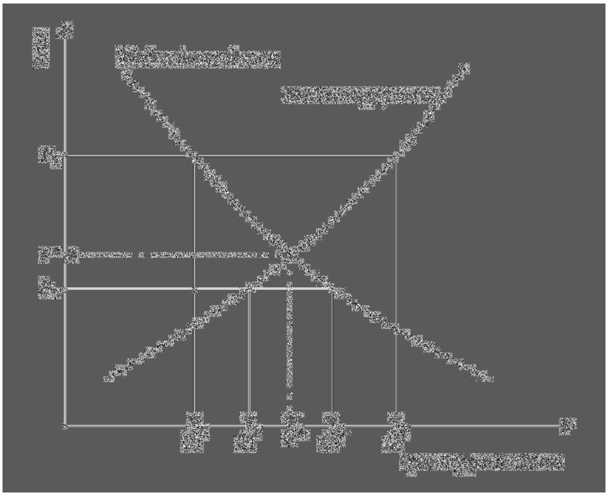


Figure 8.21 Demand-supply equilibrium.

Suppose the price is set at a value greater than P^*, for example P_1. Then the quantity supplied, Q^S_1 exceeds the quantity demanded, Q^S_1. The product therefore stockpiles and there is pressure on producers to reduce the price in order to sell their product. If the price is below P^* (e.g., P_2), the quantity demanded, Q^D_1 exceeds the quantity supplied, Q^S_1, and there are shortages of the commodity. In response to these shortages, consumers will tend to bid up the price of this scarce commodity. As the price increases, producers will tend to produce more and new producers will enter the market.

At a price of P^* the quantity demanded just equals the quantity supplied and an equilibrium state is established.

For example, consider the demand/supply equilibrium of traffic on a section of urban arterial road. In this case, each consumer may decide either to make one or more trips on the road or not to use it at all. The *price* of a trip is the total cost of petrol, oil, and tyre wear of making one traverse of the section of road. To this should be added the value of travel time of the driver and passengers. The *quantity* is the number of vehicles per hour travelling on the road.

Figure 8.2.2 shows supply and demand curves for the road. The supply curve indicates the total cost per trip as a function of the volume of traffic on the road in vehicles per hour. Typically, the cost per trip increases rapidly as the volume of traffic approaches the capacity of the road, *C*. This is due to the effects of traffic congestion and delays.

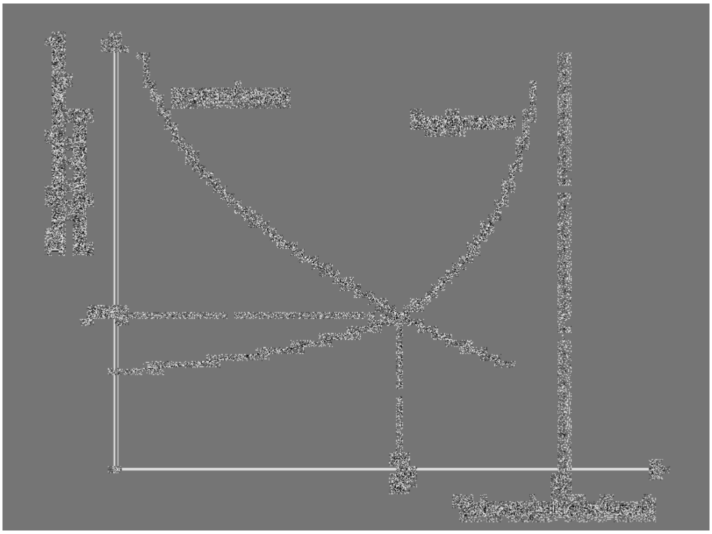


Figure 8.22 Demand-supply equilibrium for an urban arterial road.

The demand curve in Figure 8.22 indicates the number of drivers per hour wanting to use the road as a function of the total cost per trip. As the cost per trip increases, fewer drivers wish to use the road because of available substitutes such as other routes of travel or other times of the day. If the traffic conditions on the road remain steady for a sufficient period of time, equilibrium is reached with a volume of Q^* vehicles per hour on the road. The corresponding cost per trip is P^* dollars.

Shifts in demand and supply curves

As mentioned earlier, the demand curve for a particular commodity may shift with time due to changes in consumer preferences, the number of consumers, average income levels or the prices of related goods. A shift in demand leads to a new equilibrium quantity and price for the commodity. For example, suppose the demand for water in a city increases due to an increase in population. This causes the demand curve for water to shift to the right as shown in Figure 8.23. If the water supply authority charges the market price for water, one would expect the volume supplied to increase from Q_1 to Q_2 and the price of water to increase from P_1 to P_2 as less economical sources of water are used.

In a similar fashion, the construction of a new engineering system causes the supply curve to shift to the right and a new equilibrium to be established. For example, in response to increasing demand, a water supply authority builds a new dam to supply water to a city. This causes the supply curve to shift to the right as more water can now be supplied at any particular price. The situation is shown in Figure 8.24. In equilibrium, the volume of water supplied increases from Q_1 to Q_2 and the price falls from P_1 to P_2. In practice, this fall in price may take place over several years by allowing the rate of increase of the nominal price of water to fall behind the rate of inflation.

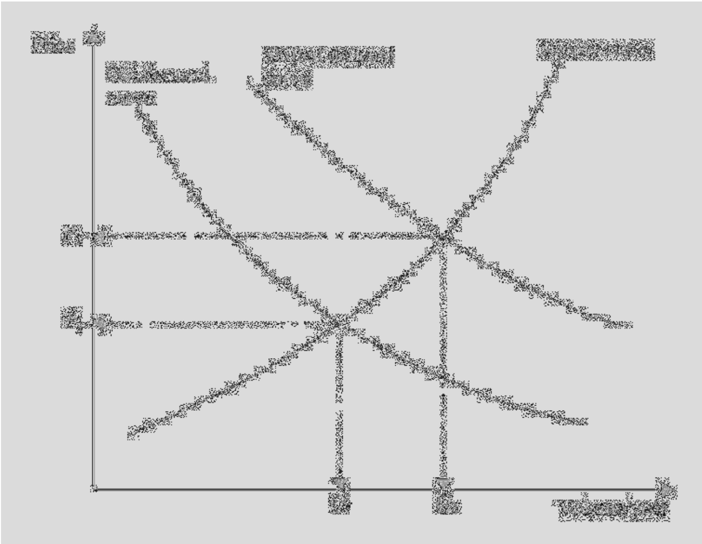


Figure 8.23 A shift in the demand curve for water.

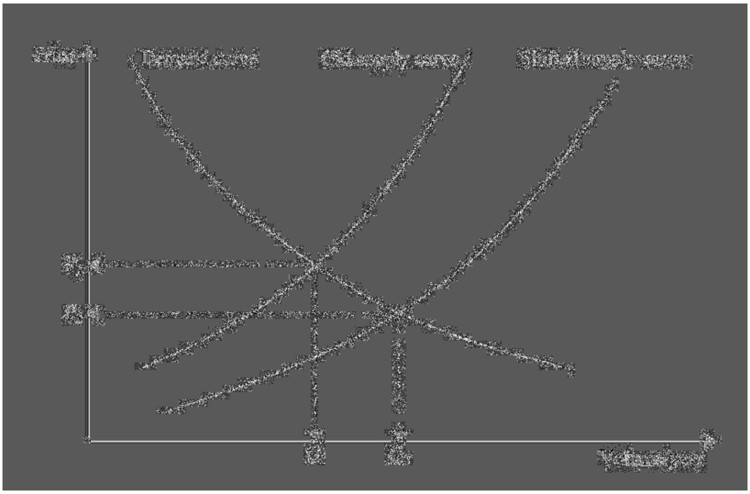


Figure 8.24 A shift in the supply curve for water.

Price elasticity of demand

In planning future engineering facilities such as transportation and water resource systems, the analyst is often concerned with the responsiveness of consumers to changes in price.

It is possible to characterize the price responsiveness of the market demand for a product at any point in time by its price elasticity of demand. This is defined as the percentage change in quantity demanded, divided by the percentage change in price, i.e. elasticity,

$$\varepsilon = \frac{\Delta Q / Q}{\Delta P / P} = \frac{\Delta Q}{\Delta P}\left(\frac{P}{Q}\right) \tag{8.30}$$

For example, the demand for water for irrigation from a particular river is 1.2 x 10^9 m^3 per year when the price of water is \$0.10/m^3. If the price increases to \$0.15/m^3 and the price elasticity of demand is -0.3, what will be the new demand? Using Equation (8.30),

$$\varepsilon = -0.3 = \frac{\Delta Q}{0.05} \times \frac{0.01}{1.2 \times 10^9}$$

Therefore

$$\Delta Q = \frac{-0.3 \times 0.05 \times 1.2 \times 10^9}{0.01} = -0.18 \times 10^9 \text{ m}^3\text{/year}$$

New demand = 1.2 x 10^9 – 0.18 x 10^9
= 1.02 x 10^9 m^3/year

Strictly speaking, the price elasticity defined by Equation (8.30) is the arc elasticity. In the limit as ΔP approaches zero, it becomes the point elasticity. This is defined as

$$\varepsilon_p = \frac{dQ}{dP}\left(\frac{P}{Q}\right) \tag{8.31}$$

For a downward-sloping demand curve the elasticity is negative. It is common to ignore the minus sign when quoting values of the price elasticity of demand. If the price elasticity of demand is less than minus one, the demand is said to be elastic, otherwise the demand is called inelastic.

Some typical estimated values of price elasticities of demand of interest to engineers are given in Table 8.12. Note that the reported values cover a wide range which would depend on a number of factors including the location and the method of estimation.

Table 8.12 Some typical values of price elasticity of demand.

Commodity	Price Elasticity of Demand		Number of Studies Considered	Reference
	Range	Average		
Urban Water	-0.11 to -1.588	-0.49	15	Brookshire et al. (2002)
Irrigation Water	-0.001 to -1.97	-0.48	24	Scheierling et al. (2006)
Public Transport	-0.002 to -1.121	-0.395	39	Hensher (2008)

CHAPTER NINE

Sustainability, Environmental and Social Considerations

This chapter explains how sustainability, environmental and social considerations inform the planning and design of engineering projects. The concept of sustainability and its assessment can be challenging. However, it can also be effectively used as a basis for the assessment of development and community projects. The concepts presented in this chapter allow for community values to be incorporated into decision making processes by placing an emphasis on the environmental and social implications of engineering projects. Methods for taking account of sustainability, social and environmental issues are briefly presented.

9.1 INTRODUCTION

Prior to the 1970s, engineering projects were often human-centric. Engineers focussed primarily on the technical and economic aspects of projects, without giving much concern to the social and environmental impacts of their work. To a large extent this was a reflection of the priorities of society at the time, and to a lesser extent, due to an insufficient understanding of the interconnectedness between engineering works and natural and social systems. Engineers now recognise that they play a critical role in the wise use, conservation and management of resources. They also recognise that they have an obligation to ensure that the needs of future generations are considered. This is reflected by professional bodies through accreditation requirements for engineering programs, and through codes of ethics for practising engineers.

In 1828 Thomas Tredgold of the Institute of Civil Engineers described engineering as the art of directing the great forces of nature for the use and convenience of man. As we have seen already in Chapter 1, nature tended to be regarded as a powerful adversary to be tamed. Today such views are tempered by the realisation that the world has finite resources and the emphasis now is on the wise use of resources, not only to protect and conserve the environment, but also to achieve a sustainable way of life into the foreseeable future.

The concept of sustainable development has been discussed extensively in past decades, with the mainstream definition being that of the Brundtland Commission of 1987 (WCED, 1987):

> *Development that meets the needs of the present without compromising the ability of future generations to meet their own needs.*

This concept has developed further, and engineering systems, policies, designs and plans can be considered and evaluated today within sustainability frameworks.

In Section 9.2 of this chapter, we look at the concept of sustainability and in Section 9.3 at the techniques which are used by engineers in planning and design to deal effectively with sustainability as an over-riding goal of engineering work. While environmental and social issues are clearly inter-related with considerations of sustainability, they are dealt with separately in this chapter. Environmental considerations and assessment programs are introduced in Sections 9.4 and 9.5, while social issues and assessment tools are dealt with in Sections 9.6 and 9.7.

9.2 SUSTAINABILITY

The systems approach provides a simple and consistent basis for investigating sustainability at all levels of society, from the global scale down to the individual. Gilman (1992) used systems concepts to provide the following definition of sustainability:

> *The ability of a society, ecosystem, or any such ongoing system to continue functioning into the indefinite future without being forced into decline through exhaustion or overloading of key resources on which the system depends.*

Based on this definition, Foley et al. (2003) have shown that to achieve sustainability it is necessary to manage appropriately all the resources that a system relies on. These include the natural, the financial, the social and the man-made infrastructure resources that are important to the functioning of the system.

Roberts (1990) has shown that self-sustaining systems in nature are generally closed-loop systems that evolved gradually over time. In the past when humans developed production systems, they have relied on an open loop, once-through use of resources, which results in much waste. To be sustainable, it is necessary to use closed-loop systems. This emphasis on a closed loop for sustainable outcomes can also be seen in discussions related to a *circular economy*. The objective of a circular economy is to maximise value at each point of a product's life (Stahel, 2016). Another closed-loop system concept is the *cradle to cradle* approach. This approach is in contrast to the linear *cradle to grave* approach. The cradle to cradle approach recognises that the perceived end of life for a given product can, and should, become the start of another product life and therefore an interconnection is introduced between product life cycles. A cradle to grave approach also recognises the product life cycles, but often includes waste at the end of the product life. Figure 9.1 shows a conceptual model of the initial extraction or use of resources, followed by the processing, transportation and consumption of the modified resources as a closed-loop system that can evolve over time. More complex models include interconnected systems and cradle to cradle connections between product life cycles.

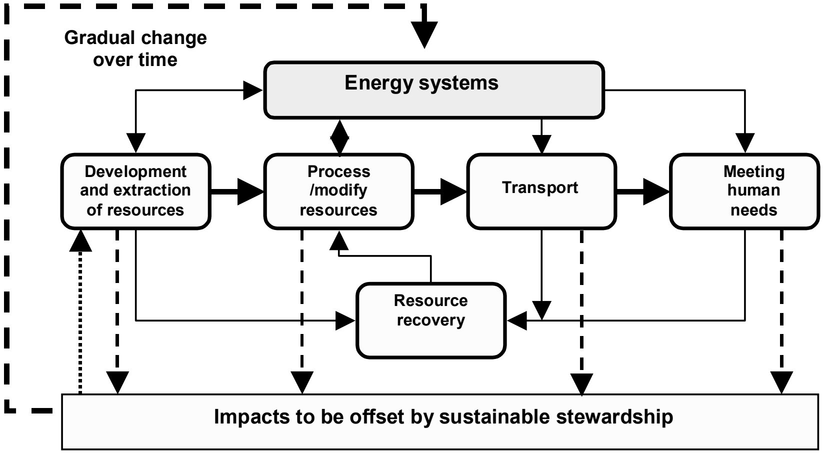


Figure 9.1 Closed-loop system representation (adapted from Roberts, 1990).

A key element of the model is the stewardship of the natural environment that is needed in order to provide a balance to the processes of extraction, modification and transportation. The main element of this closed-loop system is the recycling of waste from manufacture and consumption within a continuous loop. Energy inputs are a crucial element in the closed-loop system. For many natural systems the sun is the natural source of energy, but this is not so for man-made systems.

A conceptual model that can be used to assess the sustainability of a system or of sustainable system development (Foley et al., 2003) is shown in Figure 9.2. It outlines the flow of resources within the system. The model identifies infrastructure and other human-made resources (I) as a key element of sustainability. Infrastructure for urban development includes buildings, and the water supply system, as well as systems for waste, transport and energy. Such a systems approach provides a good platform for assessing development and sustainability, where infrastructure and resource flow are principal considerations.

Each subsystem within the larger development system can be modelled (e.g. a single house within a city development). The flow of resources such as water, energy and finance to and from the system can also be included in a more holistic way than in many other currently available tools.

A systems approach to the assessment of sustainability is not new. It was applied in early computer modelling studies in the 1970s (e.g. Meadows, et al., 1972) and in more quantitative studies which have continued up to the present. A major conclusion from such systems modelling, suggested by Suter (1999), is “that humankind needs to re-evaluate its exploitative attitude towards humans and the earth itself”.

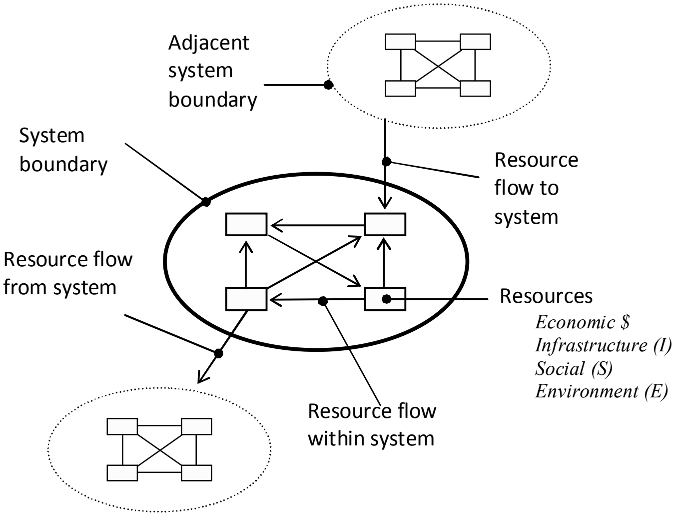


Figure 9.2 System representation (adapted from Foley et al., 2003).

Another application of the term sustainability relates to corporate social responsibility. This can impact engineers and their work through both the strategic direction and the investment decisions of the organisations within which they are employed or interact. Sustainability has been adopted in the business world to connote the principles of social and environmental responsibility. Increasingly business organisations have recognized that profits alone do not guarantee continuity of existence for their companies and that sustainability adds to their long-term business viability. Investors also see sustainability issues, such as controlling greenhouse gases, as priority concerns to be addressed by management. Since the late 1990s, the Global Reporting Initiative (www.globalreporting.org) has set the benchmark for sustainability reporting at the corporate level and the Dow Jones Sustainability Index (www.sustainability-indices.com) has become a mainstream investment consideration.

The final application of the term sustainability, as discussed here, relates to considerations of natural capital. Schmidheiny (1992) of the World Business Council on Sustainable Development suggested that population growth and economic development will eventually be constrained by environmental and social pressures. The question of what level of population is sustainable for the world is the focus of much debate amongst many scientific researchers and has been since the time of Thomas Malthus in the late 1700s. The key to the answer is in maintenance of natural capital and the level of resource use which is required for a standard of living that can be perpetuated for future generations. Sustainability has thus been explained as being the state for which all social well-being and development are within the Earth's biological capacity (Wackernagel and Rees, 1997).

Club of Rome

The Club of Rome was founded in 1968 and has been described as "a think tank and a centre of research and action, of innovation and initiative". It is composed of a group of scientists, economists, business people, senior public servants, heads of state and former heads of state from around the globe and addresses questions confronting the global society. The term "problematique" was developed by club members for the inter-related issues that were to be examined. The Club of Rome commissioned the book "Limits to Growth" which was published in 1972. Twelve million copies of the book were sold worldwide, in 37 languages. The book challenged a basic assumption of economic theory that the resources of planet Earth could be infinite. The book was based on computer modelling of the consequences of the growth of global population. Various scenarios researched pointed to a major economic crisis occurring in the early 1990s. This did not happen. Subsequent updates on the book have been undertaken and in retrospect, one needs to ask, was "Limits to Growth" a fair or a false warning?

Adapted from Suter (1999)

9.3 PLANNING AND DESIGN FOR SUSTAINABILITY

Time is a key consideration for sustainability as systems evolve and communities develop. This was highlighted by Fleming (1999) who has suggested that sustainability needs to be considered within cycles of continuous improvement. In evaluating the sustainability of engineering systems, projects, processes and operations, a key aspect to consider is the stream of resource usage. Various tools can assist in such evaluations, and are dependent on the level of society to which they are being applied. Examples of readily available tools available on the Internet are listed in Table 9.1. Some are produced by non-government organisations, some represent formal government reporting and others are project specific or proprietary tools.

The continuous improvement methods employed by industry, such as cleaner production and total quality management, evaluate the level of goal achievement over time. Many of these approaches use a systems approach to examine resources so that there is minimal waste with the conversion of linear, once-through systems, to closed-loop systems that reduce waste. These are mainly improvement processes used by industry to become more sustainable at a particular industry level, as well as processes used by governments to align with agreed global initiatives and principles. In 1992, the first set of guiding principles emphasizing sustainable development, Agenda 21, were agreed to at the Earth Summit in Rio de Janeiro. In 2000, eight Millennium Development Goals were declared, and in 2015 countries adopted 17 Sustainable Development Goals. These principles and goals are not an assessment of sustainability in their own right. They inform assessments and provide direction.

Table 9.1 Sustainability reporting, improvement processes, tools and metrics

National/regional/community level	
• Genuine Progress Indicator	• Human Development Report and Index
• Living planet report	• Measuring Australia's progress
• Compass index of sustainability	
Corporate	
• Global Reporting Initiative	• Triple bottom line
• Environmental management systems (ISO 14000)	• Social responsibility (ISO 26000)
• Sustainable, responsible and impact investing	
Industry / Project / Product / Process	
• Cleaner production	• Total quality management
• Zero emissions	• Life cycle assessment (ISO 14040)
• Environmental impact assessment	• Eco-labelling
• Factor 4	• Green procurement
• Factor 10	• Natural capital
• Ecological/water footprints	• Water sensitive urban design
• Sustainability	• Green star building assessment
• LEED® building rating system	• Arup's Sustainable Project Appraisal Routine (SPeAR®)

UN Sustainable Development Goals

The UN Sustainable Development Goals are part of a broader sustainable development agenda, agreed upon by supporting countries in September 2015. The goals are

1. no poverty
2. zero hunger
3. good health and well-being
4. quality education
5. gender equality
6. clean water and sanitation
7. affordable and clean energy
8. decent work and economic growth
9. industry, innovation and infrastructure
10. reduced inequalities
11. sustainable cities and communities
12. responsible consumption and production
13. climate action
14. life below water
15. life on land
16. peace, justice and strong institutions
17. partnerships for the goals

Each of the goals has supporting facts and figures, and importantly targets for the year 2030.

http://www.un.org/sustainabledevelopment/

In many governments and organizations, a set of sustainability criteria are used to assess the relative merits of projects, processes and products. Included within these assessments is often interaction with the stakeholders and community groups through participatory and consultation programmes.

A major push for reduction of the use of resources comes in programmes such as Factor 10, which has the goal to lower the use of natural resources for generating material wealth, using new technology, by an average factor of ten within 30 to 50 years for the purpose of approaching sustainability. This is equivalent to increasing resource productivity tenfold over the same time period (Schmidt-Bleek, 1993). Sustainability assessment requires the development of a clear set of criteria against which the assessment can be conducted. This includes considerations of environmental, economic and social factors. The Triple Bottom Line (TBL) approach establishes individual indicators to show how well a company satisfies goals in each of these three areas. Usually for TBL reporting indicators are chosen for a number of criteria for each factor and then presented using a set of bars of different colours or stars to indicate how well a company is doing against particular criteria. The establishment of principles-based criteria for assessment of sustainability can improve the standard TBL approach.

A framework for assessing the natural capital within a region or within an organization can be developed with the use of systems such as the "Compass index of sustainability" to stress the *inter-connected* nature of the elements of environment, economics, and social well-being (AtKisson and Hatcher, 2001; Hargroves and Smith, 2005). Over time, the results of the planning and engineering of cities are measured by such indicators. Predictions of changes to the development of the city infrastructure and the resource input and output for a city's existence can also be made against these criteria. The design of individual buildings involves choices of materials to be used, lighting, air conditioning and water use. Each of these choices for achieving sustainability goals will involve assessment of embodied energy within materials used, energy use over the life of products, the ability of the product to be recycled and satisfaction of performance criteria.

Sustainability assessment does not replace Environmental Impact Assessments or functional decision making based on Life Cycle Analysis or Social Impact Analyses or combinations thereof. A sustainable assessment tool needs to be incorporated within the decision-making framework to ensure that decisions in planning are, in fact, sustainable (Pope et al., 2004). Sustainability assessment can and should be applied to evaluating proposed and existing processes and projects at all levels of governmental and project decision making. Incorporating sustainability through the design and implementation phases of a development could involve the integration of social and environmental effects of a longer time horizon into analyses by the proponents of projects. Sustainability assessment requires defining clear societal goals (e.g. the Sustainable Development Goals) which can be translated into criteria, against which assessment is conducted. It is essential that the assessment method is able to discern sustainable outcomes from unsustainable ones.

Ecological Footprint

Ecological Footprint (EF) is a measure of humanity's dependence on natural resources. For a certain population or activity, the EF measures the amount of productive land and water required to sustain the current level of resource usage for the production of goods and services and the assimilation of waste required by that population or activity.

In 2001, the world average EF was 2.55 global hectares per person, >30% above the current global capacity (Venetoulis and Talberth, 2005). The lifestyle of the average global citizen can therefore be considered as unsustainable. Natural resources are being used faster than they can be regenerated.

The size of an EF can change over time, depending on population, consumption levels, technology and resource use. EFs are measured in global area (hectares or acres). Wackernagel et al. (1997) defined biologically productive areas as (a) arable land; (b) pasture; (c) forest; (d) sea space (used by marine life); (e) built-up land; and (f) fossil energy land (land reserved for carbon dioxide absorption). The current global biologically productive area is 10.8 billion hectares, of which 21% is productive ocean and 79% is productive land. This represents less than one-quarter of the Earth's surface.

An individual's resource consumption is not restricted to local resources. The resources used by an individual are from around the world, clothes from China, cars from Korea and food from many different places. From a comparison of the Ecological Footprint to the existing biologically productive area, the sustainability of an activity, lifestyle or population can be determined.

An ecological footprint of Greater London, prepared by Best Foot Forward Ltd. (http://www.citylimitslondon.com/) found that London's EF was 42 times the current capacity, or 293 times the size of London. This equated to 49 million global hectares, twice the size of the UK, and roughly the same size as Spain. Is this sustainable?

Australia has a usage of 7.0 global hectares per person, making it the 5th highest in the world. There are many websites that enable calculation of your individual footprint on the earth (http://www.footprintnetwork.org/), and to compare it with different countries.

The Matrix Evaluation of Sustainability Achievement is one such method (Fleming and Daniell, 1995; UNEP, 2002) which uses a weighting technique against a set of sustainability criteria and combines them with fuzzy set analysis. The important step is the definition of specific criteria which can be obtained to achieve the overall project goal. In the case of a building, this might be to achieve a reduction of 50% of projected CO_2 emissions over an existing conventional design. Therefore, the inclusion of sustainability as a project goal could totally change the final solution from that resulting from a short-term economic goal.

Global Reporting Initiative

The Global Reporting Initiative (GRI) provides a reporting framework for companies and organisations to guide the preparation of consistent and comparable sustainability reports. First released in 1997, the GRI has been progressively developing resources as awareness of, and expectations associated with, corporate sustainability reporting have increased. GRI gained legitimacy in 2002 when it became a collaborating centre with the United Nations Environment Program (UNEP). It is aligned with the International Standard *ISO 26000:2010, Guidance on social responsibility* and also now with the United Nations Sustainable Development Goals. The GRI G4 guidelines will be phased out in 2018 and replaced by a series of standards. These standards are modular and define disclosure requirements using the following structure:

- General Disclosures
- Management Approach
- Economic Standards
- Environmental Standards
- Social Standards

In addition, there is a Foundation Standard and Glossary as supporting material.

Global Reporting Initiative, (2016)

9.4 ENVIRONMENTAL CONSIDERATIONS

History of Environmental Concerns

While it is true that the environment and sustainability have become much more of a focus for concern among people since the 1970s, it would be quite wrong to think that prior to that time there were no problems and concerns. In ancient Greece some 2500 years ago, Plato lamented the consequences of excessive logging and grazing in the mountainous region of Attica, near Athens. Similar concerns were raised around 1000 years ago in Japan where excessive eroded granite built up in Lake Biwa, due to logging for the construction of a temple (Parker, 1999).

Environmental legislation also dates from ancient times. In Rome all wheeled vehicles were prohibited between sunrise and two hours before sunset to improve the situation for pedestrians. This law fell into disuse, and in the 3rd Century AD, writers complained about the level of noise pollution generated by traffic in the streets. The city of Florence had laws governing the polluting of the rivers Arno, Sieve, and Serchio as early as 1477 (Higgins and Venning, 2001) and in 1810 Napoleon was issuing decrees aimed at eliminating polluting industries from the centres of cities.

In Australia, some of the earliest laws dealt with environmental issues. There were laws to protect the quality of the stream which supplied water for the original settlement in Sydney and it was prohibited to fell trees within 50 feet (15 metres) of the stream. People were not allowed to throw rubbish into the stream or have pigsties within a prescribed distance. In 1839, just three years after the city of Adelaide was settled, the state of the Torrens River was so bad that laws were passed to prevent, among other things, people driving cattle through it. This was the city's only water supply and yet it was treated with scant regard by many. There are many countries today where this still happens but, as the world becomes more urbanised, there is a need to be aware of the impact of human activity on the state of the planet Earth.

The Tragedy of the Commons

The tragedy of the commons has played itself out worldwide at various times and to varying degrees. It can be explained as follows: when a group of herdsmen have access to a common pasture, it is in each individual's best interests to increase the size of their herd without reference to the overall carrying capacity of the land. However, this leads to the destruction of the common pasture, so that everyone loses. This type of situation occurs in many ways, such as when an increasing population results in increased waste disposal into the commons: rivers, lakes, oceans, and atmosphere.

A major issue in city development involves the concept of the commons in regard to the decreasing quality of air and water due to emissions from transport systems and runoff from road systems. Another issue linked to the commons on a global scale is the consumption of energy by the developed nations which has resulted in producing CO_2 emissions which have contributed to global warming. Those activities that are considered to be necessary for a high quality of life: food; energy and transport systems; communication systems; the supply of infrastructure; supply and maintenance of water and waste water systems; health systems; and the management of waste all need energy and hence generate CO_2 emissions. There is much discussion on the linkages between CO_2 production, climate change and quality of life. The linkage between the use of energy and the quality of life can be examined by inspection of emissions of CO_2 and energy use in industrialised countries in the world. The United Nations Human Development Index (HDI) is a measure of poverty, literacy, education, life expectancy, childbirth, and other factors for countries worldwide. It is a means of measuring well-being and has been used since 1993 by the United Nations Development Programme (UNDP, 2013). To compare the impacts of development across nations with respect to emissions, the Human Development Index (HDI) has been used in conjunction with CO_2 emissions per capita as given in Table 9.2.

Table 9.2 World CO_2 emissions (per capita).

HDI rank 2013	Country	Carbon dioxide emissions per capita (tonnes) 2008	Carbon dioxide emissions annual growth (%) 1970/2008
1	Norway	10.5	1.0
2	Australia	18.6	1.2
3	United States	18	-0.4
4	Netherlands	10.6	-0.1
5	Germany	9.6	-
6	New Zealand	7.8	1.1
7	Ireland	9.9	1.1
8	Sweden	5.3	-2.0
9	Switzerland	5.3	-0.6
10	Japan	9.5	0.7
101	China	5.3	4.7
136	India	1.5	3.8
153	Nigeria	0.6	1.4

Source: United Nations Development Program, 2013.

Energy Use

Energy use in developing nations is based on traditional systems of wood and coal, and not electricity. As these nations convert to electricity and develop industries that use more energy, there will be a major increase in CO_2 production across the globe. The growth in energy use of the developing nations is only just beginning and can be expected to increase twenty fold in the next 20–30 years. The need to develop better technologies for the sustainability of cities and lower CO_2 emissions are two goals which engineers need to address if there is to be equity among nations.

Of course the future cannot be predicted with certainty, but engineers will help create it. The goals that are now chosen for all development will dictate the future. In the past, individual goals were pursued independently of each other, leading to the present crises for social well-being, the environment and the economies of some countries. The need is to develop a balance between the use of resources, the environment, and social well-being to achieve sustainability into the future.

9.5 ENVIRONMENTAL ASSESSMENT PROGRAMS AND TECHNIQUES

Environmental Impact Assessment (EIA) is now an integral part of the planning of an engineering project, just as economic, financial and social issues and technical analyses are. A definition of EIA that looks at the consequences is as follows: "Environmental Impact Assessment is a tool designed to identify and predict the impact of a project on the bio-geophysical environment and on society's health and well-being, to interpret and communicate information about the impact, to analyse site and process alternatives and provide solutions to sift out, or abate/mitigate the negative consequences on man and the environment" (UNEP, 2003).

Numerous countries have implemented EIA regulations from the late 1960s onwards. The origin of EIA stems from the National Environmental Policy Act 1969 (US), from which the practice has spread around the world. The Environment Protection (Impact of Proposals) Act 1974 (Commonwealth of Australia) was the first dedicated EIA legislation in the world and has now been replaced by Environment Protection and Biodiversity Conservation Act 1999 (Commonwealth of Australia). Within Australia there is also state legislation. For example, in Victoria the Environment Protection Act 1970 provides the legislative framework for environment protection in Victoria, but it is supplemented by other acts such as the Environmental Effects Act of 1978 for performing Environmental Effects Assessment.

EIA is thus the means of including environmental factors with both economic and technical considerations at the planning stage of a project. The process of EIA:

- identifies the potential environmental effects of undertaking a project;
- presents these environmental effects with the advantages and disadvantages of the proposed project to the decision makers; and
- forces stakeholders to consider the environmental effects and informs the public of the project while giving them the opportunity to comment on the proposed project.

EIAs are necessary for large projects and are required under relevant planning legislation. They are necessary for mining projects, new transport corridor projects, major infrastructure development, large industrial factories or expansion of existing facilities. The EIA is a means of highlighting potential environmental and ecological disturbances which would be more difficult and more expensive to correct after their occurrence than before. One major benefit of EIAs is the assessment of habitats of many rare species of flora and fauna which might never have been investigated had it not been for the requirement of the EIA.

In project planning, environmental issues are now addressed at all stages of the planning process, from conception to closure and rehabilitation. There are alternatives that can be considered at an early stage to counter adverse assessments, such as:

- abandoning a project or a process;
- proposing alternatives to projects which have extreme detrimental impacts on the environment; and
- proposing alternatives for projects not economically or financially viable.

Very few projects have been deemed not viable merely because of increased costs of environmental controls. For example, at a new paper pulp mill the additional cost has been assessed to be less than 3% of the initial investment (UNEP, 2003). The environmental controls that need to be put in place will vary from industry to industry, with significantly higher costs for some industries. Major research projects are being undertaken on the sequestering of carbon dioxide from energy production. A detailed assessment of the practices and methods involved in EIAs is contained in Canter and Sadler (1997) but the key elements of an EIA include:

- Scoping: identify key issues and concerns of interested parties;
- Screening: decide whether an EIA is required based on information collected;
- Identifying and evaluating alternatives: list alternative sites and techniques and the impacts of each;
- Forwarding measures that deal with risks and uncertainties of the proposed project and to minimise the potential adverse effects of the project; and
- Issuing a report called an environmental impact statement, which covers the findings of the EIA.

Environmental Impact Statements (EISs)

The EIA process produces a document, called the EIS, which provides information on the existing environment and predictions about the environmental effects which could flow from the proposal. Whether an EIS is to be done in its entirety or whether some other report is warranted by the proposal is decided by government legislation. In South Australia there is an independent statutory authority, the Major Developments Panel, which controls the level of reporting required: an Environmental Impact Statement which is required for the most complex proposals; or a Public Environmental Report which is required for a medium level of assessment, sometimes referred to as a “targeted EIS”; or a Development Report required for the least complex level of assessment.

The EIS is the document that reports the findings of the EIA and, depending on each country’s legislation, is now often required by law before a new project can proceed. A typical EIS has three parts with different levels of detail: an Executive Summary in a style that can be understood by the public; the main document containing relevant information regarding the project; and a volume containing the detailed assessment of significant environmental effects. If there are no significant effects either before or after mitigation, this volume will not be required. The EIS should

- describe the proposed action as well as alternatives;
- predict the nature and magnitude of the environmental effects;
- identify the relevant human concerns;
- list the impact indicators and determine the total environmental impact; and
- make recommendations for inspection procedures and alternatives to the plan.

As part of the EIA, as shown in Figure 9.3, there is usually a review of the EIS by the public and government to consider its accuracy and to recommend whether the proposal should go ahead or not. Different authorities and governments at the local, state or federal level have their own processes and procedures indicating timeframes for each review.

The EIS is normally carried out by the developer or proponent of the proposed development. The developer generally uses a consulting firm that would assemble a multi-disciplinary team to undertake the EIS. Although it might be argued that the developer might be biased, there are advantages in that the developer must pay for the EIS to be carried out, and since the developer has the relevant information it is more efficient for the developer to undertake the EIS. There is also the advantage

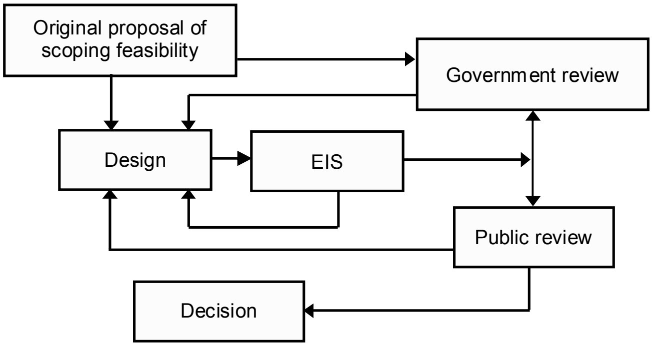


Figure 9.3 Simplified flowchart for an environmental impact assessment.

that the proponent can modify the design to reduce adverse impacts during the preparation of the EIS.

Most governments have regulations regarding EIAs and the reporting thereof, and give a formal procedure for the requirement of an EIA. For any significant project, an EIA would be required as part of the process of getting development approval. Generally a preliminary investigation, feasibility or scoping stage is performed prior to an EIA. It is very difficult to compare projects in terms of their environmental and social effects because of the many diverse factors involved. This contrasts with an economic comparison in which a single criterion such as NPV (see Chapter 8) gives an unambiguous ranking of projects. A number of techniques have been developed to assist in evaluating environmental and social factors. Some of these are listed in Table 9.3. Two of these that will be discussed here are the Battelle environmental index method and the Leopold matrix display technique. The Battelle method is a checklist method that uses scoring and weights. The Battelle and the Leopold methods have both been in use for many years. They are discussed here in order to illustrate the fundamental aspects of environmental assessments. Advanced modelling techniques allow greater complexity in assessments and modelling of scenarios.

The Battelle method

The Battelle method or Environmental index method was first designed for water resource development, (Dee et al., 1972; Dee et al., 1973; Martel and Lackey, 1977) but can be adapted for evaluation of a variety of types of projects. The principle lies in splitting the environmental impacts into four major categories: ecology; physical/chemical pollution; aesthetics; and human interest. These categories are divided into thematic data, as shown in Table 9.4. The full list is included in Appendix 9A.

Table 9.3 Main advantages and disadvantages of impact identification methods.

Impact Methods	Advantages	Disadvantages
Checklists -simple ranking and weighting	Simple to use and understand. Good for site selection. Good for priority setting.	Do not distinguish between direct and indirect impacts. Do not link action and impact. The process of incorporating values can be subjective, hence controversial.
Matrices	Links action and impacts. Good method for visual display of EIA results.	Difficult to distinguish the direct and indirect impacts. Potential for double-counting of impacts.
Networks	Links action to impacts. Useful in simplified form for checking for second order impacts. Handles direct and indirect impacts.	Can become very complex if used beyond simplified version.
Overlays	Easy to understand. Good display method. Good siting tool.	Address only direct impacts. Do not address impact duration or probability.
GIS and computer expert systems	Excellent for impact identification and analysis. Good for examining different scenarios-'experimenting'	Heavy reliance on knowledge and data. Often complex and expensive.

Source: (adapted from UNEP EIA Training Manual Edition 2, 2002).

These thematic data are divided into environmental indicators. For example, in a project discharging wastewater, the water pollution could be represented by: BOD; dissolved oxygen; faecal coliforms; inorganic carbon; pH; temperature; total dissolved solids; and turbidity of the receiving water and/or of the waste stream.

Table 9.4 Thematic types for the Batelle method.

Ecology (240)	Physical/ Chemical (402)	Aesthetics (153)	Human interest (205)
Species and populations	Water pollution	Land	Educational/scientific packages
Habitats and communities	Air pollution	Air	Historical packages
Ecosystems.	Land pollution	Water	Cultures
	Noise pollution	Biota	Mood/atmosphere
		Manmade objects	Life patterns

Note: the numbers in brackets are the weightings for each category.

Once the environmental indicators are chosen using Appendix 9A, the method follows three steps:

Step 1: transform environmental indicators into an environmental quality (EQ) rank. This usually requires expert advice to convert the environmental measurement to a scale of 0 to 1 (0 for poor quality and 1 for good quality). It is then possible to quantify environmental deterioration or improvement for the given project. A sample transformation is shown in Figure 9.4 for turbidity measured in Nephelometric Turbidity Units, transformed to the environmental quality index between 0 and 1.

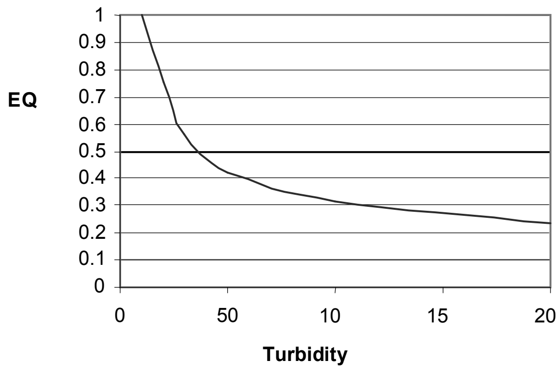


Figure 9.4 Environmental quality transformation.

Step 2: use the distributed Parameter Importance Units (PIU) which were developed by Dee et al. (1972). The relative importance of each parameter is reflected by distributing a total of 1,000 points among the indicators (e.g., a total of 240 for Ecology).

Step 3: complete the analysis for both the situation with and without the project in Environmental Impact Units (EIU). It can even reflect benefits or losses in terms of environmental conditions. The environmental impact is calculated using:

$$EI = \sum_{i=1}^{m} w_i \left[(V_i)_1 - (V_i)_0 \right] \tag{9.1}$$

where V_i is Environmental Quality (EQ) for each indicator i; the 1 subscript refers to the EQ with the project, and the 0 to the EQ without the project; w_i is the relative weight of indicator i; and m is the total number of indicators.

An advantage of this method is that it gives a comparative analysis of alternatives. Therefore, it is effective when a choice is to be made between projects. The problem of dealing with qualitative data is encountered in all assessments, and synthesis of information depends on the experience and professional judgement of the team undertaking the assessment and the stakeholders who have been consulted.

The Leopold matrix display technique

Matrix display techniques use a large matrix to summarize the environmental and social impacts of a proposed project, which may be given in a quantitative form. The Leopold matrix (Leopold et al., 1971) is an example. A disadvantage of the matrix display technique is that it can become difficult for users to absorb all of the information contained in the matrix.

The matrix lists a set of project actions in the first row and a set of environmental characteristics in the first column. Part 1 of Table 9.7 in Appendix 9A lists the project actions, while Part 2 lists the environmental characteristics. Other actions or environmental characteristics can be added as appropriate to give over 8800 possible interactions. For any project only some of the possible interactions will be relevant, but the actions and environmental characteristics do provide a useful checklist to ensure that all possible effects have been considered.

Each cell of the matrix for which an impact is likely to occur is identified and indicated by a diagonal line. An assessment is then made of the magnitude of the likely impact and a number between 1 and 10 is placed in the upper left-hand corner, with 1 representing the smallest magnitude and 10 the greatest (Leopold et al., 1971). This should be a factual assessment and not a value judgement. A number between 1 and 10 is placed in the lower right-hand corner of each cell to indicate the importance of the impact. This is a value judgement made by the evaluator. Leopold used the example of a proposal for the construction of highways and bridges. One consideration was that bridges may cause a large amount of bank erosion because geologic materials in an area are poorly consolidated. This may lead to the evaluator marking the magnitude of impact of highways and bridges on erosion at 6 or more. However, if the streams involved have high sediment loads and appear to be capable of carrying such loads without appreciable secondary effects, the effective importance of bridges through increased erosion and sedimentation will be relatively small and so marked as 1 or 2 in the lower right hand of the box. This would mean that while the magnitude of impact is relatively high, the importance of impact is not great.

Of course a matrix is just one example of summarizing the environmental impact of a proposed project. It must be supported by an environmental impact statement which discusses all impacts identified in the matrix. For example, Figure 9.5 shows a reduced impact matrix for a proposed phosphate mining lease in Los Padres National Forest, California (Leopold et al., 1971). The mining was to be an open-cut operation with ore processing on site, including crushing, leaching, and neutralization. The most important impact of the proposal was identified as the likely effect on the California condor, a rare and endangered species which exists in the region. The primary actions of concern were blasting and the increase in truck traffic, both of which were likely to disturb the nesting of the condor. In addition, sulphur fumes from mineral processing could have prevented the birds from landing to catch prey and hence present a danger to them.

These effects are shown in the row labelled “rare and unique species” as having a magnitude of 5 and an importance of 10. Other effects considered to be of moderate importance include the impact of industrial sites and buildings, highways and bridges, surface excavation, trucking, and the placement of tailings on the “wilderness qualities” of the area. The values in the boxes of a Leopold matrix are on an ordinal scale and cannot therefore be added or averaged. However, two or more projects can be compared in terms of the entries in an individual cell in the matrix to see which project will have the greater impact on that particular characteristic. Another approach, which avoids the use of numbers, is to use colour coding of the rankings and highlight those of importance with red flags.

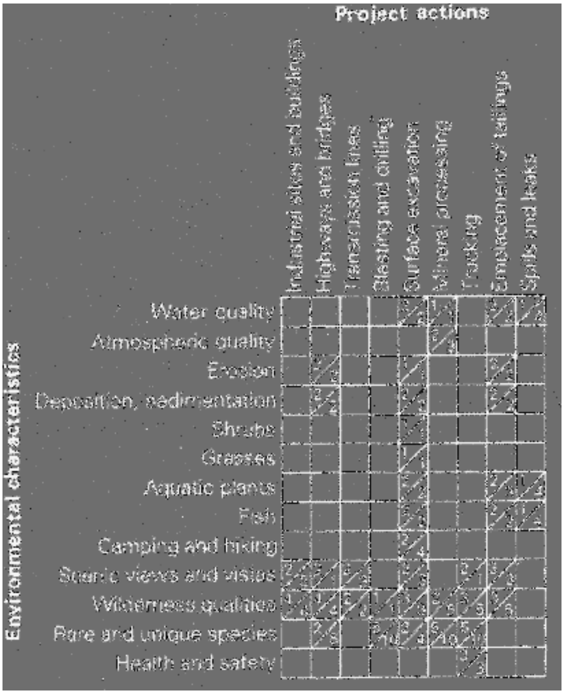


Figure 9.5 Abbreviated Leopold matrix for a phosphate mining lease adopted from (Leopold et al., 1971).

Multi-objective and multi-criteria assessment approaches

Multi-objective approaches have been developed to quantify economic, environmental and social aspects of a project. These methods use data in an integrated way and have the capability of dealing with trade-offs among objectives.

Sustainability principles have yet to permeate the full range of methods and techniques used for project evaluation and assessment but they are becoming increasingly important. In planning and design situations where more than one objective is relevant, it is not usually possible to identify a single best solution. A design which is better in terms of its economic performance might not be as good in terms of its environmental and social impact. A design which has low environmental impact may involve low economic and social benefits. Compromise is the essence of good planning and design and is reflected in the push towards sustainability.

The role of the evaluator in multiple objective planning is to identify the most efficient designs (in terms of all objectives) and elucidate the tradeoffs between them. A central concept is that of *inferior* and *non-inferior* designs or plans. An example, using two objectives, may be a project where national economic development and environmental quality are considered as the objectives. Using these objectives, all feasible designs for a project could be evaluated and plotted in two-dimensional space, as shown in Figure 9.6.

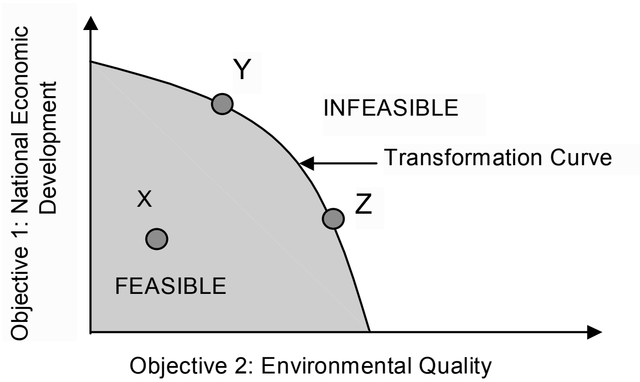


Figure 9.6 Multi-objective trade-off between alternatives.

By definition, an alternative A is inferior to alternative B if B ranks higher or equal to A for all objectives and higher for at least one objective. If A ranks higher than B for one objective and B higher than A for another, they are both non-inferior (for this pairwise comparison).

In Figure 9.6 design X is clearly inferior to design Y and can be discarded from further analysis. Designs Y or Z are non-inferior to each other and should therefore both be retained in the evaluation. Designs which are not inferior to any other are called "non-inferior" designs. A line joining all non-inferior designs is called the "Pareto Optimal frontier" or "transformation curve" (or "transformation surface" if there are more than two objectives). The transformation curve represents the boundary between feasible and infeasible designs in objective space.

The preferred design should lie on the transformation curve, provided all relevant objectives have been included in the analysis. The choice of a final design involves a value judgment as to the relative importance of the objectives and, in many cases, will involve a political decision.

The role of the evaluator in multiple objective planning is to identify the most efficient designs in terms of the objectives, and recognize the trade-offs among them. There are many techniques and software available for different applications and further considered in Section 12.7. An example of the development of a Multi-Objective Decision Support System (MODSS) for transport situations is described in some detail in Thoresen et al. (2001).

9.6 SOCIAL IMPACT

The environmental impact statement usually includes social impacts as well as effects on the physical and biological components of the environment. On the other hand, the USA Principles and Standards for Planning Water and Related Land Resources (USWRC, 1973) considered social well-being as a separate category of objectives. The social objective was defined as follows:

> *To enhance social well-being by the equitable distribution of real income, employment and population, with special concern for the incidence of the consequences of a plan on affected persons or groups; by contributing to the security of life, health and property; by providing educational, cultural and recreational opportunities; and by contributing to national security.* (USWRC, 1973)

If a full economic, environmental and social evaluation of a project is carried out, it is most important to avoid the double-counting of benefits or costs under more than one account. If, for example, an economic value is placed on lives saved through accident reduction and this is included in the *economic* evaluation of the project, it would not be appropriate to also highlight the number of lives saved in the *social* evaluation. Other examples of double-counting are discussed later in this section. The following types of social effects may need to be considered for a particular project:

- distribution of income;
- population distribution;
- employment;
- life, health, and safety;
- educational, cultural, and recreational opportunities;
- national security and emergency preparedness;
- displacement and relocation; and
- neighbourhood disruption and intrusion.

As with environmental effects, each social parameter can have its own ordinal or cardinal scale of measurement. The above effects are considered in turn.

Distribution of Income

In theory it is possible for a government to redistribute the benefits and costs of any public sector project among the various individuals or groups in society. This being the case, it can be argued that the role of the engineer or planner is to devise projects which maximise the total net benefits to the community and let the government decide how these benefits (and costs) will be distributed. However, this redistribution of benefits and costs is often not practicable because of the difficulties in identifying all affected groups and because of the administrative cost of carrying out the transfers. Redistribution of the benefits of engineering projects rarely occurs in practice.

It is therefore important that the distributional consequences of any major project be explicitly considered. Of particular relevance is the effect of the project on the income levels of certain target groups. These target groups may be distinguished on the basis of income (e.g. those living below the poverty level), race, sex, or geographic region.

It has been suggested by Weisbrod (1972) that the distributional question can be approached by placing explicit weights on the benefits and costs received by each group in the community. Traditional benefit-cost analysis gives the same weight to a dollar of benefits received by a pauper and by a millionaire. It is conceptually possible to give a higher weight to benefits or costs incurred by low-income earners and thus develop a single index of social welfare, using the following formula:

$$SW = \sum_{i=1}^{m} w_i [B_i - C_i] \qquad (9.2)$$

where SW is the index of social welfare; B_i is the present value of benefits received by group i; C_i is the present value of costs incurred by group i; w_i is the "social" weight of benefits or costs to group i; and m is total number of community groups. The difficulties of deciding on a set of weights and implementing this approach should be apparent.

Population Distribution

It has been argued that the redistribution of population throughout a region or the nation may be a major objective of government policy. This is particularly true in sparsely populated nations such as Australia. Major engineering projects in remote areas may contribute positively to the decentralization of population. The benefits of decentralization are related to enhanced national security and the benefits of rural versus urban living. High concentrations of population and industry are particularly vulnerable to attack by conventional or nuclear weapons. On the other hand, the settlement of remote areas leads to additional infrastructure costs for electricity supply, roads, railways, water supply, ports, and airfields.

Projects such as major irrigation developments which contribute to maintaining the size of the rural community may be supported on the grounds of the "healthier" lifestyle of country living. Continued growth of large cities at the expense of the surrounding country areas may also involve considerable increases in infrastructure costs in the city, as well as increases in the crime rate and other social problems. The negative aspects of population redistribution on individual families also need to be considered. These include the weakening of friendships and family ties and the stress associated with moving, setting up a new home, and adjusting to a changed lifestyle. The benefits of urbanization versus decentralization can vary from family to family depending on the independence of the family members.

Employment

One of the benefits cited by governments when new public works are announced is the effect that it will have on unemployment. A large engineering project may involve considerable employment in the implementation phase and also in the operations phase. Some people consider this an achievement in itself. In an economy with little or no unemployment, workers employed on a specific project will be taken from productive activities elsewhere in the economy. Therefore, the opportunity cost of their labour will be assessed as a *cost* to the project in the economic evaluation. In such a case, the number of people employed on the project is not a social benefit.

In an economy with high unemployment, some of the workers employed on the project may have been previously unemployed. As discussed in Section 8.8, previously unemployed labour has an opportunity cost to the project considerably less than the wage rate. If this effect is correctly taken into account in the economic evaluation of all projects, there is no need to further describe the employment implications of the projects, as this would be a double-counting of the same benefit.

The foregoing applies to public sector projects only. In the economic evaluation of private sector developments, all labour is costed at the actual wage rate paid. Thus, at times when there is less than full employment in the economy, the true net benefits of the project to society will be understated. In such cases it *is* valid to include the additional employment opportunities created by the project as part of the social evaluation.

The distributional consequences of employment opportunities may also be important; for example, in the creation of employment opportunities in depressed regions of the country or among disadvantaged groups such as ethnic, or low-income groups.

Life, Health, and Safety

Many engineering works contribute to saving lives, improving health, or increasing safety in the community. Typical examples are flood mitigation works, water treatment plants, sewage treatment, improved road alignments, grade separation, and coastal protection works. Attempts have been made to include the benefits of lives saved into the economic evaluation of public sector projects by placing an economic value on human life. Various methods have been suggested for estimating the economic benefits of saving human life. These include:

- the present value of the person's expected future earnings;
- a value imputed from political decisions which involve public expenditure aimed at reducing the number of deaths. For example, the government spends \$50 m in equivalent annual worth to upgrade railway crossings which will save an estimated 100 lives per year from fatal accidents. The implied value of saving one life is greater than or equal to \$50 m/100 (i.e. \$0.5 m per live saved); and
- a dollar amount which people are prepared to accept as compensation in order to put up with the additional risk of death involved in the project. For example, a proposed new chemical plant will increase the probability of death per year to each person in an adjoining town by 1 in 10,000. The economic cost of this increased risk is the sum over all people in the town of the minimum level of compensation which each person is prepared to accept in order to tolerate the increased risk. The same basis can be used for projects which *reduce* the probability of death for part of the community.

Practical and theoretical difficulties exist with all of these economic approaches. An alternative approach is to estimate the number of lives saved per year and treat this as a social benefit. Similar arguments apply to projects which increase or reduce the risk to public health or the chance of accidents occurring. Either a dollar value may be used in the economic evaluation or an estimated reduction in the risk may be included in the social evaluation.

Educational, Cultural, and Recreational Opportunities

It has already been stated that the distribution of income, and in particular to certain disadvantaged groups, is an important social effect. Similarly the distribution of educational, cultural, and recreational opportunities among regions and socio-economic groups may be an important consideration in evaluating a major project. For example, a major dam may provide a new recreational resource for fishing, boating, picnicking, and swimming (depending on the final end use of the water). The social benefits of these recreational uses may be as important as the economic benefits of the water provided by the dam. In some cases it may be advantageous to choose a dam site close to a population centre in order to increase the recreational opportunities for a particular group of people.

Educational and cultural values may be enhanced by engineering projects providing improved access to sites of historical, archaeological, or scientific interest. For example, a new road along the coast may provide improved access to some unique geological formations which are of considerable interest to the public, but controls need to be enforced; otherwise, the project could lead to the ruin of the formation.

National Security and Emergency Preparedness

A new road or railway could have important consequences for national security. Such benefits are difficult to quantify in economic terms and, in fact, only become apparent in times of armed conflict. However, they may be an important social consideration in the evaluation of a project.

Emergency preparedness requires, for example, the provision of a flexible water supply, electricity grid, and road and rail network. The provision of a single dam to supply water to a city or a single power station to supply electricity may be economically efficient under normal circumstances but could be disastrous in the event of a catastrophe such as major technical failure, earthquake, hurricane, nuclear explosion, or act of sabotage. The provision of some redundancy in all engineering systems is a good rule to follow.

Displacement and Relocation

Engineering projects in urban areas may involve the displacement and relocation of some houses, and commercial or industrial buildings. A typical example arises from the land acquisition associated with an improved transport link, such as a freeway, arterial road, tramway, or railway. It is common practice for the households or firms displaced to be suitably compensated for the loss of their land, home, shop, or factory.

It is often argued that the loss of physical assets is suitably compensated for by the payment of fair market value. However, such an argument ignores the fundamental distinction between value in exchange and value in use. Consider the market for houses of a particular age and quality. If the current market price for such houses is P_o, this may be called the “value in exchange of the property”. Clearly all homeowners who are prepared to sell at this price or less will put their houses on the market. However, there are many homeowners who are not willing to sell at the price P_o. To these people, the value in use of the house exceeds its value in exchange. If a road authority were to compulsorily acquire such a house and pay compensation equal to P_o, the homeowner would clearly suffer a loss of utility. Full compensation should cover the value in use of the asset and would usually exceed P_o by a significant margin. A similar argument applies to commercial and industrial properties.

The difference between value in use and value in exchange of a house includes an allowance for the stress of moving, disruption of social ties, the cost of finding a new house, and changes in accessibility and travel costs at the new location. In practice such costs are difficult to assess and considerable negotiation may be required with the property owners.

The compensation costs for displaced households and firms are a true opportunity cost to a project and should be included in the economic evaluation, rather than the social evaluation.

Neighbourhood Disruption and Intrusion

An engineering project may involve significant disruption and/or intrusion into a community. This is particularly true of transportation corridors which can sever a cohesive community or produce intrusive impacts in the form of noise, vibration, air pollution, or aesthetic degradation. A sense of community is often very strong and may be centred on facilities such as schools, shopping centres, the town hall, or parks. Proposals which isolate one part of the community from another may receive considerable opposition from local community groups. Transport corridors should be planned to provide the minimum disruption to communities and ensure that adequate safe crossings such as overpasses or underpasses are provided for pedestrians and cyclists.

The principle of compensation described in the previous section can be applied to neighbourhood disruption and intrusion, but it is very difficult to define the loss of utility for an entire community. Attempts have been made to assess the economic effects of noise due to airports, highways, or railways by assessing the lower property values adjacent to these facilities. If such an approach is taken, these social effects can be included as a cost in the economic evaluation.

9.7 SOCIAL ASSESSMENT TOOLS AND METHODS

An extremely important facet of undertaking any project analysis involves the participation of interested stakeholders. This is an area where many engineers will be called on to manage and participate with specialists who work in the social assessment area. A basic knowledge of the methods available for undertaking such project analyses is given in the UNEP EIA Training Manual (UNEP, 2002). A number of methods are described below which allow social concerns to be discovered and to determine local knowledge concerning a proposed project.

Analytical tools

Stakeholder analysis is an entry point to Social Impact Assessment (SIA) and participatory work. It addresses strategic questions, e.g. Who are the key stakeholders? What are their interests in the project or policy? What are the power differentials between them? What relative influence do they have on the operation?

This information helps to identify institutions and relationships which, if ignored, can have negative influence on proposals or, if considered, can be built upon to strengthen them.

Gender analysis focuses on understanding and documenting the differences in gender roles, activities, needs and opportunities in a given context. It highlights the different roles and behaviour of men and women. These attributes vary across cultures, class, ethnicity, income, education, and time; and so gender analysis does not treat men or women as a homogeneous group.

Data Review of information from previous work is an inexpensive, easy way to narrow the focus of a social assessment, to identify experts and institutions that are familiar with the development and context, and to establish a relevant framework and key social variables in advance for the project.

Community-based methods

The participatory approach aims to ascertain local knowledge and actions. It uses group exercises to enable stakeholders to share information and to develop plans. These techniques have been employed successfully in a variety of settings to enable local people to work together to plan community-appropriate developments. Attributes of self-esteem, associative strength, resourcefulness, action planning and responsibility for follow-through are important to achieve a participatory approach to development. Generally there is a philosophy of empowerment of the stakeholders to enable people to adopt responsibility for outcomes. It can best be described as development of teambuilding skills and learning from local experience rather than from external experts. Other participatory consultation methods include selecting a sample of stakeholders to ensure that their concerns are incorporated into the assessment. This selection is for the purposes of giving voice to the poor and other disadvantaged stakeholders.

Other Participatory Methods

Role playing helps people to be creative, open their perspectives, understand the choices that another person might face, and make choices free from their usual responsibilities. This exercise can stimulate discussion, improve communication, and promote collaboration at both community and agency levels.

Wealth ranking (also known as "well-being ranking" or "vulnerability analysis") is a visual technique to engage local people in the rapid data collection and analysis of social stratification in a community (regardless of language and literacy barriers). It focuses on the factors which constitute wealth, such as ownership of or right to use productive assets/resources, their relationship to locally powerful people, labour and indebtedness.

Mapping is an inexpensive tool for gathering both descriptive and diagnostic information. Mapping exercises are useful for collecting baseline data on a number of indicators as part of a beneficiary assessment or rapid appraisal, and can lay the foundation for community ownership of development planning by including different groups and making them aware of the implications of the project.

Needs Assessment draws out information about people's needs and requirements in their daily lives. It raises participants' awareness of development issues and provides a framework for prioritising actions and interventions. All sectors can benefit from participating in a needs assessment, as can trainers, project staff and field workers.

Tree Diagrams are multi-purpose, visual tools for narrowing and prioritising problems, objectives or decisions. Information is organized into a tree-like diagram. The main issue is represented by the trunk, and the relevant factors, influences and outcomes are shown as roots and branches of the tree. Other techniques such as mind mapping, as discussed in Chapter 6, can also be used for this purpose.

Observation and interview tools

Focus group meetings are a rapid way to collect comparative data from a variety of stakeholders. They are brief meetings - usually one to two hours - with many potential uses, e.g. to address a particular concern; to build community consensus about implementation plans; to cross-check information with a large number of people; or to obtain reactions to hypothetical or intended actions.

Workshop-based methods encourage participatory planning and analysis throughout the project life cycle. A series of stakeholder workshops tend to be held to set priorities, and integrate them into planning, implementation and monitoring. Building commitment and capacity is an integral part of this process (UNEP, 2002). These stakeholder workshops are in common use for many infrastructure and resource projects throughout the world.

9.8 SUMMARY

Engineers have a primary responsibility to society when developing infrastructure and this can override their responsibility to the client, depending on the Code of Ethics being applied. The responsibility entails evaluating environmental and social effects of the project. Many consulting engineering firms are required to perform Environmental Impact Assessments for projects that they plan and design. In many cases, the engineer will be actively involved in stakeholder consultations, giving presentations to community groups as well as professional groups. To do this, communication skills and knowledge of the environmental assessment process are essential. Future work for both private development and for government infrastructure will be undertaken using sustainability principles to guide the development of the project. The role of the engineer in planning is to identify the most efficient designs, in terms of all the objectives, and recognize the trade-offs among them.

Professional engineering institutions worldwide have embedded sustainability within their charters and codes of ethics, but progress to achieve these ideals has in the past been thwarted by little political and legislative assistance, but this is changing. The production of CO_2 and the linkage to climate change are the main driving forces for the way industry and government are viewing sustainability. The integration of sustainable energy systems with a process of closed-loop systems for resource use is seen as a key element in approaching sustainability.

PROBLEMS

9.1 Locate three reports which are aligned with the Global Reporting Initiative. Compare and contrast the data contained within them. How readily does the data enable comparisons between the different organisations?

9.2 Select two of the 2015 Sustainable Development Goals and explain how progress towards the targets could be measured. How could these measurements inform the overall assessment of progress?

9.3 A study is being carried out on alternative energy sources for the future generation of electricity in your state. The energy sources include coal, natural gas, nuclear, solar and wind energy (and combinations thereof). List and describe the environmental and social effects which would need to be considered when comparing these alternatives. Illustrate use of the Leopold matrix display technique by deriving the matrix relevant to the use of coal as the primary energy source for your city, mined from a regional centre and transported by rail.

9.4 A winery is proposing to expand its operations with the development of a new processing facility. The winery is confident with its proposal since it includes some additional treatment of the wastewater before it is discharged into a nearby waterway, although the machinery used will create increased noise in the quiet rural setting and some venting of NO and NO_2 (reported as nitrous oxides, NO_x). Details of the air and water discharges, and the noise produced are given in Table 9.5. The results of an expert panel assessing the environmental quality of the relevant parameters are included in Figure 9.7.

Table 9.5 Discharge estimates for winery and its proposed development.

Water Quality Parameter	Existing Operations	Option 1
Turbidity (NTU)	30	20
Temperature (°C)	20	22
DO (mg/L)	8	6
Ph	8	7
Noise (dB)	5	20
Nitrous Oxides (NO_x) (mg/L)	5	15

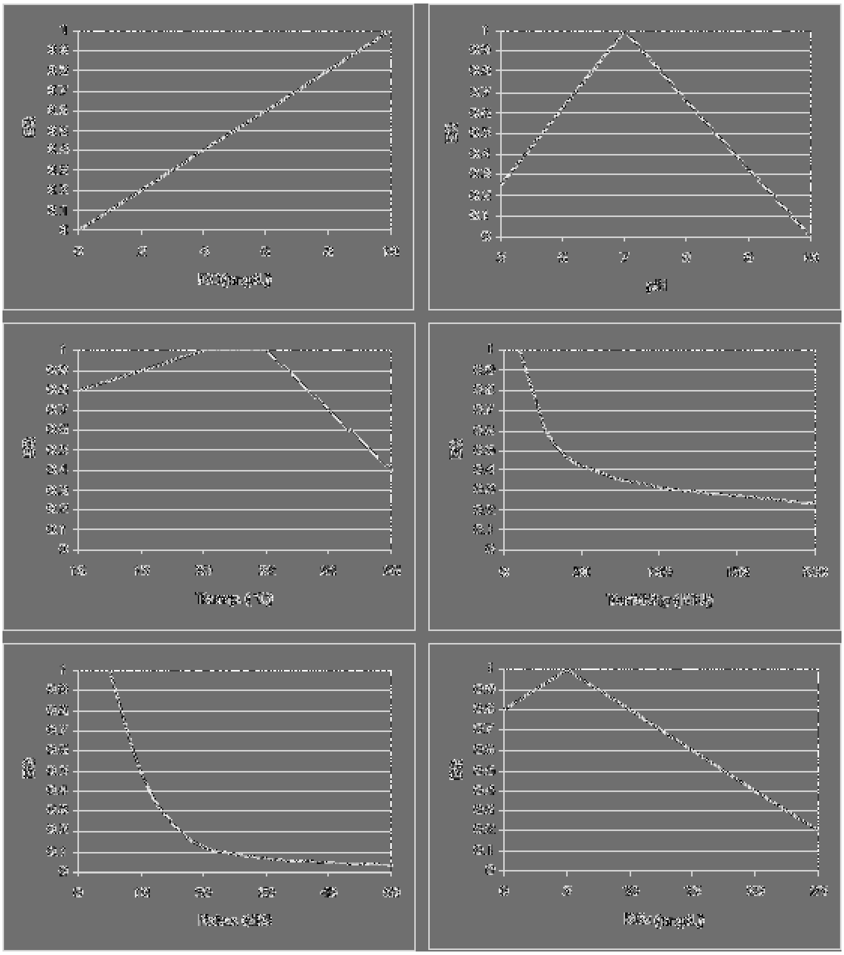


Figure 9.7 Environmental quality of various environmental parameters.

Use the Environmental Index System to suggest whether, on the basis of the environmental information provided, the winery should be allowed to expand its operations. List, and discuss in 20 words or less (each), the six most important social issues that should be considered in the environmental assessment of the proposal.

9.5 In 1974, Ehrlich and Ehrlich published a book, *The End of Affluence*, in which they proposed a "nutritional disaster that seems likely to overtake humanity in the 1970s (or, at the latest, the 1980s). Before 1985 mankind will enter a genuine age of scarcity" in which "the accessible supplies of many key minerals will be nearing depletion." Comment on why this did not happen and whether it will occur by 2020 as many environmental activists are predicting?

9.6 A wave energy farm is to be constructed 1 km off shore to supply both electrical energy and desalinated water to a coastal community of 2,000 people. You are to prepare a scoping study which would include the activities of the development with the potential environmental impacts and suggestions of mitigating solutions to these impacts. There are a number of wave energy systems being trialled around the world. Choose two of these systems and compare with each other.

9.7 Search environmental modelling journals and find two papers, published since 2015, which describe an environmental modelling approach. How do these compare with the early models described in the chapter?

REFERENCES

AtKisson, A. and Hatcher, R.L., 2001, The compass index of sustainability: Prototype for a comprehensive sustainability information system, *Journal of Environmental Assessment, Policy and Management*, 3(4), 509–532.

Canter, L. C. and Sadler, B.B., 1997, *A Tool Kit for Effective EIA Practice — Review of Methods and Perspectives on their Application*, International Association for Impact Assessment, Fargo, North Dakota, U.S.A.

Dee, N., Baker, J.K., Drobny, N. L., Duke, K.M. and Fahringer, D.C., 1972, Environmental Evaluation System for Water Resource Planning, Final Rep., Batelle Columbus Labs, Columbus, Ohio, USA. 188pp.

Dee, N., Baker, J., Drobny, N. and Duke, K.M., 1973, An Environmental Evaluation System for Water Resource Planning, *Water Resources Research* 9, 523–535.

Fleming, N.S., 1999, *Sustainability and Water Resources Management for the Northern Adelaide Plains*, *South Australia*, PhD Dissertation at the University of Adelaide.

Fleming, N.S. and Daniell, T.M., 1995, Matrix for Evaluation of Sustainability Achievement (MESA): Determining the Sustainability of Development, National Environmental Engineering Conference 1995, Towards a Sustainable Future: Challenges and Responses, The Institution of Engineers Aust. Preprints of Papers, 99–106.

Foley, B.A., Daniell, T.M., and Warner, R.F., 2003, What is Sustainability and can it be measured*? Australian Journal of Multidisciplinary Engineering*, 1(1), 1–8.

Gilman, R., 1992, *Sustainability,* from the 1992 UIA/AIA Call for Sustainable Community Solutions. Retrieved 4 February 2003. http://www.context.org/ICLIB/DEFS/AIADef.htm

Global Reporting Initiative, 2016, GRI Standards, https://www.globalreporting.org/standards/, accessed November 2016.

Hargroves, K., and Smith, M. H., 2005, *The Natural Advantage of Nations (Vol. I): Business Opportunities, Innovation and Governance in the 21st Century,* The Natural Edge Project, Earthscan/James & James.

Higgins, J. and Venning, J. (eds), 2001, *Towards Sustainability: Emerging Systems for Informing Sustainable Development,* Sydney, Australia: University of New South Wales Press, 240 pp.

Leopold, L. B., Clarke, F.E., Hanshaw, B.B., and Balsley, J.R., 1971, *A Procedure for Evaluating Environmental Impact*. US Geological Survey Circular 645.

Martel, Gary and Lackey, Robert, 1977, A computerized Method for Abstracting and Evaluating Environmental Impact Statements, Bulletin 105 Virginia Water Resources Research Center, Virginia Tech, Dec.

Meadows, D.H., Meadows, D.L., Randers, J. and Behrens III, W.W., 1972, *The Limits to Growth*, A Report to The Club of Rome, Universe Books, New York. Available at: https://www.clubofrome.org/report/the-limits-to-growth/

Parker, G., 1999, Progress in the Modeling of Alluvial Fans. *Journal of Hydraulic Research,* 37(6), 805–825.

Pope, J., Annandale, D. and Morrison-Saunders, A., 2004, Conceptualising sustainability assessment, *Environmental Impact Assessment Review*, 24, 595–616.

Roberts, D.V., 1990, *Sustainable Development: A Challenge for the Engineering Profession*. Paper presented at the International Federation of Consulting Engineers Annual Conference, Oslo, Norway, September.

Schmidheiny, S., 1992, *Changing Course: A Global Business Perspective on Development and the Environment*, MIT Press, Cambridge, MA.

Schmidt-Bleek, F., 1993, Wieviel Umwelt braucht der Mensch – mips, das ökologische Mass zum Wirtschaften, Birkhäuser, Basel, Boston, Berlin, English version in www.factor10-institute.org under the title "The Fossil Makers", accessed 23 Sept 2006.

Stahel, W.R., 2016, Circular Economy, *Nature*, 531, 435-438.

Suter, K., 1999, About the Club of Rome, Australian Broadcasting Commission, Retrieved Sept 11, 2005. http://www.abc.net.au/science/slab/rome/rome.htm.

Thoresen, T., Tsolakis, D., Houghton, N., 2001, *Austroads Project Evaluation Compendium,* ARRB Transport Research, Austroads Project No. RSM.F.N.021, Austroads Publication No. AP–R191/01

United Nations Development Program (UNDP), 2013, Human Development Report 2013, Available at http://hdr.undp.org/en/content/human-development-report-2015-work-human-development.

UNEP, 2002, EIA Training Resource Manual, 2nd Edition 600pp.

UNEP, 2003, Environmental Management Tools, Life Cycle Analysis, Retrieved 21 Sept 2005, http://www.uneptie.org/pc/pc/tools/lca.htm.

USWRC, 1973, *Principles and standards for planning water and related land resources*. United States Water Resources Council, Federal Register, 38, Sept. 10, Washington D.C.

Venetoulis, J. and Talberth, J., 2005, Ecological Footprint of Nations, 2005 Update, Redefining Progress, www.RedefiningProgress.org, 16p.

Weisbrod, B.A., 1972, Deriving an implicit set of governmental weights for income classes. In Cost Benefit Analysis, R. Layard (ed.), 395–428, London, Penguin.

Wackernagel, M. and Rees, W., 1997, *Our Ecological Footprint: Reducing Human Impact on the Earth*, New Society Publishers.

World Commission on Environment and Development (WCED), 1987, *Our Common Future*, Australian Edition. New York: Oxford University Press.

APPENDIX 9A THE BATTELLE CLASSIFICATION AND LEOPOLD MATRIX

Table 9.6 The Battelle Environmental Classification for Water-Resource Development Projects (Dee et al., 1973). The Bracketed Numbers are the Distributed Parameter Importance Units, i.e. Relative Weights.

ECOLOGY (240)
Terrestrial Species & Populations
-Browsers and grazers (14)
-Crops (14)
-Natural vegetation (14)
-Pest species (14)
-Upland game birds (14)
Aquatic Species & Populations
-Commercial fisheries (14)
-Natural vegetation (14)
-Pest species (14)
-Sport fish (14)
-Water fowl (14)
Terrestrial Habitats & Communities
-Food web index (12)
-Land use (12)
-Rare & endangered species (12)
-Species diversity (14)
Aquatic Habitats & Communities
-Food web index (12)
-Rare & endangered species (12)
-River characteristics (12)
-Species diversity (14)
Ecosystems
-Descriptive Only

PHYSICAL/CHEMICAL (402)
Water Quality
-Basin hydrologic loss (20)
-Biochemical oxygen demand (25)
-Dissolved oxygen (31)
-Faecal coliforms (18)
-Inorganic carbon (22)
-Inorganic nitrogen (25)
-Inorganic phosphate (28)
-Pesticides (16)
-pH (18)
-Streamflow variation (28)
-Temperature (28)
-Total dissolved solids (25)
-Toxic substances (14)
-Turbidity (20)
Air Quality
-Carbon monoxide (5)
-Hydrocarbons (5)
-Nitrogen oxides (10)
-Particulate matter (12)
-Photochemical oxidants (5)
-Sulphur oxides (10)
-Other (5)
Land Pollution
-Land use (14)
-Soil erosion (14)
Noise Pollution
-Noise (4)

AESTHETICS (153)
Land
-Geologic surface material (6)
-Relief & topographic character (16)
-Width and alignment (10)
Air
-Odour and visual (3)
-Sounds (2)
Water
-Appearance of water (10)
-Land & water interface (16)
-Odour and floating material (6)
-Water surface area (10)
-Wooded /geologic shoreline (10)
Biota
-Animals -domestic (5)
-Animals -wild (5)
-Diversity of vegetation types (9)
-Variety within vegetation types (5)
Man-Made Objects
-Man-made objects (10)
Composition
-Composite effect (15)
-Unique composition (15)

HUMAN INTEREST /SOCIAL (205)
Education/Scientific
-Archeological (13)
-Ecological (13)
-Geological (11)
-Hydrological (11)
Historical
-Architecture and styles (11)
-Events (11)
-Persons (11)
-Religions and cultures (11)
-'Western Frontier' (11)
Cultures
-Indians (14)
-Other ethnic groups (7)
-Religious groups (7)
Mood/ Atmosphere
-Awe/inspiration (11)
-Isolation/solitude (11)
- Mystery (4)
-'Oneness' with nature (11)
Life Patterns
-Employment opportunities (13)
-Housing (13)
-Social interactions (11)

Table 9.7 The Leopold Matrix (Leopold et al., 1971). Part I Lists the Project Actions; Part 2 Lists the Environmental Characteristics and Conditions

PART 1: Project Actions (Horizontal Axis of Matrix)

A. MODIFICATION OF REGIME
a) Exotic flora or fauna introduction
b) Biological Controls
c) Modification of habitat
d) Alteration of ground cover
e) Alteration of ground-water hydrology
f) Alteration of drainage
g) River control and flow codification
h) Canalization
i) Irrigation
j) Weather modification
k) Burning
l) Surface or paving
m) Noise and vibration

B. LAND TRANSFORMATION AND CONSTRUCTION
a) Urbanization
b) Industrial sites and buildings
c) Airports
d) Highways and bridges
e) Roads and trails
f) Railroads
g) Cables and lifts
h) Transmission lines, pipelines and corridors
i) Barriers, including fencing
j) Channel dredging and straightening
k) Channel revetments
l) Canals
m) Dams and impoundments
n) Piers, seawalls, marinas, and sea terminals
o) Offshore structures
p) Recreational structures
q) Blasting and drilling
r) Cut and fill
s) Tunnels and underground structures

C. RESOURCE EXTRACTION
a) Blasting and drilling
b) Surface excavation
c) Sub-surface excavation
d) Well drilling and fluid removal
e) Dredging
f) Clear cutting and other forestry
g) Commercial fishing and hunting

D. PROCESSING
a) Farming
b) Ranching and grazing
c) Feed lots
d) Dairying
e) Energy generation
f) Mineral processing
g) Metallurgical industry
h) Chemical industry
i) Textile industry
j) Automobile and aircraft
k) Oil refining
l) Food
m) Lumbering
n) Pulp and paper
o) Product storage

E. LAND ALTERATION
a) Erosion control and terracing
b) Mine sealing and waste control
c) Strip mining rehabilitation
d) Landscaping
e) Harbour dredging
f) Marsh fill and drainage

F. RESOURCE RENEWAL
a) Reforestation
b) Wildlife stocking and management
c) Groundwater recharge
d) Fertilization application
e) Waste recycling

G. CHANGES IN TRAFFIC
a) Railway
b) Automobile
c) Trucking
d) Shipping
e) Aircraft
f) River and canal traffic
g) Pleasure boating
h) Trails
i) Cables and lifts
j) Communication
k) Pipeline

H. WASTE DISPOSAL AND TREATMENT
a) Ocean dumping
b) Landfill
c) Emplacement of tailings, spoil and overburden
d) Underground storage
e) Junk disposal
f) Oil-well flooding
g) Deep-well emplacement
h) Cooling-water discharge
i) Municipal waste discharge including spray irrigation
j) Liquid effluent discharge
k) Stabilization and oxidation ponds
l) Septic tanks
m) Stack and exhaust emission
n) Spent lubricants

I. CHEMICAL TREATMENT
a) Fertilization
b) Chemical de-icing of highways, etc.
c) Chemical stabilization of soil
d) Weed control
e) Insect control (pesticides)

J. ACCIDENTS
a) Explosions
b) Spills and leaks
c) Operational failure

PART 2: Environmental Characteristics and Conditions (Vertical Axis of Matrix)

A. PHYSICAL AND CHEMICAL CHARACTERISTICS

1. *Earth*
a) Mineral resources
b) Construction material
c) Soils
d) Landform
e) Force fields and background radiation
f) Unique physical features

2. *Water*
a) Surface
b) Ocean
c) Underground
d) Quality
e) Temperature
f) Snow, Ice, and permafrost

3. *Atmosphere*
a) Quality (gases, particulates)
b) Climate (micro, macro)
c) Temperature

4. *Processes*
a) Floods
b) Erosion
c) Deposition (sedimentation, precipitation)
d) Solution
e) Sorption (ion exchange, complexing)
f) Compaction and settling
g) Stability (slides, slumps)
h) Stress-strain (earthquake)
i) Recharge
j) Air movements

B. BIOLOGICAL CONDITIONS

1. *Flora*
a) Trees
b) Shrubs
c) Grass
d) Crops
e) Microflora
f) Aquatic plants
g) Endangered species
h) Barriers
i) Corridors

2. *Fauna*
a) Birds
b) Land animals including reptiles
c) Fish and shellfish
d) Benthic organisms
e) Insects
f) Microfauna
g) Endangered species
h) Barriers
i) Corridors

C. CULTURAL FACTORS

1. *Land use*
a) Wilderness and open spaces
b) Wetlands
c) Forestry
d) Grazing
e) Agriculture
f) Residential
g) Commercial
h) Industrial
i) Mining and quarrying

2.*Recreation*
a) Hunting
b) Fishing
c) Boating
d) Swimming
e) Camping and hiking
f) Picnicking
g) Resorts

3. *Aesthetics and Human Interest*
a) Scenic views and vistas
b) Wilderness qualities
c) Open space qualities
d) Landscape design
e) Unique physical features
f) Parks and reserves
g) Monuments
h) Rare & unique species or ecosystems
i) Historical or archaeological sites and objects
j) Presence of misfits

4. *Cultural Status*
a) Cultural patterns (lifestyle)
b) Health and safety
c) Employment
d) Population density

5. *Man-Made Facilities and Activities*
a) Structures
b) Transportation network (movement, access)
c) Utility networks
d) Waste disposal
e) Barriers
f) Corridors

D. ECOLOGICAL RELATIONSHIPS
a) Salinization of water resources
b) Eutrophication
c) Disease-insect vectors
d) Food chains
e) Salinization of surficial material
f) Brush encroachment
g) Other

CHAPTER TEN

Ethics and Law

As engineering work is concerned with creating and improving the physical infrastructure, it is inevitable that engineering decisions often involve value judgements that raise questions concerning ethics, integrity and the law. In this chapter we discuss how ethics and legal obligations influence decisions that engineers make. The difference between the civil laws of tort and contract law is explained. In order to make responsible decisions, professional engineers must consider their obligations to society, employers, and to fellow employees. They also must take account of the many laws and regulations that apply to engineering work. In some circumstances the different obligations and laws and regulations can be mutually incompatible. Sometimes, litigation becomes the last resort to solving problems involving ethics and the law.

10.1 ETHICS IN ENGINEERING

Many professions, including the medical, the engineering and the legal professions, have developed codes of ethics to guide the behaviour of practitioners. Some of these codes are entrenched in law, so that violations of the code are subject either to civil or criminal penalty, while other codes involve an organisational penalty such as loss of licence or membership. Other codes are advisory or subject to enforcement by the promulgating institution.

Throughout the world there are many different codes of ethics that set standards for professional engineers. The Centre for the Study of Ethics in the Professions lists more than 75 organisations or institutions with codes of ethics that are applicable to the engineering profession (The Illinois Institute of Technology, 2016). Interest boxes in this chapter list the key elements in ethical codes that are used by the engineering profession in Australia (Engineers Australia, 2010), the United Kingdom (Engineering Council UK, 2014), and the United States (NSPE, 2007).

In Australia, Engineers Australia's Code of Ethics (2010) is largely derived from the World Federation of Engineering Organisations (WFEO) Model Code of Ethics. It states that engineers have obligations to society, to employers and clients, and to other engineers. Engineers are expected to adopt the standards spelt out in the code, to regulate their work habits and relationships, as well as the legal and moral standard that a professional engineer should adhere to.

In the United Kingdom, Chartered Engineers and Incorporated Engineers are expected to observe the Code of Conduct of the engineering institution that they join. Each UK Institution (Civil, Mechanical, Chemical, Structural, etc.) follows the Engineering Council guidelines for the Code of Conduct which stresses that engineers should act with integrity regarding all issues associated with the profession, from dealing with the public, their fellow workers, the environment and all business operations.

Code of Ethics
(Engineers Australia, 2010)

The Engineers Australia Code of Ethics is designed to define the values and principles that shape decisions made by engineers. The code is complemented by Guidelines on Professional Conduct. Members of Engineers Australia can be held accountable to the code through Engineers Australia's disciplinary regulations. The code states:

In the course of engineering practice we will:

1. Demonstrate integrity
   1.1 Act on the basis of a well-informed conscience
   1.2 Be honest and trustworthy
   1.3 Respect the dignity of all persons

2. Practise competently
   2.1 Maintain and develop knowledge and skills
   2.2 Represent areas of competence objectively
   2.3 Act on the basis of adequate knowledge

3. Exercise leadership
   3.1 Uphold the reputation and trustworthiness of the practice of engineering
   3.2 Support and encourage diversity
   3.3 Communicate honestly and effectively, taking into account the reliance of others on engineering expertise

4. Promote sustainability
   4.1 Engage responsibly with the community and other stakeholders
   4.2 Practise engineering to foster the health, safety and wellbeing of the community and the environment
   4.3 Balance the needs of the present with the needs of future generations

https://www.engineersaustralia.org.au/ethics

In the United States, the Code of Ethics of the National Society of Professional Engineers is stated in terms of Fundamental Cannons, Rules of Practice and Professional Obligations.

In other countries, national interests can also be seen within the specific codes. The Japan Society of Civil Engineers Code of Professional Conduct (2014) has a requirement to *ensure the security of society and to mitigate disasters.* While many of the national codes have some similar content, their variation demonstrates the need for a professional engineer to appreciate the expected conduct within the given operating environment and culture.

10.2 ETHICS AND MORAL PHILOSOPHY

Ethics deals with the interpretation of moral values, laws and principles which sway the decisions and behaviour of individuals and groups. It is part of the field of moral philosophy, which attempts to distinguish between that which is right and that which is wrong. A brief overview of moral philosophy provides a useful basis for the discussion of the ethical questions that arise in engineering.

Questions of morality include (Newall, 2005):

- What is right and what is wrong?
- What are rights? Who or what has them?
- Is there an ethical system that applies to everyone?
- What are the differences between community interests and self-interest?
- What do we mean by honour, integrity and dignity?

Statement of Ethical Principles
(Engineering Council, United Kingdom, 2014)

The Engineering Council, together with the Royal Academy of Engineering, in the United Kingdom provides the following four fundamental principles, to supplement codes of conduct by individual engineering institutions.

Accuracy and rigour - Professional engineers and technicians have a duty to ensure that they acquire and use wisely and faithfully the knowledge that is relevant to the engineering skills needed in their work in the service of others.

Honesty and integrity - Professional engineers and technicians should adopt the highest standards of professional conduct, openness, fairness and honesty.

Respect for life, law and the public good - Professional engineers and technicians should give due weight to all relevant law, facts and published guidance, and the wider public interest.

Responsible leadership: listening and informing - Professional engineers and technicians should aspire to high standards of leadership in the exploitation and management of technology. They hold a privileged and trusted position in society, and are expected to demonstrate that they are seeking to serve the wider society and to be sensitive to public concerns.

Issues of moral reasoning are investigated in the field of moral philosophy, as well as the meaning of the terms used in moral discourse. Moral philosophy deals with both analytic ethics (or metaethics) and normative ethics. In the field of analytic ethics, the questions that are discussed include whether moral judgements are possible, whether moral values exist objectively or only subjectively, and whether they are linked to cultures or individuals. Normative ethics deal with the ethics of specific actions. In so far as we are concerned here with ethics in engineering: “The purpose of analytic and normative ethics is to enable us to arrive at a critical, reflective morality of our own choosing” (Taylor, 1972).

An awareness of alternatives is essential when deliberating any decision, as this is what enables us to express a freedom of choice. The principle of freedom is essential to a theory of ethics. People cannot be responsible for their actions if there is no freedom of choice. Any action will determine a consequence which partly results from the set of laws and partly from the principles governing the decision.

Code of Ethics for Engineers
(United States National Society of Professional Engineers, 2007)

The National Society of Professional Engineers in the United States requires members of the profession to exhibit the highest standards of honesty and integrity. The Code of Ethics includes fundamental canons, rules of practice and professional obligations.

Fundamental Canons
Engineers, in the fulfillment of their professional duties, shall

- Hold paramount the safety, health, and welfare of the public.
- Perform services only in areas of their competence.
- Issue public statements only in an objective and truthful manner.
- Act for each employer or client as faithful agents or trustees.
- Avoid deceptive acts.
- Conduct themselves honorably, responsibly, ethically, and lawfully so as to enhance the honor, reputation, and usefulness of the profession.

There is a relativity in ethics which explains that conduct called good or bad and acceptable or unacceptable now has varied across time and among societies. Therefore, some people call things "right" that many others have a problem with. The Dalai Lama stated in a radio interview, "Of course, there are different truths on different levels. Things are true relative to other things; "long" and "short" relate to each other, "high" and "low," and so on. But is there any absolute truth? Something self-sufficient, independently true in itself? I don't think so." (Thurman, 1997).

Much of the moral reasoning of ethics in western countries in the past has been influenced by the Greek philosophers and, in particular, Socrates, Plato and Aristotle. Other major ethicists and philosophers through the ages have included instigators of various religions: Buddha (B.C. 483-563) the Indian philosopher and spiritual leader spreading Buddhism through Asia**;** Lao-zi, or Lao Tze*,* the founder of Taoism and the supposed author of Tao Te Ching; Confucius (Kong Fu-zi), whose philosophy and religion came to dominate China for more than two millennia; the Hebrew Bible with Judaism; Christianity based on the teachings of Jesus; and Mohammed with the Koran and Islam.

A myriad of writings is available to provide a basis for the study of ethics. A person thinks through a problem and then makes ethical decisions could be considered an individualist, even when the reasoning leads to society's norms. This is because the person has reasoned the decision and not just acted an existing value system.

Ullmann (1983) condemns people, including some engineers, who act as the implementers of others' goals without questioning them, when these goals might be to design weapon systems of mass destruction, or gas chambers and crematoria of concentration camps. The supporting argument from the condemned parties could be "that I was only obeying orders or doing my job". It is necessary as an engineer to collect information for any decision and weigh up the consequences of the outcomes resulting from that decision.

Socrates (470–399 BC)

Socrates was a philosopher who took his own life by drinking hemlock after being found guilty by an Athenian people's court of interfering with the religion of the city. Socrates "was accused of corrupting the youth" because he taught his students to question thinking. In his method, questions were posed to assist a person or group to examine their beliefs and their knowledge. The Socratic method uses hypothesis elimination, by eliminating those hypotheses that lead to contradictions and hence leaving the better hypotheses. This method forced people to examine their own beliefs as well as those of authorities. Socrates seemed to believe that wrongdoing resulted from ignorance, in that those who did wrong knew no better.

He did not recant and was sentenced to death. He could have escaped through the help of his friends but refused. In fact, Socrates said, "The life untested by criticism is not worth living."

Ullmann (1986)

10.3 ETHICS AND WHISTLE BLOWING

Historically, codes of ethics have sometimes been used to stifle dissent within a profession (Johnston et al, 1995). This is because codes of ethical behaviour, in general, emphasize refraining from criticizing fellow professionals and maintaining the profession's reputation. Situations where major disasters have occurred have led to the perception that there have been a lot more "whistle blowers" in recent years in engineering. Perhaps as technology has become more complex and environmental concerns have increased, there is a greater concern about ethics in decision making. To examine a problem from many angles is a useful way to develop an ethical approach, as the choice to take a certain action will only exist if we are aware of alternatives.

When does the dilemma of obligation to society override the obligation to the company or to peers? In some engineering situations, a conflict of obligations can arise. How does one address the obligations that a professional engineer has? Is there a natural hierarchy of obligations? Generally, it is acknowledged that an obligation to society takes precedence over obligation to the employer. What are the responsibilities, of both a personal and professional nature, which complicate ethical decisions? The term "blowing-the-whistle" refers to disclosing information outside of standard reporting lines, when a wrong has been identified and it is not felt that standard reporting will be sufficient to address the concern. Information is often disclosed to prevent or stop continuation of certain activities. When it comes to "blowing-the-whistle", there are numerous concerns, not the least being career prospects and family life.

The story of the Russian mining engineer, Peter Palchinsky, may serve as an inspiration for engineers today to follow moral principles, perhaps with less drastic consequences. Graham (1996) in his book, *The Ghost of the Executed Engineer – Technology and the Fall of the Russian Empire*, relates how Peter Palchinsky (1875–1929) spoke up against the misuse of resources. The misuse of technology and squandering of human resources by the Russian government incensed Palchinsky and he attempted to make known the facts of the misuse. He was imprisoned, but was released shortly thereafter by the efforts of his wife and friends. Palchinsky continued his professional engineering career and went on to criticize several of Stalin's projects: the White Sea Canal; Dnieper River Dam; and the Steel City of Magnitogorsk. His days were numbered, and he was executed in May 1929.

It was shown by Graham that there are different consequences of voicing what the individual believes to be morally correct ideas or demonstrating an obligation to society, from those of politically correct ideas depending on the political environment at the time. The book also shows the dangers of major projects which ignore human values.

The whistle blowing exploits of engineer Roger Boisjoly, concerning the Challenger disaster of January 1986, are still important for study on ethical decision making. The details about how the launch went forward in unusually cold temperatures against the recommendation of engineers are well known. Examining the events in ethical terms leads us to a better understanding of “whistle blowing”. The term portrays the sharp sound of a whistle, warning of harm or trying to halt actions which are about to go outside of acceptable bounds. Roger Boisjoly's action did not fit that scenario. What branded him as a whistleblower was his decision to relate to a presidential commission *after* the disaster the story of events that led to the disaster. The tragedy he had warned about had already occurred. He told the commission the history of the O-ring problems, and the decision to launch, even though he was being pressured by managers of his own company not to do so. Boisjoly's account was forward looking in view of the design of newer shuttles. It was really the high visibility of Boisjoly's disclosure about the failure that labelled his report as “whistle blowing”.

When a discloser deliberately steps outside of approved organizational channels to reveal a significant problem, this could be termed “whistle blowing”. In this case, the warning against launching in the meeting before the disastrous launch was acting within approved channels. Can it be said that all those decision makers above Boisjoly acted ethically? His testimony to the presidential commission did not contravene the approved channels, but rather, his testimony was honest and demonstrated the obligation he had to society.

Non-ethical behaviour seems to be driven by the advance of industry and the desire for success, power and money. Over time the need to choose between right and wrong, survival of a family or company, use of the environment now or preserving it for future generations is becoming more important for the individual in daily life. Ethical decisions reflect the conscious decision to do what is the right thing.

Generally, people use ethics to impose constraints on one’s own behaviour, whereas law could be described as a system of externally imposed or legislated constraints on the behaviour of the individual and society. The question of "what to do?" arises when there is conflict between ethics and law.

10.4 THE LEGAL SYSTEM

Roman law was developed in the City of Rome and in the Roman Empire as it expanded across Europe and beyond. In the Middle Ages, following the demise of Roman rule over Europe, Roman law was largely discarded. However, it continued to influence legal thinking and legal practice and through the ages has been revived, transformed and reinvented many times. For the first two centuries of the Common Era, Roman legal science was very fertile. This age is considered to be the classical period of Roman Law, as the law which was taught and practised best exemplified the Roman legal tradition (Rufner, 2006). From the 10th Century on, there was a renewed interest Roman Law. It was studied and taught at universities in Europe and this led to it being applied in the area of civil law throughout most of Europe, except in England. However, the Roman-based law in Europe from the 16th Century was very different from the Roman law of antiquity and it was called *Ius Commune* (common law). The modern European legal systems are derived from the *Ius Commune* as it was interpreted and rewritten by medieval jurists.

National legal codes, developed in the 18th and 19th Centuries, superseded the *Ius Commune*. The most important of these was the Code of Napoleon in 1804 in France, which became the basis for the rule of law in Belgium, the Netherlands, Portugal, Spain, and their former colonies. However, in Germany, Roman law was maintained in most of the regions until 1900 when the German Civil Code was introduced. The Roman-Dutch law based on *Ius Commune* still remains the basis of the legal system in the Republic of South Africa. The influence on the English legal system from the teaching of Roman law at Oxford and Cambridge amounted to ways of reasoning and concepts developed by continental legal scientists.

The English legal system has been followed in Australia, New Zealand and other former member countries of the former British Empire, as well as, to a large part, in the United States. It is a "common law" system which is based on the results of legal cases over a long period of time, whereas the codified civil law system on the European continent, as described above, emanated from Roman law. The term "common law" loosely describes the whole of the English legal system and has been used to distinguish between common law rules and equity, although the term in its broadest sense now embraces equitable principles.

There are two major sources for law, one being case law from judgements and the other from legislation through parliament. Both new and amended legislation and case law result in the law continually changing. In the building industry, engineers must be familiar with acts, codes of practice, ordinances and regulations which control the various facets of planning, design and construction of urban infrastructure. There are also some statutory modifications to common law, such as establishing the requirement of written contracts for sale of real estate.

Acts that will impinge upon an engineer's work are numerous and relate to all facets of the planning and development of infrastructure in a city and environs. For different countries and states these tend to be diverse and are continually changing under different governments.

In Australia, there are both Federal laws (covering everyone in the country) and State laws (covering people in the State or Territory which passes the law). A State or Territory Parliament can pass a law in an area where there is no Federal law, or it can pass a law that adopts what is said in a Federal law. If there is a clash between a Federal law and a State or Territory law, the Federal law overrides the State law. The Constitution of Australia outlines the laws the Federal Government can pass. Similarly, in the United States the Constitution defines the nature of laws which can be introduced within jurisdictions, whether at a federal or state level. Hence engineers will find themselves working under both Federal and State laws.

Most engineering and building works are subject to one or more of a range of Government Acts and Regulations, at both the State and Federal level. These may affect areas of planning development, design, construction, materials used, employment, environmental impacts, occupational health and safety, water supply and sewerage, water resources, natural resources, public health, energy and transport.

Different states have different Acts in Australia but some codes such as The Building Code of Australia (BCA), which is a set of technical provisions for the design and construction of buildings and other structures, are common to all states. The BCA is produced and maintained by the Australian Building Codes Board (ABCB) on behalf of the Australian Government and each State and Territory Government. Similar codes and Boards (or Councils) can be found throughout the world, including the International Code Council and the Canadian Code Centre.

Criminal law is the body of law which deals with crime and the legal punishment of criminal offences. It is that part of law that deals with matters of public interest. For example, the public has an interest in seeing that people are protected from being robbed or assaulted. Before the 18th and 19th Centuries, legal systems did not clearly define criminal and civil law. Codification of criminal laws progressed and separation of civil and criminal law occurred. Criminal law distinguishes crimes from civil wrongs such as tort or breach of contract, where there is a dispute between private individuals or organisations. The origin of tort comes from the French for “wrong” and is a civil wrong, other than a breach of contract. In engineering work, civil law is of prime interest, although criminal law can be relevant in the case of negligence. The following sections discuss aspects of the legal system most relevant to engineers.

10.5 LAW OF CONTRACT

The Law of Contract is of particular relevance as engineers provide services to customers, and procure services from others. Each of these transactions made externally to an organisation is likely to involve a contract.

A *contract* is an "agreement" or a "promise" that is recognized by the law. “The common law assumes that a contract is an agreement between parties of equal bargaining power” (Cooke, 2001). A contract may be simply defined as a voluntary agreement where one party agrees to do work for another for a stated payment, which the law can enforce. For a contract to be legally enforceable it must include the following elements:

- An offer must be made by one party to the other;
- Acceptance of the offer must be made by the second party;
- There must be consideration, such as a promise of payment; and
- There must be a mutual intention to create a legal relationship.

It is important to note that a contract will only be valid if the work to be done is legal and if both parties are legally competent. Therefore, no contracts for terrorism or illegal drug deals will hold up in a court of law.

Any contract that uses words, spoken or written, is labelled a verbal contract. An informal exchange of spoken promises can be binding and regarded as legally valid as a written contract, but will generally lead to more disputation if large sums of money are involved. A spoken contract is often called an "oral contract".

Courts in the United States have generally ruled that if the parties have the same intent and they act as though there were a formal, written and signed contract, then a contract exists. However, most jurisdictions require a signed written form for certain kinds of contracts such as real estate transactions and some building legislation (e.g. Home Building Act, NSW).

In Australia, there is no requirement for the entire contract to be in writing, although there must be evidence in the form of a note or memorandum of the contract, which must be signed. In England and Wales there is a “Statute of Frauds” legislation for guarantees which must be evidenced in writing, although the agreement may be made orally. Other kinds of contract (such as for the sale of land) must be in writing, otherwise they are void. If a binding agreement or an oral exchange is not honoured by one or more of the parties involved in the contract, by non-performance or interference with the other party's performance, then a “breach of contract” is deemed to have been committed.

Engineering contracts should be in writing to avoid ambiguities, and to simplify enforcement, if it is necessary. In the engineering and construction industries, the word *tender* is used when describing an offer by one of the parties. A contract for an engineering project may include drawings and specifications, which will include details of the full works required by the client/owner from the contractor. Contracts need to be specific to the agreement under consideration. Contracts do not need to be constructed by a lawyer, although the more complex the contract the more likely a lawyer will be involved. Engineering companies may have standard terms and conditions prepared by a lawyer and then these are used for all contracts that fall within pre-determined conditions. Alternatively, standard contracts conditions developed for common project types may be agreeable to both parties. Examples of such contracts include those based on the Australian Standard *AS 2124:1992 General conditions of contract* and *AS4300:1995 General conditions of contract for design and construct.*

Contract termination

A contract between parties is said to be discharged when the agreement is terminated. This can occur in several ways and usually for different reasons. Contracts can be terminated by agreement, when contract reaches completion, or because of a breach of contract, as described below:

- Agreement: The parties terminate the contract on mutual agreement by either party waiving their rights and/or releasing the other party from any obligations.
- Completion: The contract is discharged completely when it has been completely fulfilled in every respect by both parties.
- Breach: Whenever a party fails an obligation which was included in the contract, then a breach of contract has been committed.

When a breach occurs, action will of course be taken to determine the cause of the breach and/or how the contract can recover from it. Breaches can arise where a contract becomes seriously behind schedule. For a construction project, this is likely to follow many site meetings from the contract administrator to improve performance. It will also occur usually after on-site work has slowed down, or event stopped. In such cases, a possible cause will be that the contractor has financial problems. Clauses in the contract may allow the owner/client to terminate the contract. Also, the contractor may determine that the owner has not provided certain certificates or has refused to supply sufficient instructions on the work. In either case, if the contract is terminated without good cause, either party will be able to make a claim for damages for wrongful repudiation of the contract. In this kind of case there are bound to be lengthy delays and cost escalations on the contract, as well as large legal fees.

Within the contract process, a notice of intention to terminate by either party needs to be made under the procedures of the contract (Cooke, 2001). Considerations include:

- lapse of time: actions on a contract must be commenced within a certain time period decided by law; and
- impossibility of performance: this may arise when the contract works change so markedly from what was originally envisaged that a party can be exonerated from their obligations.

Most engineering contracts are terminated by completion or agreement. However, if termination occurs by breach of the contractor, then contract clauses usually permit the owner to retain all equipment and materials on site, completing the work at the contractor's expense. If the owner were in breach, then the contractor can take out an action against the owner. Failure to complete the contract within a specified time exposes the contractor to a claim of liquidated damages which could be included as a provision in the contract as a specific sum per day. Liquidated damages can be construed as a debt owed by the contactor to the owner and can be deducted from contract payments. Obviously, the owner cannot delay the works and then recover damages. All obligations on both parties must therefore be strictly observed if their rights under the contact are to be preserved. Contracts can run late due to (a) the lack of supply of certain materials, (b) long periods of rainfall not allowing the contractor to undertake work on site, and (c) disputes between subcontractors and the main contractor.

The engineer will, in many cases, be in the position of an arbitrator to settle minor variations on the original contract. If there cannot be an amicable solution between the contractor and the administrator (acting for the client), then all the parties could become involved in a dispute that leads to litigation. It is therefore essential for the engineer, or the administrator of the contract, to keep detailed and accurate records of all contract variations and actions by the contractor.

10.6 LAW OF TORT

William Prosser (1971) in his treatise, *Handbook of the Law of Torts*, defined "tort" as "a term applied to a miscellaneous and unconnected group of civil wrongs other than breach of contract for which a court of law will afford a remedy in the form of an action for damages." Besides damages, tort law will tolerate self-help in a limited range of cases. For example, using reasonable force to expel a trespasser. Furthermore, in the case of a continuing tort, or even where harm is merely threatened, the courts will sometimes grant an injunction to minimise the continuance or threat of harm. Some engineering disputes, where negligence or damages are involved, are decided under the law of tort.

Negligence is the broadest of the torts, forming the basis of many personal injury cases. The engineer has to be aware that there is a minimum *standard of care* that must be achieved in everyday activities. The engineer owes a *duty of care* to the client and *breaches* that duty by doing or not doing certain actions which are considered to be negligent. A breach of the duty of care can also occur if certain actions are not undertaken. Such a breach must lead to loss or damage, and it is generally considered to be fair and reasonable for the negligent party to pay compensation to the plaintiff/claimant.

When claiming for damages in tort, due to deceit or negligence, the plaintiff has to justify to the court the amount that is being claimed (Cooke, 2001). “The object is to restore the plaintiff to the position in which he would have been placed if the wrongful act had not been committed” (South Australia v Johnson (1982) 42 ALR 161 at 169–170).

In a case in South Australia, in which excessive cracking occurred in a house and was caused by negligent footing design, the court found that 50% was the responsibility of the builder, 25% the local authority and 25% the consulting engineer who advised the builder. The damages awarded in this case were the cost of demolishing and rebuilding the house. The judgement was appealed and overturned at the Supreme Court of South Australia but the High Court of Australia reinstated the original decision. Such cases have resulted in a strict set of guidelines for footing design in South Australia. These judgements have also set a precedent for other cases involving damages for defective building work (Thomas, 2004).

10.7 THE PROCESS OF DISPUTES

Disputes in the engineering construction industry are of many different types. Historically, most disputes are related to defective design and poor workmanship. As projects have become more complex, the nature of disputes has involved more parties from different disciplines. Traditionally, the mechanism for resolving construction disputes has followed a “tiered” process involving:

- negotiation;
- facilitated negotiation or mediation;
- assisted negotiation or conciliation;
- expert determination or neutral evaluation;
- arbitration; and
- litigation.

As the dispute escalates through the stages, then so do the time and resources required for resolving the dispute.

Negotiation is a relatively inexpensive approach to dispute resolution. It generally is an informal process where the parties involved want to achieve outcomes as expeditiously as possible. The parties negotiate between themselves until a resolution is reached.

Mediation or *conciliation* can be seen as the next most expensive approach and can offer distinct advantages, in that the duration of the process can be pre-determined with the parties involved deciding the rules under which they and the mediator will operate, and the informality of proceedings may make the necessity for legal representation less likely. The benefits are a resolution achieved more expeditiously and economically.

Expert determination is also favoured as a resolution process, particularly for disputes relating to a single issue of a technical nature. Compared with litigation or arbitration, expert determination is generally faster and less expensive. In the case of complex technical disputes it is arguable that a Tribunal, which does not have sufficient (or any) technical literacy, can appropriately handle such disputes. In the usual contract situation, the Superintendent, most likely an engineer, is the administrator of the Contract between the Principal (client/owner) and Contractor, and gives decisions on disputes. Increasingly the Superintendent has been viewed as being potentially compromised in this role with disputing parties, particularly those aided by legal advisors.

Arbitration previously offered a number of advantages over litigation. These entailed: privacy, less rigid procedures and formality; a shorter hearing time; lower costs; confidence in using a technical arbitrator; and a significant degree of finality, since a sound award is generally difficult to overturn. Arbitration involves each party presenting their points of view to an independent arbitrator, who then makes a decision. In the present climate in Australia, arbitration has become burdened with procedures similar to litigation, involving legal representation and procedural formality. The key benefits of time taken to bring a dispute to conclusion, and the cost, have escalated within these processes.

Litigation is both costly and time consuming and involves formal court processes. Disputes that arise during a project and go to arbitration or litigation are now unlikely to be resolved until after construction is complete, because of the lengthy procedures.

Because arbitration and litigation have become so costly, alternative methods of resolution are being called for in the construction industry. In the meantime, conciliation or mediation is being used as an intermediate step, before embarking upon the more costly and formal arbitration procedures.

Engineers can find themselves in arbitration, civil court actions, or even criminal court actions as an expert witness. An expert witness is usually asked by one or other of the parties contesting the case to act as independent engineer to investigate a failure or inadequate design or just advise on some dispute where an expert engineering opinion is required. However, an expert witness's general duty is to assist the court impartially on matters relevant to the expert's area of expertise. The expert has a duty only to the court and not to the party retaining the expert (i.e., the client). Accordingly, an expert witness is not an advocate for the client. The expert witness is required to follow the appropriate Court Guidelines for Expert Witnesses in fulfilling their duty. After collecting evidence, analysing the data and preparing a report, the expert will produce an expert report. It is common for both the plaintiff and the defendant to call their own expert witnesses and they present differing expert reports. In this case, the court needs to decide the facts of the case.

It is appropriate for lawyers and experts to collaborate on the form of an expert report, but not acceptable for litigants or their lawyers to influence the content. There is a need to take care and record all communications which are undertaken as an expert witness with the litigant, and the instructing lawyers. Generally, such communications are protected from disclosure until the expert report is served on the other party.

In essence, the expert, because of his or her independence and expertise, is in the privileged position of expressing opinions in evidence, in contrast to normal witnesses, who deal in facts.

Quite often, when clients receive an impartial report on a matter, they are inclined to withdraw the claim. It is likely also that when the prepared report has been viewed by the judge this can lead to an order to negotiate settlement outside the court.

10.8 LEGAL RESPONSIBILITY IN MANAGING STAFF

In many countries in the world, employers are responsible for ensuring that staff within their organisations have a healthy and safe place in which to work and do not suffer from discrimination. This is applicable to all occupations and the implications need to be understood by both the manager, and the staff being managed. National laws usually apply to Equal Opportunity, Discrimination, Human Rights and Work Health and Safety.

Equal opportunity and human rights

In many countries, including the United States, the member states of the European Union, the United Kingdom and Australia, there are national laws for equal employment opportunity and human rights. In the US, this is administered by the Equal Employment Opportunity Commission. In the UK, the national legislation requires organisations to have a written equal opportunities policy with procedures covering equal opportunities in recruitment, promotion, transfer, training, dismissal and redundancy. In the US, the Federal Laws are usually supplemented by State Laws. The Australian Constitution does not give any power to the Federal Government to make laws in the area of discrimination. However, once the Commonwealth of Australia signs a treaty with an international organization (such as the United Nations or associated bodies like the International Labour Organization), the Federal Government has an international duty to draw up laws based on the principles in the treaty. Federal laws in Australia based on principles of human rights include:

- The Racial Discrimination Act 1975;
- The Sex Discrimination Act 1984;
- The Australian Human Rights Commission Act 1986 (formerly called the Human Rights and Equal Opportunity Commission Act 1986);
- The Privacy Act 1988;
- The Disability Discrimination Act 1992; and
- The Age Discrimination Act 2004.

Each state government has an Act to promote equal opportunity in organisations and communities. For example, in South Australia it is the Equal Opportunity Act (SA), 1984 which is an act to: "promote equality of opportunity between the citizens of this State; to prevent certain kinds of discrimination based on sex, sexuality, marital status, pregnancy, race, physical or intellectual impairment or age; to facilitate the participation of citizens in the economic and social life of the community; and to deal with other related matters."

Equal Opportunity is about a person's behaviour towards others, or how decisions are made in public areas of life. Put simply, it is about the respect to be shown to fellow workers.

Anti-discrimination or Equal Opportunity means that all people have a right to be treated fairly regardless of irrelevant personal characteristics. For example, it is unlawful under the Equal Opportunity Act (SA) 1984 for anyone to be treated unfairly based on age; sex; marital status; pregnancy; sexuality; physical or intellectual impairment; or race. However, these laws only apply in areas of public life (including employment), and private life is excluded.

The work areas covered by the Act include: accommodation; advertising; clubs and associations; conferral of qualifications; disposal of land; education; employment; and provision of goods and services. In the employment area, the law applies to paid full-time, part-time and casual workers, as well as unpaid or contract employees. It applies to all stages of employment: job advertisements and applications; interviews; promotions; training; and dismissal. In employment, Commonwealth laws also cover the following personal characteristics: religion; political opinion; medical record; irrelevant criminal record; social origin; and trade union activity. This means that some of the current State legislation would have to be read in conjunction with the Federal legislation.

Application of legislation to employment

Maria sought a promotion to Team Leader of the roadwork's team at a council depot. Although she was an excellent worker and had previous supervisory experience, the depot manager decided not to consider her application, believing that the all-male staff would not respect her. Maria complained to the Commissioner. A representative held discussions with the manager, and Maria was given an interview for the team leader job.

(Equal Opportunity Commission of South Australia, 2006)

Work health and safety

Work health and safety legislation aims to provide a healthy and safe work environment for all, whether they are employees, contractors or visitors. In Australia, a model Work Health and Safety Act (2012) was introduced for enactment by the individual states. In the Act, and accompanying regulations, all employers must ensure that each person employed or engaged by the employer is safe from injury and risks to health while at work by providing and maintaining a safe working environment, safe systems of work, plant and substances in a safe condition, adequate facilities for the welfare of employees; and must provide such information, training and supervision that are necessary to ensure that each employee is safe from injury and risks to health.

Worker fatalities involving a fall from height
(Safe Work Australia, 2013)

While worker health and safety has improved over time, fatalities and injuries still occur. Unfortunately, industries related to the engineering profession such as the construction industry are over-represented in the statistics.

Between the years 1989 and 1992, there were 214 fatalities (54 per year) recorded. Of those fatalities, 12% involved a fall from height. Worker fatalities fell to 112 deaths between 2008 and 2011 (28 per year). However, the percentage of deaths involving a fall from height remained similar during the period, at 11%.

It is necessary that all employees are aware of the procedures and processes involved in creating and maintaining healthy and safe work environments. Engineers are often responsible for staff, and they must ensure that breaches of the relevant legislation do not occur. Many organisation have management systems which are designed to provide processes and templates for use. These processes and templates ensure that consistent approaches to induction, training and risk management are undertaken. While these processes and templates do assist an organisation to demonstrate compliance with the associated legislation, their primary aim is to maintain a healthy and safe work environment.

For engineers, responsibilities within the Act extend beyond the standard workplace and include a responsibility for designs to integrate control measures early in the design process to eliminate or, if this is not reasonably practicable, minimise risks to health and safety throughout the life of the system under consideration. Health and safety issues are a key consideration in the planning and design of any engineering system. This is often referred to as Safe Design Practice, Safe by Design, Safety in Design or Prevention by Design.

Engineers need to develop a mindset which holds paramount the need for their work to be carried out safely, and that designs are safe for users. This relates to the entirety of any project and not just the part that they are directly involved with. This mindset extends from the problem formulation phase through detailed planning and design to implementation, and ultimately, decommissioning. The requirements of legislation of course need to be satisfied, but, more importantly, the safety mindset needs to be ingrained in employees through training and awareness programs.

10.9 SUMMARY

The issues raised in this chapter are of special importance to engineers who are responsible for safety procedures in the workplace. They are also relevant to any engineer who is making decisions that impinge upon the community, is involved in choosing new technology, or is applying existing technology. Engineers must take responsibility for their decisions and defend their actions, especially those actions and decisions that cannot be supported by scientific and technological argument. Engineering has found itself in the middle of some of the largest disasters that have occurred both in terms of social distress and provision of a solution. The questions of what is justified as a whistle blower or what is obligatory under the code of ethics need to be answered by taking into account all the ramifications of a moral dilemma for the decision maker.

PROBLEMS

10.1 "Relativity applies to physics, not ethics" (Albert Einstein). Discuss how this statement is important for ethical decisions a professional engineer might make.

10.2 An engineer learns by accident that his employer, a major chemical fertilizer company, has been disposing of toxic wastes in an area where the frequency of cancer and birth defects is on the rise. He also learns that a TV documentary involving local citizens will be aired next week and these citizens are preparing a legal suit against the company. He immediately sells all his shares in the company and following advice that they have been sold, then tells his CEO what he has learned. The legal suit is filed two days later. Has he done anything unethical? Please discuss. (Based on an example in Ullmann, 1986.)

10.3. How much should you change your behaviour to fit with the beliefs, values and norms of those with whom you are interacting with? Whose responsibility is it to change: the visitor; sojourner; newcomer; or the host? What about the language one speaks? Should a person who comes to Australia be required to speak English?
Issues: At what point do you give up your cultural identity and moral integrity? At what point does adoption of another culture offend or insult? It is necessary to be aware of one's own moral stance and at the same time the need to display respect for others. Discuss with someone and write a paragraph on your outcomes.
(adapted from http://www2.andrews.edu/~tidwell/bsad560/Ethics.html)

10.4 Compare the Engineering Code of Ethics applicable to the country where you are studying with two other countries. If you were to be employed in one of these two countries would be there be a need to change your professional conduct? Would there be any implications if you were employed in one of the countries, but remotely undertaking your engineering work in another?

10.5 Manager's dilemma (adapted from the Equal Opportunity Commission for South Australia, 2006).

An employee comes to you to complain about offensive behaviour she has been putting up with from another co-worker. She insists, however, that she does not want to make a formal complaint. What are some possible steps you might consider? Consider and debate the following ideas:

(a) Ask her why she insists on not proceeding further with her complaint;
(b) Respect her right not to proceed with a formal complaint and inform her of your duty of care to protect her and other workers;
(c) Monitor the situation and follow up with her to see if the problem gets any worse; and
(d) Consider recirculating relevant policies or codes of conduct to all employees.

10.6 There are many resources available on the web. Search your local Equal Opportunity website and one in another country and compare the differences in rulings on a number of different aspects, such as religion, age and racial discrimination.

10.7 Identify and discuss a legal case where a prosecution has occurred for a design which was not considered safe. What lessons could be learned for this case?

10.8 A professional engineer is walking past a building site (in which he or she has no professional involvement) and notice that the scaffolding is quite unsafe. Discuss whether he or she has a legal or an ethical obligation to report this? If so, what action should he or she take?

REFERENCES

Beaton, G., 2010, Why professionalism is still relevant. http://www.professions.com.au/about-us/what-is-a-professional.

Cooke, J.R., 2001, *Architects, Engineers and the Law*, Federation Press, 3rd Edition.

Dym, C.L. and Little, P., 2000, *Engineering Design. A Project-Based Introduction*. John Wiley & Sons.

Engineering Council UK, 2014, Statement of Ethical Practice, 4pp. http://www.engc.org.uk/engcdocuments/internet/Website/StatementofEthical Principles.pdf.

Engineers Australia, 2010, Our Code of Ethics, 3pp. http://www.engineersaustralia.org.au/sites/default/files/shado/About Us/Overview/Governance/codeofethics2010.pdf.

Equal Opportunity Commission of South Australia, 2006, Understanding Equal Opportunity, http://www.eoc.sa.gov.au/default_multitext.jsp?xcid=170.

Graham, L.R., 1996, The Ghost of the Executed Engineer – Technology and the Fall of the Russian Empire, Harvard University Press.

Halsall, P.M., 1998, *Ancient History Source Book, Code of Hammurabi* (c.1750 BCE). http://www.fordham.edu/halsall/ancient/hamcode.html#horne

Japan Society of Civil Engineers, 2014, Code of Professional Conduct, http://www.jsce-int.org/about/p_engineer.

Johnston, S., Gostelow, P., Jones, E., Fourikis, R., 1995, *Engineering and Society. An Australian Perspective*, Harper.

Kelly, G.D., 2013, Construction Law: An Introduction for Engineers, Architects and Contractors. John Wiley & Sons.

Morgan, Roy, 2016, *Roy Morgan Image of Professions Survey 2016: Nurses still easily most highly regarded – followed by Doctors, Pharmacists & Engineers.* Finding No. 6797. http://www.roymorgan.com/findings/6797-image-of-professions-2016-201605110031.

National society of Professional Engineers, 2007, Code of Ethics for Engineers, http://www.nspe.org/resources/ethics/code-ehtics.

Newall, P., 2005, *Ethics*, Galilean Library, www.galilean-library.org/int11.html.

Oberg, J., 2000, Houston, We Have A Problem. *New Scientist*, 165(2234), 26–29.

Prosser, W., 1971, *Handbook of the Law of Torts*. 4th ed. St. Paul: West Publishing Co.

Rufner, T., 2006, *Roman Law*, http://www.jura.uni-sb.de/Rechtsgeschichte/Ius.Romanum/english.html.

Safe Work Australia, 2013, Work-related injuries and fatalities involving a fall from height, Australia, 57pp.

South Australia v Johnson, 1982, 42 ALR 161 at 169–170.

South Australian Government, 2005, *Occupational Health, Safety and Welfare Act* 1986, amended 2005. http://www.parliament.sa.gov.au/Catalog/legislation/Acts/o/1986.125.un.htm.

Taylor, P. W., 1972, *Problems of Moral Philosophy*, 2nd ed., Dickenson Publishing.

The Illinois Institute of Technology, 2016, The Center for the Study of Ethics in the Professions, Ethics Code Collection, http://ethics.iit.edu/ecodes/ethics-area/10.

Thomas, B., 2004, The assessment of damages for breach of contract for defective building work, *The Building and Construction Law Journal*, August, 20(4), 230–256.

Thurman R., 1997, The Dalai Lama interview with Robert Thurman, 1997, http://www.motherjones.com/news/qa/1997/11/thurman.html.

Ullmann, J.E. (ed), 1986, *Handbook of Engineering Management*, John Wiley & Sons.

Ullmann, J.E., 1983, The responsibility of Engineers to their Employers in J.H Schaub and Karl Pavlovic, *Engineering Professionalism and Ethics*, John Wiley & Sons.

CHAPTER ELEVEN

Risk and Reliability

In this chapter we consider the concepts of risk and reliability and how they can be considered in engineering planning and design. The concepts of risk and reliability are defined and illustrated in relation to engineering problems. Reliability-based design and the selection of safety coefficients are explained in detail. The concepts of resilience, vulnerability and robustness are also introduced. Methods for evaluating the reliability of engineering components and engineering systems are explained.

11.1 INTRODUCTION

Risk is inherent in all human activity. For example, we risk personal accident, injury and even death whenever we travel in a car, go for a jog, or cross the street. The levels of such everyday risks depend, in part, on the way individuals choose to conduct their lives. Generally, such risk levels are very low in modern society, unlike the risks faced daily by our ancestors in pre-historic times. The probabilities associated with the risks associated with various everyday activities have been estimated in a number of studies (US Nuclear Regulatory Commission, 1975; Stone, 1988; Morgan, 1990; BC Hydro, 1993). Some of these probabilities are given in Table 11.1.

Risk is also inherent in all engineering work, as was explained in Chapter 3, and part of the role of the engineer is to manage risk levels and to keep them to the very low levels that are acceptable to society.

11.2 LEVELS OF RISK

Dougherty and Fragola (1988) define *risk* as the combination of the probability of an abnormal event or failure and the consequences of that event or failure to a project's success or a system's performance. This combination of probability (or likelihood) and consequences is used in tables of risk ratings such as Table 11.2.

In Table 11.2 catastrophic consequences would involve loss of life, major consequences include serious injury or major economic loss, moderate consequences include minor injuries or moderate economic loss, and minor consequences include minor economic loss. Efforts to minimise risk should give priority to the activities that have an extreme or high risk rating.

Table 11.1 Annual probability of death to an individual.

Source of risk	Annual probability of death	Source of information
Car Travel	1 in 3,500	Morgan (1990)
	1 in 4,000	US Nuclear Regulatory Commission (1975)
	1 in 10,000	Stone (1988)
Air Travel	1 in 9,000	Morgan (1990)
	1 in 100,000	US Nuclear Regulatory Commission (1975)
Drowning	1 in 30,000	US Nuclear Regulatory Commission (1975)
Drowning (UK average)	1 in 100,000	Morgan (1990)
Fire (UK average)	1 in 50,000	Morgan (1990)
Household Electrocution (Canada)	1 in 65,000	Morgan (1990)
Electrocution	1 in 160,000	US Nuclear Regulatory Commission (1975)
Fire (U.K. average)	1 in 50,000	Morgan (1990)
Lightning	1 in 2,000,000	US Nuclear Regulatory Commission (1975)
	1 in 5,000,000	Morgan (1990)
	1 in 10,000,000	Stone (1988)
Nuclear Reactor Accidents	1 in 5,000,000	US Nuclear Regulatory Commission (1975)
Structural Failure	1 in 10,000,000	Morgan (1990)

Table 11.2 Risk ratings. (Reproduced with the kind permission of Neill Buck and Associates)

Likelihood	Consequences				
	Insignificant	Minor	Moderate	Major	Catastrophic
Almost Certain	Moderate	High	Extreme	Extreme	Extreme
Likely	Low	Moderate	High	Extreme	Extreme
Possible	Low	Moderate	Moderate	High	Extreme
Unlikely	Insignificant	Low	Moderate	Moderate	High
Rare	Insignificant	Insignificant	Low	Low	Moderate

For example, the risk levels that are applied in the planning, design and construction of a city office building vary widely. The structural design objectives are to provide a system which will have adequate safety against collapse, as well as good serviceability under normal working loads. These objectives are quantified using minimum performance levels. In regard to serviceability, a maximum allowable deflection is prescribed for the floors and is not to be exceeded when the design working loads are applied. In regard to safety against collapse, minimum levels of overload due to live load, wind and earthquake are prescribed that the structural system must be able to withstand. It will be clear that the consequences of exceeding the deflection criterion will be at most minor, whereas not meeting the minimum strength requirements could well be catastrophic. For this reason the acceptable probability of excessive deflection is much higher than the acceptable probability of collapse.

Risk of gastroenteritis among young children who consume rainwater in South Australia

Eighty-two percent of rural households in South Australia have rainwater tanks as their main source of drinking water. There are some risks associated with the consumption of rainwater as it may be contaminated by pathogens from birds or animals. A study was carried out to assess whether there was a significant difference in the risk of gastroenteritis among 4- to 6-year old children who drank rainwater compared to those who drank treated mains water.

Over 1000 children were included in the study which was carried out over a six-week period. Parents of the children were asked to keep a diary of gastrointestinal symptoms. The occurrence of gastroenteritis was based on a set of symptoms (defined as a highly credible gastrointestinal illness or HCGI). It should be noted that most incidences of gastroenteritis were mild.

Based on the raw data, the incidence of HCGI was 4.7 episodes per child-year for those who only drank rainwater compared to 7.3 episodes per child-year for those who only drank mains water. However, when adjustments were made for other risk factors, there was no significant difference in the incidence of HCGI between the two groups. The study concluded that consumption of rainwater did not increase the risk of gastroenteritis compared with mains water for 4- to 6-year old children. One possible reason for this conclusion is the acquired immunity by the children who drank rainwater, as they had all done so for at least a year prior to the study and could have developed immunity to a number of microorganisms during this period.

Source: Heyworth et al., 2006

11.3 RELIABILITY BASED DESIGN

The *reliability* of an engineering component or system is defined as the probability that the system will be in a non-failure state. The possible modes of failure need to be defined for each particular system under consideration. For example, for a water supply system, one failure mode would correspond to running out of water, although a more likely mode would occur with the imposition of extreme water restrictions on household consumers and industry. Yet another mode would be the provision of water of an unacceptable quality to consumers. For a freeway system, failure could correspond to extreme congestion associated with low vehicle speeds and long delays.

In this section, a simplified example of reliability-based design will be explained using a simple structural element in tension. The same principles apply to many other types of engineering systems.

In structural engineering, probabilities of failure are rarely calculated for design calculations. The usual practice is apply overload factors to the load quantities and understrength terms to the resistance quantities. However, the probability of failure is taken into account in evaluating these “safety factors”. This approach will be demonstrated for the design of simple tension members. Some examples of these are shown in Figure 11.1.

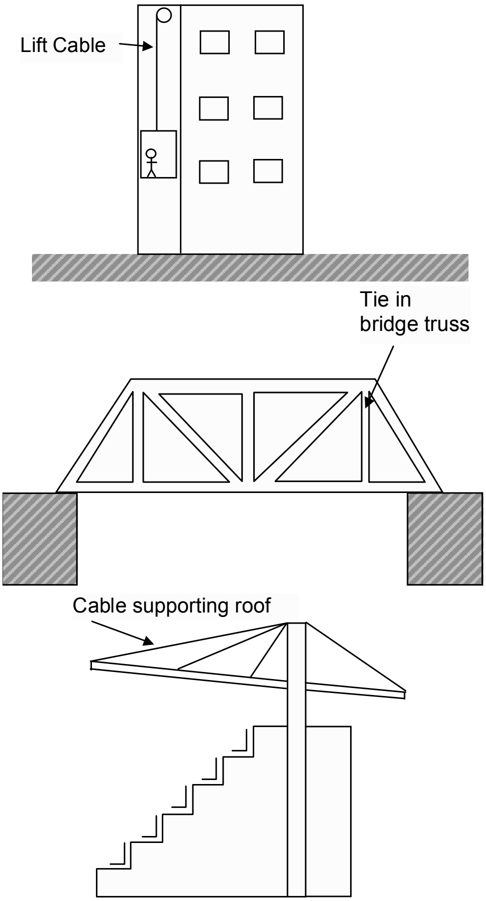


Figure 11.1 Examples of tension members (Lift cable, tie in a bridge truss, and cable supporting a roof).

Consider the analysis of a tension member (or tie) of cross-sectional area, A, as shown in Figure 11.1. If the maximum force that will be applied to the member during its lifetime, S, is known, the maximum stress in the member, σ, can be calculated using the equation:

$$\sigma = \frac{S}{A} \tag{11.1}$$

where A = the cross-sectional area of the member. If σ is less than the stress that will cause failure of the member, σ_f, the member will not fail. In the design situation, the failure stress of the steel and the maximum applied force are assumed to be known, hence the required cross-sectional area of the member, A, can be determined by rearranging Equation 11.1 to give:

$$A \geq \frac{S}{\sigma_f} \tag{11.2}$$

However, this approach does not take into account uncertainty in S, σ_f, and A. Clearly the maximum force that will be applied to the member during its lifetime is not known with certainty. There can also be variations in the failure stress of the steel that comprises the member due to variations in the manufacturing process. Finally, the actual cross-sectional area of the member may vary from the value specified in the design due to variations in production of the member.

One approach that provides some margin of safety in the design is to use a *safety factor*. If we define T, the resistance of the member, as:

$$T = \sigma_f A \tag{11.3}$$

We require T to exceed S by some margin to allow for the possibilities of overloads and understrength. We define the safety factor, γ, as follows:

$$\gamma = \frac{T}{S} \tag{11.4}$$

where γ is typically in the range 1.5–3.0 depending on the loading conditions, material and likely mode of failure. The cross-sectional area of the member is then determined by combining Equations (11.3) and (11.4) as follows:

$$A \geq \frac{S\gamma}{\sigma_f} \tag{11.5}$$

The single safety factor approach is no longer commonly used in structural engineering design, but is still used in geotechnical engineering. Structural engineers now use a reliability-based approach to structural design. This approach explicitly recognises the uncertainties in the various factors involved in the design. In the reliability approach, we define the difference between T and S as the *safety margin*, Z.

$$\text{i.e.: } Z = T - S \tag{11.6}$$

We expect variations in T and S due to some of the factors described previously. We will assume that T and S are normally distributed random variables that are independent of each other. Typical distributions of these variables are shown in Figure 11.2. This figure shows some small but finite probability that either T or S is negative. Clearly negative values of T or S are physically impossible, as the normal distribution is only an approximation to the true distributions of the resistance T and maximum applied force S.

Let the means of T and S be designated by $\mu(T)$ and $\mu(S)$ (respectively) and the standard deviations of T and S be designated by $\sigma(T)$ and $\sigma(S)$ (respectively). Then, if both T and S have normal distributions, so will Z. The mean and standard deviation of Z can be determined using Equations (11.7) and (11.8).

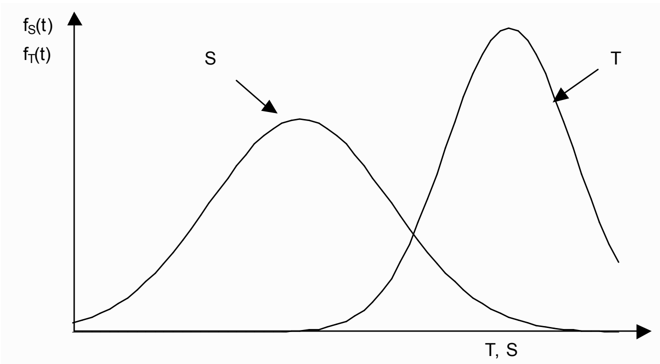


Figure 11.2: Probability distributions of resistance *T* and applied force *S*.

$$\mu(Z) = \mu(T) - \mu(S) \tag{11.7}$$

$$\sigma(Z) = \sqrt{\sigma(T)^2 + \sigma(S)^2} \tag{11.8}$$

The probability distribution of Z is shown in Figure 11.3.

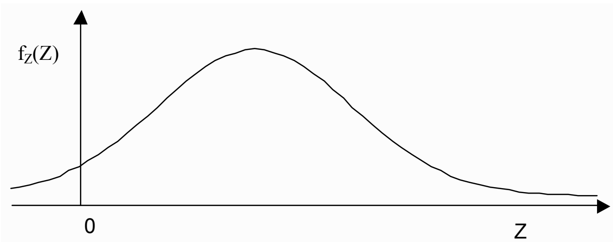


Figure 11.3 Probability distribution of the safety margin, *Z*.

If Z is negative, the member will fail, if it is positive, the member will not fail. The probability of failure, P_f, is given by Equation (11.9):

$$P_f = P[Z < 0] \tag{11.9}$$

As stated earlier, the reliability of a system, R is defined as the probability that the system will be in a non-failure state. Hence:

$$R = 1 - P_f = 1 - P[Z < 0] \tag{11.10}$$

A commonly used measure of safety is the safety index, β, which is defined as:

$$\beta = \frac{\mu(Z)}{\sigma(Z)} = \frac{1}{V(Z)} \tag{11.11}$$

where $V(Z)$ = the coefficient of variation of $Z = \sigma(Z)/\mu(Z)$. For the case where Z follows a normal distribution, β is related to the probability of failure as indicated in Table 11. 3.

Table 11.3 Relationship between safety index and the probability of failure for a normal distribution (adapted from Warner et al., 1998).

Safety Index β	2.32	3.09	3.72	4.27	4.75	5.20
Probability of failure, P_f	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}

As T and S are now both considered to be random variables, the safety factor defined in Equation (11.4) may be expressed in terms of the mean values of T and S.

i.e., $$\gamma_0 = \frac{\mu(T)}{\mu(S)} \tag{11.12}$$

where γ_0 is called the central safety factor. Warner et al. (1998) show that β is related to γ_0 by the following expression:

$$\beta = \frac{\gamma_0 - 1}{\sqrt{\gamma_0^2 V(T)^2 + V(S)^2}} \tag{11.13}$$

where $V(T)$ and $V(S)$ are the coefficients of variation of T and S (respectively).

Hence in order to maintain the same safety index (and hence probability of failure) γ_0 must vary depending the values of $V(T)$ and $V(S)$.

As noted previously, a major weakness of using γ_0 is that it does not take into account variations in T and S. This may be overcome to some extent by using more extreme values of T and S. For example, define S_k as the value of the applied load S that has a 5% chance of being exceeded. Furthermore, define T_k as the value of the resistance T that has a 95% chance of being exceeded. Then the “nominal” safety factor is given by the following equation:

$$\gamma = \frac{T_k}{S_k} \tag{11.14}$$

thus, we use a high value of the applied load and low value of the resistance in assessing the nominal safety factor.

Figure 11.4 shows the relationship between the distributions of T and S and the values of T_k and S_k. Modern structural codes of practice use partial safety coefficients to allow for the variabilities in T and S separately. The partial safety coefficients are applied to T_k and S_k to obtain design values T_d and S_d as follows:

$$T_d = \frac{T_k}{\gamma_T} \tag{11.15}$$

$$S_d = \gamma_S S_k \tag{11.16}$$

Thus a safe design requires:

$$T_d > S_d \tag{11.17}$$

$$T_k > \gamma_T \gamma_S S_k \tag{11.18}$$

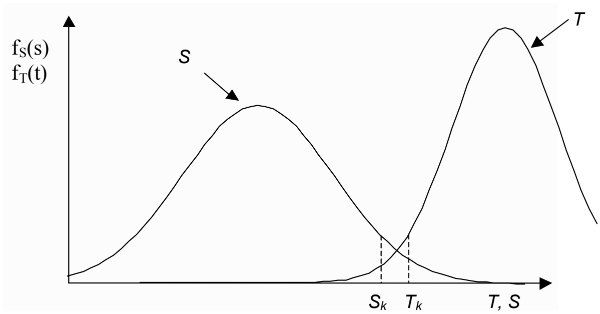


Figure 11.4: Relationship between *T* and *S* and the values of T_k and S_k.

Values of the partial safety coefficients, γ_T and $\gamma_{S,}$ are selected to ensure that that the probability of failure, P_f, given by Equation (11.9) lies within an acceptable range.

It should be noted that the approach outlined above is based on a number of simplifying assumptions such as that *T* and *S* are independent and have normal distributions. Furthermore, even in the simple case of a tensile member, the variability in *T* can be due to variations in materials properties as well as variations in the actual dimensions of the member compared to those assumed in the design. For this reason, the probability of failure determined is usually referred to as a nominal probability of failure.

A similar approach to that outlined above can be used for other structural elements such as beams, columns and frames or other engineering components such as electrical transformers, items of machinery or distillation columns.

Of course, the analysis of an entire system is more complicated. For example, a building structure may have many components, alternative load paths and inbuilt redundancies as well as many different modes of failure.

Furthermore, it should be noted that the above analysis assumes that structural failure will occur due to unusually high loads or low resistance of the structural member itself (due to lower than expected material properties or dimensions). In reality, it has been observed that most structural failures occur due to human error caused by factors such as:

- gross errors in the design of the structure;
- loads applied to the structure that were not anticipated in the design (e.g. an aircraft crashing into a building); or
- inappropriate construction methods.

This underlines the need for proper quality control in engineering design and construction including appropriate work practices, supervision and checking.

11.4 SELECTION OF SAFETY COEFFICIENTS

As outlined in Chapter 3, most engineering design work is guided by codes of practice that specify minimum levels of safety to ensure that the frequency of failure is within limits that are acceptable to the community. The setting of these levels of safety involves a balancing act between overly conservative design on the one hand and unconservative design on the other. If the levels of safety chosen lead to overly conservative design, money will be spent on achieving these high standards that could be better employed elsewhere in the economy (e.g. building new hospitals or purchasing additional medical equipment). If the levels of safety are too low, failures (and the consequent loss of life and/or economic loss) will occur more frequently than is acceptable to the community and there will be a public outcry.

Following Warner et al. (1998), it is appropriate to consider the choice of the optimum probability of failure for a particular class of structure (e.g. highway bridges). Clearly as the probability of failure, P_f increases the cost associated with the failure of all bridges also increases as shown in Figure 11.5. On the other hand, increasing the probability of failure is associated with reducing the margin of safety and hence the cost of constructing all bridges as shown in Figure 11.5. The total cost is the cost of failures plus the cost of construction of all highway bridges. The value of P_f that corresponds to the minimum total cost is the optimum probability of failure. In practice, this value may be quite difficult to identify and values of the partial safety coefficients, γ_T and γ_S are often specified in codes of practice based on experience with previous successful design procedures and comparison with a number of standard design cases.

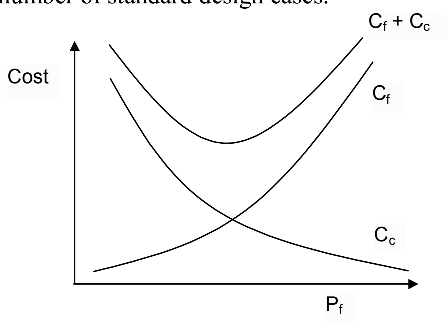


Figure 11.5 Tradeoff between cost and probability of failure, where C_c = cost of construction and C_f = cost of failures (adapted from Warner et al., 1998).

11.5 RELIABILITY, RESILIENCE, VULNERABILITY AND ROBUSTNESS

The **reliability** of an engineering system is defined in Section 11.3 as the probability that it is in a non-failure state at any point in time. Reliability is clearly an important performance indicator for most engineering systems. However, it does not give the total picture. Hashimoto et al. (1982) identified the following three performance measures of a water resource system: reliability, resiliency and vulnerability.

Hashimoto et al (1982) defined *resiliency* as the probability of a system returning from a failure state to a non-failure state in a single time step. This is equal to one divided by the average length of time that a system's performance remains unsatisfactory after a failure. Clearly this definition of resiliency does not apply if system failure corresponds to a catastrophic collapse (e.g., of a building or other structure) from which recovery is impossible. However, it is applicable to systems where failure corresponds to economic loss or inconvenience. For example, the imposition of water restrictions in a city, flooding of an urban area or severe traffic congestion that causes significant delays on a road network. In these cases, it is possible for the system to recover from the failure and perform in a satisfactory manner for an extended period thereafter.

Note that the terms "resilience" and "resiliency" tend to be used interchangeably in the literature. From this point onwards, we will use the term resilience throughout. A more general definition of *resilience* is the ability of a system to recover from large perturbations without changing its basic structure (Fiksel, 2003). This definition is commonly applied to ecological (Holling, 1973) or socio-economic systems (Resilience Alliance, 2010). This is related to the concept of *stability* for mechanical systems. A system is said to be in a stable state if, when perturbed, it returns to that state.

The consideration of resilience in the planning or design of an engineering system leads to a fundamentally different approach to that based purely on system reliability. Reliability-based design considers the system's performance during periods when it is subject to standard loading conditions and the required performance criteria are expected to be met. Hence, reliability-based design is aimed at avoiding failure or providing "fail-safe" performance. Conversely, resilience considers performance during periods when the required performance criteria are not met (i.e. the system is in a failure state). Hence the aim of resilience-based design is to recover from failure and to ensure that the system is "safe to fail" (Butler et al, 2016).

The structural design of buildings in areas with significant likelihood of the occurrence of earthquakes is an example of resilient design. Such structures are designed to withstand the loads imposed by a specific design earthquake with minimal damage. However, in recognition that it is possible that a larger earthquake could occur, the beams, columns and joints of the structure are designed to be ductile, so that, should failure occur, it will not result in a catastrophic collapse of the building. As shown in Figure 11.6, large deformations of the structure may occur, but, provided its structural integrity is maintained, there will be a good chance that the occupants will survive the earthquake.

Figure 11.6 A building in Kobe, Japan that failed in the 1995 earthquake but did not collapse due to ductility in its beams and columns (© M. C. Griffith).

Hashimoto et al. (1982) defined *vulnerability* as the average magnitude of failure of a system, given that it does fail. This is sometimes expressed as a percentage of the average demand or load on the system, so that it is dimensionless. It is desirable for a system to have high values of reliability and resilience but a low value of vulnerability. However, Hashimoto et al. (1982) showed that, for a water resource system there is usually a trade-off between the three measures, so that the system that has the lowest vulnerability is unlikely to have the highest reliability or the highest resilience and vice versa.

Another concept related to system performance is *robustness*. Robustness is a measure of the insensitivity of the performance of a given system or plan to uncertain future conditions (Maier et al., 2016). Hence a robust system or plan is one that performs well over a wide range of future conditions or scenarios. The concept is illustrated by the example given in the following interest box.

Reliability, Vulnerability and Robustness of Southern Adelaide's Water Supply System

Paton et al. (2014) carried out a multi-objective planning study of future water supply options for Southern Adelaide, south Austrialia. The population of the area is approximately 600,000 people. In 2009 the city's water supply was provided primarily from reservoirs in the nearby Mount Lofty Ranges and inter-basin transfers from the River Murray (70 km to the east of the city). Future water supply options considered were combinations of the following: a desalination plant (of various capacities), a number of stormwater harvesting schemes (for non-potable supply) and household rainwater tanks. The planning horizon was from 2010 to 2050. The objectives considered were to: (a) minimise the present value of capital and operating costs of the system; (b) minimise the total greenhouse gas emissions; and (c) minimise vulnerability of the system (maximum shortfall in supply). Each of these was evaluated for the full planning horizon.

The system reliability was constrained to be greater than or equal to 95% and the average maximum duration of failure (corresponding to the imposition of severe water restrictions) was constrained to be less than or equal to 365 days. The average maximum duration of failure is the inverse of the resilience of the system according to the definition of Hashimoto et al. (1982).

The optimisation was carried out for the most likely climate change scenario and a medium level of population growth. A post-optimality analysis of system robustness was carried out for selected plans that were identified in the multi-objective optimisation. The aim of this analysis was to consider how the various plans performed under changes in meteorological factors due to climate change and changes in demand due to different population projections. Each plan was evaluated under 252 possible scenarios. These were combinations of 7 possible global circulation models, 6 emissions scenarios and 6 projections of future population for the city. Robustness for each plan was defined as the percentage of the 252 scenarios under which the plan exhibited acceptable performance. Acceptable performance was defined as follows: (a) reliability greater than or equal to 95%; (b) maximum duration of failure less than or equal to 365 days; and (c) maximum vulnerability less than or equal to 27% of demand.

A summary of the objectives and performance measures for the six selected plans is given in the table below. The selection of a final plan requires the use of multi-criteria analysis methods (Section 12.7) considering all six objectives and performance metrics.

Solution number	2010 NPV total system cost ($billion)	Total system GHG emissions (millions tonnes CO_2e-)	Average maximum annual vulnerability (% of demand)	Reliability (%)	Average maximum resilience (days of failure)	Robustness (% of scenarios exhibiting acceptable performance)
1	3.16	6.28	23.7	95.1	92.9	64.8
2	4.31	5.09	23.5	95.5	116.6	43.9
3	3.27	6.82	14.7	95.1	73.6	63.9
4	4.14	8.94	0.0	100.0	0	78.7
5	5.08	7.59	0.0	100.0	0	80.6
6	6.23	5.27	7.3	97.3	28.5	78.7

11.6 RELIABILITY OF ENGINEERING COMPONENTS

Many engineering systems contain a large number of components, any one of which can fail at a particular point in time with varying consequences to the overall system performance. For example, a water supply system for a city may consist of thousands of pipes, hundreds of pumps and valves and tens of tanks. The impact of the failure of each one of these will vary depending on its function and location in the system, its capacity, the time of day and of the year, the number of customers affected and so on. We now want to consider how the reliability of system components may vary over time.

Figure 11.7 shows a commonly observed pattern for failure rates of engineering components (Agarwal, 1993; Hyman 1998). This is the so-called "bath tub" model of component failures.

The probability of failure per unit time is high initially during the "break in period", as some components fail due to poor materials or workmanship as they are placed under stress for the first time. During the mature phase of operation, a small number of failures occur due to unexpected causes such as occasional high loads or localised weak points. During this period, the failure rate per unit time is approximately constant. As the components begin to reach the end of their useful lives, the failure rate starts to increase as the effect of corrosion and fatigue take effect. This is the so-called "wear out period".

A simple analysis can be carried out for the mature phase by assuming that the probability of a component failure per unit time is constant and equals λ. Under these conditions, Hyman (1998) derives the following equation for the reliability of a component:

$$R(t) = e^{-\lambda t} \tag{11.19}$$

where *R(t)* is the probability that the component does not fail in the time period 0 to *t*.

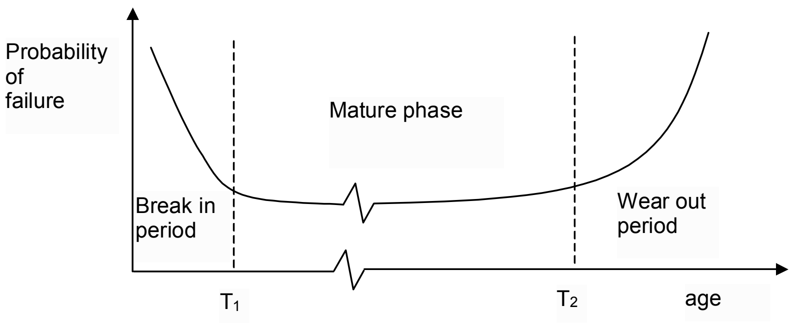


Figure 11.7 Typical failure rates for engineering components.

Another important measure is the mean time to failure, *MTTF*, which is defined as the average time that a component will be in service before failure occurs. Hyman (1998) shows that the following relationship holds:

$$MTTF = \frac{1}{\lambda} \tag{11.20}$$

Example: Light bulbs manufactured by a particular company have a mean time to failure of 250 hours. Assuming that the bulbs have a constant probability of failure, what is their reliability for a period of 100 hours? 250 hours?

Answer: Applying Equation (11.20),

$$\lambda = \frac{1}{MTTF} = \frac{1}{250} = 0.004 \text{ hour}^{-1}$$

therefore:

$$R(100) = e^{-0.004x100} = 0.670$$
$$R(250) = e^{-0.004x250} = 0.368$$

Thus, only 36.8% percent of the light bulbs are expected to last to the mean time to failure and 67% will last 100 hours or more. The apparently high failure rate in the first 100 hours (33%) is a consequence of the assumption of a constant failure rate. In reality, they are more likely to have a low initial failure rate that gradually builds up over time.

11.7 SYSTEM RELIABILITY

As noted earlier, an engineering system is composed of a large number of non-identical components that interact in some defined way. The failure of individual components can have different consequences on the performance of the system as a whole, depending on the nature of the components and how they are connected together. We will consider two simple cases to demonstrate this effect.

Series System

Consider a system that consists of a set of components in series. For example, consider a chain comprised of n links. The failure of any one of the links will cause the chain to fail. From elementary probability theory, the probability that the system does not fail is the product of the individual probabilities of non-failure (assuming that the failure of each component is independent), i.e.,

$$R_{series} = R_1 \times R_2 \times R_3 \times \ldots R_i \times \ldots R_n \tag{11.21}$$

where R_{series} is the reliability of the series system and R_i is the reliability of component i. Therefore, a series system has a reliability that is less than the lowest reliability of any of its components.

In the special case that each component has a failure rate that is constant with time, Equation (11.19) gives:

$$R_i = e^{-\lambda_i t} \tag{11.22}$$

and hence Equation (11.21) becomes:

$$R_{series} = e^{-\lambda_1 t} \times e^{-\lambda_2 t} \times e^{-\lambda_3 t} \times \ldots e^{-\lambda_i t} \times \ldots e^{-\lambda_n t} \tag{11.23}$$

hence:

$$R_{series} = e^{-\lambda_s t} \tag{11.24}$$

where the failure rate of the series system per unit time is given by:

$$\lambda_s = \lambda_1 + \lambda_2 + \ldots + \lambda_i \ldots\ldots\ldots + \lambda_n \tag{11.25}$$

Thus, in a series system where each component has a constant failure rate, the system failure rate will also be constant and will equal the sum of the individual failure rates.

Parallel Systems

A parallel system will not fail unless every component of the system fails. An example is a set of parallel pipes supplying water to part of a city. Clearly if one pipe fails, water can still be supplied to the city (presumably at a reduced pressure or level of service). However, if failure is defined as having no water available, all pipes must fail before system failure occurs.

For a parallel system with *n* components, the probability of system failure is the probability that all components fail. Therefore, (assuming independence of the individual components) the probability of failure for the system as a whole is given by:

$$P_{f.parallel} = P_{f1} \times P_{f2} \times \ldots P_{fi} \times \ldots P_{fn} \tag{11.26}$$

where $P_{f.parallel}$ is the probability of failure of a parallel system and P_{fi} is the probability of failure of component *i*. As reliability is defined as one minus the probability of failure, Equation (11.26) becomes:

$$\left(1 - R_{parallel}\right) = (1 - R_1) \times (1 - R_2) \times \ldots (1 - R_i) \times \ldots (1 - R_n) \tag{11.27}$$

It can be shown that the reliability of a parallel system is always greater than the reliability of its most reliable component. Given this high level of reliability, why don't we design all engineering systems as parallel systems? The answer is that there is a high cost of providing redundancy. Parallel systems are used when high levels of reliability are required. For example, all hospitals have an emergency power supply consisting of a motor-generator set that will start if the regular supply of electricity fails. This ensures that patients on life support systems are not at risk in the event of a power failure. Most pumping stations are designed to have a number of pumps in parallel. This not only allows for the flow provided by the pumps to vary as the demand on the system varies, but also allows for redundancy should one of the pumps break down.

Unlike series systems, parallel systems do not have constant failure rates per unit time, even if all of their components do.

Mixed Systems

Many real engineering systems do not correspond to simple series or parallel systems. They are described as mixed systems. A mixed system can be analysed by breaking it into a set of series and parallel systems and calculating the reliability one step at a time. For example, Figure 11.8(a) shows a mixed system. Figures 11.8(b) through 11.8(e) show thesteps involved in computing its reliability.

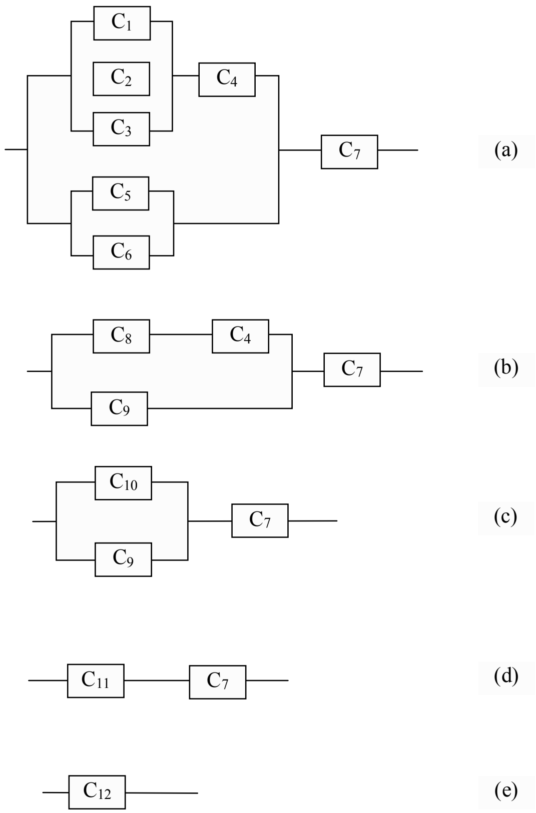


Figure 11.8 Steps involved in reducing a mixed system to a simple system.

11.8 SUMMARY

Knowledge of the concepts of risk and reliability are fundamental to the planning and design of engineering systems. Risk involves two components: the likelihood of an abnormal event or failure occurring and the consequences of that event or failure. All humans face risks on a daily basis. While the probability of failure of engineering systems cannot be reduced to zero, it is the role of the engineer to ensure that this probability is kept within bounds that are acceptable to the community.

The reliability of a system at a particular point in time is the probability that it will be in a non-failure state at that time. Engineering systems can be designed to achieve a specified level of reliability due to known hazards, although there are often unexpected hazards that are impossible to quantify. In the latter case, planning and design must rely on past experience or obtaining additional information that can help quantify the risks. Identification of the appropriate level of reliability of an engineering system may be viewed as a trade-off between the cost of construction or manufacture and the expected cost due to failure.

In addition to reliability, resilience, vulnerability and robustness of a system may also be important in planning and design. Resilience is the probability of the system returning from a failure state to a non-failure state in a single time step. Vulnerability is the average magnitude of failure of a system, given that it does fail. A system with a high reliability does not necessarily have a high resiliency or low vulnerability and vice versa. Robustness is a measure of the insensitivity of the performance of a given system or plan to uncertain future conditions.

Most engineering components wear out over time. Knowledge of this behaviour can be used to assess the reliability of a system or component over a specified period and hence to plan the repair or replacement of components in an optimal fashion.

An engineering system is usually composed of many different components, each with its own reliability. The reliability of the system can be determined based on the reliabilities of its individual components and the manner in which they are connected. However, the calculations can be very complex. Series systems and parallel systems represent simple cases of engineering systems.

PROBLEMS

11.1 A cable is used to raise and lower a hopper in a mine shaft. The properties of the cable are given in Table 11.5.

Table 11.5 Properties of cable for mine hopper.

Characteristic	Mean value	Standard deviation
Cross-sectional area (mm^2)	10	0.8
Failure stress (MPa)	300	30

The maximum load applied to the cable has a mean value of 2000 kN with a standard deviation of 400 kN. Assume all random variables are normally distributed and independent of each other.

(a) Estimate the mean and standard deviation of the resistance of the cable. Note if X and Y are two independent random variables with means m_x and m_y and coefficients of variation V_x and V_y (respectively) and $W = XY$, then the mean of W, m_w, and the coefficient of variation of W, V_w, can be found using the following equations (Benjamin and Cornell, 1970):

$$m_w = m_x m_y$$
$$V_w^2 = V_x^2 + V_y^2 + V_x^2 V_y^2$$

(b) What is the mean and standard deviation of the safety margin, Z?
(c) What is the value of the safety index, β?
(d) What is the probability of failure?
(e) If the coefficient of variation of the cross-sectional area remains constant at 8%, what is the mean value of cross-sectional area that corresponds to a probability of failure of 0.01? (Hint: Try trial-and-error using a spreadsheet.)

11.2 A computer part has a per unit failure rate of 0.002 failures per day.

(a) What is the probability that this part will fail in it first year of operation?
(b) What value of the per-unit failure rate will ensure a probability of failure of less than 5% in the first year?

11.3 A power generating system has four components with probabilities of failure of 0.003, 0.0025, 0.004 and 0.001 (respectively). The system only fails if all of the components fail. What is the reliability of the system?

11.4 Four system components (A, B, C and D) have the following reliabilities: 0.99, 0.90, 0.95 and 0.92 (respectively). Determine the reliabilities of the following systems:

(a) A and B combined in series.
(b) C and D combined in parallel.
(c) A and B combined in parallel and the resulting system combined in series with C and D.

11.5 A water pump has a reliability of 0.85. How many pumps should be combined in parallel so that the total system has a reliability of 0.95?

REFERENCES

Agarwal, K.K., 1993, Reliability engineering. (Kluwer Academic Publishers, Dordrecht, The Netherlands).

BC Hydro, 1993, Guidelines for Consequence-Based Dam Safety Evaluations and Improvements, Report number H2528, (BC Hydro: Hydroelectric Engineering Division).

Benjamin, J.R. and Cornell, C.A., 1970, Probability, statistics and decision for civil engineers. (McGraw-Hill, New York, NY.)

Bernstein, P.L., 1998, *Against the Gods. The Remarkable Story of Risk,* (New York, John Wiley & Sons).

Buck, N. And Associates, 2006, *Company Directors Course. Module 8: Risk Issues for the Board*, Australian Institute of Company Directors, Sydney.

Butler, D, Ward, S, Sweetapple, C, Astaraie-Imani, M, Diao, K, Farmani, R and Fu, G, 2016, Reliable, resilient and sustainable water management: the Safe & SuRe approach, *Global Challenges*, Open Access, John Wiley & Sons Ltd.

Dougherty, E.M. and Fragola, J.R., 1988, *Human Reliability Analysis,* (New York, John Wiley & Sons).

Fiksel, J., 2003. Designing resilient, sustainable systems. *Environ. Sci. Technol.*, 37(23), 5330–5339.

Hashimoto, T., Stedinger, J.R. and Loucks, D., 1982. Reliability resiliency and vulnerability criteria for water resource systems performance evaluation, *Water Resources Research,* 18(1), 14-20.

Heyworth, J.S., Glonek, G., Maynard, E.J., Baghurst, P.A. and Finlay-Jones, J. , 2006, Consumption of untreated tank rainwater and gastroenteritis among young children in South Australia, *International Journal of Epidemiology*, 35, 1051–1058.

Holling, C. S., 1973, Resilience and stability of ecological systems, *Ann. Rev. Ecol. Systems*, 4, 1-23.

Hyman, B., 1998, *Fundamentals of Engineering Design,* (Upper Saddle River, New Jersey, Prentice Hall).

Maier, H.R., Guillaume, J.H.A., van Delden, H., Riddell, G.A., Haasnoot, M. and Kwakkel, J.H., 2016, An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together?*, Environmental Modelling and Software*, **81**, 154-164, DOI: 10.1016/j.envsoft.2016.03.014

Morgan, G.C., 1990, Quantification of risks from slope hazards, Proceedings of GAC Symposium on Landslide Hazard in the Canadian Cordillera.

Paton, F.L., Maier, H.R. and Dandy, G.C., 2014, Including adaptation and mitigation responses to climate change in a multiobjective evolutionary algorithm framework for urban water supply systems incorporating GHG emissions, *Water Resources Research,* 50 (8), 6285-6304.

Reid, S.G., 1989, Risk acceptance criteria for performance-oriented design codes, Proceedings of ICOSAR'89, the 5th International Conference on Structural Safety and Reliability, (San Francisco).

Resilience Alliance, 2010, Assessing resilience in social-ecological systems: Workbook for practitioners. Version 2.0. Online: http://www.resalliance.org/3871.php

Stone, A. 1988. The tolerability of risk from nuclear power stations. *Atom 379,* May, 8–11.

U.S. Nuclear Regulatory Commission, 1975. *Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants.* WASH-1400 (NUREG 75/014).

Warner, R.F., Rangan, B.V., Hall, A.S. Faulkes, K.H., 1998, *Concrete Structures*, (Melbourne: Longman).

CHAPTER TWELVE

Engineering Decision Making

Decision making is an essential part of engineering work, and, in a real sense, planning and design are sequential decision-making processes. This chapter begins with a short introduction to mathematical decision theory, which is based on the criterion of maximising monetary gain, or, more generally, maximising utility. A short review is then given of game theory, which studies decisions that are made in competitive situations. Although engineering decision problems are often too complex to be amenable to analysis using decision theory and game theory, these disciplines provide a useful framework for approaching and structuring engineering decision problems. One complicating factor in engineering decision making is that there will often not be just one decision criterion to apply, but several, or even many. A further complication is that there can be many stakeholders involved in the decisions. Although it is not possible to find rigorous solutions for complex decision problems, there are techniques that can lead to good and practical (if not optimal) outcomes. These are discussed at the end of this chapter.

12.1 INTRODUCTION

Decision making is an essential activity in the work of engineers and other professionals, as well as in our personal life. Each day we make countless personal decisions: they begin when we wake up in the morning and end only when we go to sleep at night. Most are routine and are undertaken almost automatically, with little thought or pre-planning. For example, when to get up in the morning, which clothes to wear, and what to eat for breakfast, are routine decisions for most people. Some engineering decisions are also minor and can be treated routinely, as for example when a specific type of vibrator is to be chosen from the commercial range available, for use in placing and compacting fresh concrete.

On the other hand, some decisions, both private and professional, require more careful attention, perhaps because there are long-term consequences or because large sums of money are involved. The purchase of an automobile is a personal decision that would not be routine for many people. We need to have at hand a good, reliable decision-making procedure for such circumstances. A common-sense approach consists of the following steps:

identify the alternative courses of action that are available;
work out the consequences that can follow from each of these actions, and their relative likelihood of occurrence; and
compare and evaluate the consequences, using some criterion such as minimum cost or personal preference, and hence identify the best action.

This approach to decision making is sound and rational, and bears similarities to the basic approach to planning and design that was discussed in Chapter 3. However, it cannot be implemented unless appropriate *decision criteria* are available for evaluating the alternative outcomes, and hence ranking them. Such criteria depend on the type of decision being made and possibly also on the specific details of each decision.

Decision theory

Decision problems have been studied mathematically since the 17th and 18th Centuries, when Blaise Pascal and Daniel Bernoulli made the first pioneering contributions to the emerging field of decision theory. Many of the early theoretical studies were of gambling decisions in which maximising monetary outcome was the objective. Today, there is a large body of knowledge that provides decision criteria and methods of analysis for many different decision problems. Although these problems tend to be simple and rather idealised when compared with the real-world situations that occur in engineering practice, they do have practical applications. Even when no theoretical approach is directly applicable to a specific decision, a useful insight into the decision-making process, and into the structuring of a problem, can be obtained from a study of decision theory. Decision theory deals with a range of problem types, but usually only a single criterion is employed, such as monetary value. To provide a brief introduction to decision theory, we will look at the following types of decisions:

One-off and *many-off decisions:* A one-off decision is made when there is a unique set of circumstances that will not be repeated. An example is the siting of a dam. In contrast, a many-off decision applies when a single decision is applied to many identical situations. For example, when an automobile manufacturer chooses a new electro-mechanical device for the automatic parking of vehicles, the intention is to make many thousands of installations, and it is in this sense that the decision is many-off. The distinction between many-off and one-off decisions is important because different decision criteria may be required to deal with each case.

Single-stage and *multi-stage decisions:* A single-stage decision is simply a one-off decision, while a multi-stage decision consists of a sequence of inter-related decisions, so that the outcome of the n-th decision could depend on previous decisions and outcomes. Planning and design are, in effect, multi-stage decision processes. An example of a single-stage decision is the siting of a new airport for a city. An example of a multi-stage decision sequence is a week-by-week program for operating a system of reservoirs to supply water and hydro-electricity to a city.

Decisions with risk, certainty, uncertainty and ignorance of outcome: An important distinction to be made among decisions depends on the amount of information available to the decision maker. In a decision with *risk of outcome* all the possible outcomes for each possible course of action are known in advance, as well as the probability of occurrence of each outcome. If more information is available, so that the decision maker knows which single outcome will necessarily follow on from each possible action, we say that there is *certainty of outcome*. This is a no-risk situation. *Uncertainty of outcome* exists when much less information is available, so that although all possible outcomes for each action are known, there is not enough information available to evaluate their relative probabilities. In an even less-desirable situation for the decision maker, the information available is so meagre that the outcomes from a possible course of action are unknown. In such a situation, which we will refer to as *ignorance of outcome*, the decision maker should consider postponing the decision and “buy” (or in some other way obtain) additional information and so improve the situation to one of uncertainty, or possibly risk. Deferral of a decision for the purpose of obtaining additional information is often appropriate, also for uncertainty problems and even for risk problems, and is an option always to be considered.

The types of decisions just mentioned are studied in decision theory using the criterion of dollar value, or more generally utility, and are discussed in Sections 12.2 to 12.6 of this chapter.

Game theory

Engineering and business decisions are often made in situations where there are other participants in competition with the decision maker. For example, when several construction companies tender for the same project, each company chooses a specific tender price from a range of feasible values, knowing that the outcome of the tendering process will depend on the courses of action of all the competitors. *Game theory* is a relatively new field of mathematics which extends the ideas of decision theory to deal with such competitive situations. A brief introduction is given to game theory in Section 12.7.

Complex engineering decisions with multiple criteria and multiple participants

Engineering decision problems become complex when there is not just one criterion, such as cost, but many relevant and even conflicting ones to consider. Engineering decisions can become even more complicated, and sometimes intractable (in the sense discussed in Chapter 3) when various persons and/or organisations are affected by the outcomes. As *stakeholders*, they can take part in the decision making. Researchers have found that optimal mathematical solutions are rarely feasible in these circumstances. On the other hand, there are pragmatic approaches that can provide realistic and acceptable solutions that are “good”, if not optimal. A discussion of practical approaches to complex engineering decision making is presented in Sections 12.8 and 12.9 of this chapter.

12.2 SINGLE-STAGE DECISIONS WITH RISK OF OUTCOME

In a single-stage decision with risk of outcome, the possible alternative courses of action are known and are denoted as a_1, .. a_i, .. a_m. The possible outcomes from each action a_i are also known, and are denoted as O_{i1}, .. O_{ij}, .. O_{in}. The outcome that results from any action a_i depends not on the decision maker, but on chance and outside circumstances. The number of outcomes may differ from action to action, but if we take n to be the maximum number of outcomes among all the actions, we can introduce nil entries as needed to create an mxn format for the problem.

It is sometimes convenient to recast this situation by imagining various possible *states of nature* that can follow on after a course of action has been chosen, and which will determine the outcome of the action. The states of nature are denoted as S_1, .. S_j .. S_n, and one of these, say S_j, is the one that in fact occurs and determines the outcome of decision a_i, which is $O(a_i, S_j)$, or, as previously, O_{ij}. Problem formulation in terms of states of nature can be advantageous in some situations and will be used in later sections of this chapter.

Risk of outcome means that the decision maker has sufficient information available to estimate the relative probabilities of occurrence of the outcomes of each action. For action a_i the probability of occurrence of outcome O_{ij} is p_{ij}.

The following example illustrates the decision-under-risk problem: A construction company is considering hiring new equipment to speed up its construction operations. The alternative is the status quo. The hiring cost of the new system is estimated to be an additional \$5m (million) per year. After checking past records of the company's operations, it is considered that the success or otherwise of the new equipment will depend on the type of work coming in.

After some simplification and idealisation, three possible outcomes are chosen to apply if the new equipment is hired: in the best case, there will be an increase in efficiency which leads to profits of \$19.5m (million) per year, with a 60 percent probability. The second possibility is that a profit of \$15m is obtained, with a likelihood of 25 percent. The third, and worst case scenario, is a loss of \$15m at a probability of 15 percent. The hire cost of \$5m for the new equipment must be subtracted from each of these outcomes. For the option of staying with the current equipment, it is estimated that the profit for the year will either be \$5m, with a probability of 40 percent, or \$8m with a probability of 60 percent.

There are two actions to compare, with five possible outcomes that are stated directly in dollar values. The actions are:

> a_1: hire the new equipment on a yearly basis; or
> a_2: stay with the current equipment arrangement.

The dollar outcomes for action a_1, including the hiring cost of \$5m, are

> O_{11}: a profit of \$14.5m, with a probability of 0.6;
> O_{12}: a profit of \$10.0m, with a probability of 0.25;
> O_{13}: a loss of \$20.0m, with a probability of 0.15.

The outcomes for action a_2 are:

> O_{21}: a profit of \$5.0m, with a probability of 0.4;
> O_{22}: a profit of \$8.0m, with a probability of 0.6.

Decision tree representation of a decision with risk of outcome

This type of problem can be represented graphically using what is known as a *decision tree*, which can be easier to comprehend than the verbal statement of the problem. Figure 12.1 below shows the decision tree for this example.

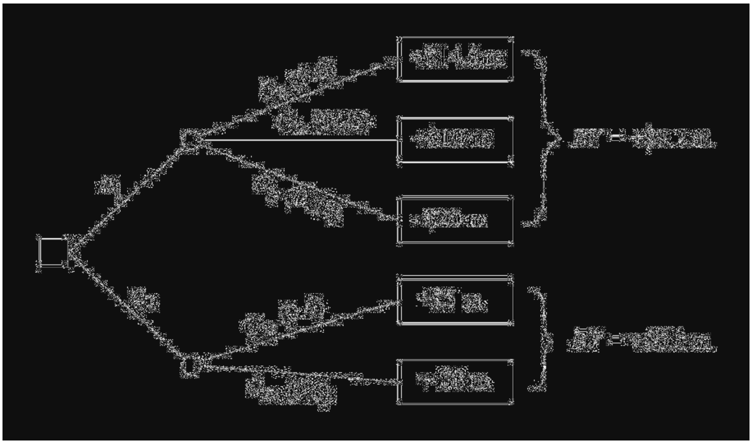


Figure 12.1 Decision tree for the building company decision.

In such a tree, a small square box represents a decision node and a small circle represents an outcome node. The possible actions are represented by lines emanating from decision nodes. In Figure 12.1, the two action lines a_1 and a_2 each end in an outcome node because this is a one-step decision. Lines emanating from a (circular) outcome node represent the possible outcomes from a course of action. Each outcome line is tagged (in general, as O_{ij}) and the probability of occurrence of the outcome is also shown. In Figure 12.1, the action lines a_1 and a_2 have three and two possible outcomes, respectively. In this example, each outcome is expressed as a dollar value, $O_{ij} = V(O_{ij})$, and the dollar values are given in the boxes on the right side at the end of the outcome lines. Note that the sum of probabilities for lines emanating from any outcome node must add up to unity.

Matrix representation of the decision

The numerical data for a decision with risk of outcome can also be presented using matrices for the payoffs and the probabilities. Table 12.1 below contains the information in Figure 12.1. It will be seen that a nil entry applies for action a_2 and S_3. This leads to a zero entry for the probability p_{23}. Decision trees become unwieldy when the numbers of actions and possible outcomes are large, and the matrix representation then becomes more useful. The matrix representation will be used for decision problems later in this chapter.

Expected value decision criterion

To proceed, we observe that the *expected value criterion* is often used as the criterion for decision making when monetary values can be placed on all possible outcome. The expected value of an outcome is calculated as its monetary value, multiplied by its probability. The expected value of a course of action a_i is the sum of the expected values of the possible outcomes:

$$EV(a_i) = \sum_{j=1}^{n} V(O_{ij}) p_{ij} \tag{12.1}$$

Here $V(O_{ij})$ is the monetary value of the j-th possible outcome of action a_i, and p_{ij} is the probability of occurrence of O_{ij} when action a_i has been taken. Altogether there are n possible outcomes for action a_i. The expected value criterion chooses the action with the highest expected value as the best one. In our example, the expected values of the actions are

$$EV(a_1) = 0.6 \times 14.5 + 0.25 \times 10.0 + 0.15 \times (-20.0) = +\ \$\ 8.2\ \text{m};$$
$$EV(a_2) = 0.4 \times 5.0 + 0.6 \times 8.0 = +\ \$\ 6.8\ \text{m}.$$

These values are shown on the right of Figure 12.1, beside the boxes associated with the action. According to the expected value criterion, action a_1 is preferred.

Table 12.1 Matrix representation of example in Figure 12.1, decision with risk of outcome.

(a) Payoff matrix

		States of Nature		
Payoffs in \$m:		S_1	S_2	S_3
	a_1	+14.5	+10	-20
Courses of action:	a_2	+ 5	+ 8	*

(b) Probability matrix

		States of Nature		
Probabilities:		S_1	S_2	S_3
	a_1	0.6	0.25	0.15
Courses of action:	a_2	0.4	0.60	0.00

This example may seem to be overly idealised and therefore unrealistic. In the real situation, there would certainly not be just three or two possible outcomes, but a wide range of profits and losses to consider. There would also be more than two possible actions. Any number of actions and outcomes can of course be accommodated in the expected value calculations, but, as in all engineering modelling, idealisation and simplification are appropriate. The use of just three possible outcomes for decision a_1 approximates a more complex situation. Such idealisations should give an acceptably close representation of the real situation. This aspect of modelling has been discussed previously in Chapter 2.

Applicability of the expected value criterion

The obvious question that arises is: how appropriate is the expected value criterion in this particular situation? Or, more generally, under which circumstances is it appropriate to use the expected value criterion for decision making?

If the decision is many-off, that is, repeated many times, the expected value of an action will be equal to its long-term return, and it is then highly appropriate to choose the action with the highest expected value. The choice of the type of non-structural cladding panel to be used in the construction of a multi-storey building is a many-off decision, in that a very large number of panels will be used, and the expected value criterion would be appropriate.

If many one-off decisions with risk have to be made, which are all rather similar in nature and with similar ranges of values of outcomes, then the expected value criterion will also be appropriate. Here again, the long-term return for the class of decisions will be equal to the expected value.

But how appropriate is the expected value criterion for the example in Figure 12.1? The answer will depend on the size of the construction company and the number of similar decisions that are being regularly made. For a large company which undertakes comparable work in many cities and regions, the expected value criterion will be suitable for reasons already given. On the other hand, if the company is small and has limited financial backing, the possible loss of $20m could be of real concern, even though the probability of this outcome is relatively low, at 15 percent. The expected value criterion is not appropriate in this case.

Risk aversion and risk acceptance

There are many situations where the expected value criterion will not be appropriate and should not be used. In the above example, a loss of $20m, even at a relatively low probability, would be unacceptable because it could lead to bankruptcy of the company. The expected value criterion is not acceptable because the decision maker is averse to risk. *Risk aversion* usually exists when the stakes are high in comparison with the financial resources that are available to the decision maker.

Risk aversion can also be seen in personal decision making, for example when a person takes out insurance on a house or car. The expected value here will always favour the insurance company (otherwise it would not be in business). Nevertheless, the individual willingly pays a premium, which is a small proportion of the cost of the house or car, to be assured that replacement or repair will occur in the very unlikely event of damage or loss.

Almost the opposite situation can arise when an individual is willing to gamble a small amount of money in the hope of making a very large gain, as occurs with the purchase of a lottery ticket. This behaviour is called *risk acceptance*. The expected value criterion also advises against such an action, but the near-certain loss of a few dollars (the purchase price of the ticket) is acceptable to many people in return for the very unlikely chance of winning lots of money.

It is not irrational for risk aversion and risk acceptance to be displayed simultaneously by the same person, who is averse to large, unaffordable losses, yet finds quite acceptable a small loss with the possibility of a large, but unlikely, win.

Utility and the utility function

Risk aversion and risk acceptance can be taken into account using the concept of *utility*, which was first proposed by John von Neumann and Oskar Morgenstern in 1947 (Neumann & Morgenstern, 1953). A valid decision criterion is obtained by setting up an appropriate utility function that suits the personal preferences of the decision maker in regard to risk aversion and risk acceptance. This function translates the monetary values of the possible outcomes in the decision situation into units of utility. Maximum expected utility then becomes a valid decision criterion for choosing the best course of action.

The essential step is the creation of the utility function. It is emphasised that this function will be subjective and it will only apply to the decision maker for whom it is set up. Also, it will apply only to the decision situation at the time it is set up. This is because the decision maker's attitude can change significantly over time, and as external circumstances change.

To set up the utility function, the possible outcomes for the particular decision situation are inspected and the most favourable and the least favourable are identified and are denoted as O^* and O_*. Their monetary values are $V(O^*)$ and $V(O_*)$, respectively. An arbitrary range of utility values is chosen. This could for example be from 0 to 100 units, or 0 to 10 units, or –100 to +100 units. The highest value of utility is allocated to outcome O^*, and the lowest value to O_*.

A "game" for the decision maker is now created. An intermediate outcome is chosen, say, O_x, which has a monetary value $V(O_x)$ that lies between $V(O^*)$ and $V(O_*)$. The purpose of the game is to evaluate the utility of O_x, which is denoted as $U(O_x)$. The decision tree in Figure 12.2 specifies the game. There are two possible courses of action, a_1 and a_2.

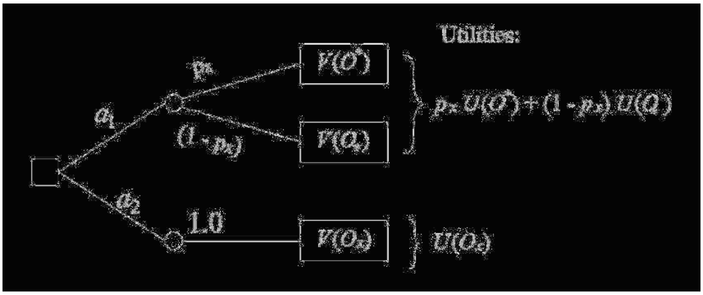


Figure 12.2 Decision tree for evaluating utilities.

Action a_1 has two possible outcomes, O^* and O_*, with values $V(O^*)$ and $V(O_*)$ and allocated utilities $U(O^*)$ and $U(O_*)$. The associated probabilities, p_x and $(1 - p_x)$, are initially unknown. Action a_2 has only one possible (and therefore certain) outcome, which is O_x with the value $V(O_x)$. The decision maker is asked to consider the decision and choose the specific value p_x, such that decisions a_1 and a_2 become equally acceptable. The decision maker can try out several values of p_x in turn. For the extreme values, we can see that $p_x = 1$ leads to the choice of action a_1, but $p_x = 0$ will result in a_2 being preferred. At some intermediate value, p_x, the actions a_1 and a_2 become equally acceptable, and a_1 and a_2 then have the same value of utility, $U(a_1) = U(a_2)$. The utility of O_x is obtained as:

$$U(O_x) = p_x\, U(O^*) + (1 - p_x) U(O_*) \qquad (12.2)$$

The game is replayed using differing outcomes O_x so that a functional relation between utility and dollar value is created. A numerical example is provided below.

Expected utility as decision criterion

The utility function is an appropriate criterion for decision making under risk for the case where the decision is one-off. This is because the transformation of dollar value into utility units and the use of the utility measure of value allows for the attitude of the decision maker to risk, including both risk aversion and risk acceptance. The expected utility for an action a_i is

$$EU(a_i) = \sum_{j=1}^{n} U(O_{ij})p_{ij} \qquad (12.3)$$

If decisions are being made on behalf of a corporation or an organisation or a business, an individual or a small group of people can be chosen to represent the best interests of the organisation in order to set up the appropriate utility function.

With the expected values of the outcomes replaced by expected utility values, the decision tree can be modified to treat utilities rather than dollar values. The utility criterion can be used in other decision situations, and will be discussed again in later sections of this chapter.

Numerical example using the utility criterion

A re-analysis of the problem of Figure 12.1 will illustrate the method. We choose a range of utilities between + 20 and – 20, so that the upper and lower values are

$O^* = O_{11}$; $V(O^*) = +$ \$14.5m; $U(O^*) = +20$
$O_* = O_{13}$; $V(O_*) = -$ \$20.0m; $U(O_*) = -20$

We now consider several intermediate outcomes, the first being $V(O_x) = +$ \$10m. To obtain equal value for actions a_2 and a_1, where a_2 leads to a certain gain of +\$10m, the decision maker representing the small firm would require a very high value of p_x, because a loss of –\$20m could be disastrous. A value of $p_x = 0.98$ is chosen. The corresponding utility value is then:

$U(V(O_x) = +$ \$ 10m$) = 0.98 \times 20 + (1 - 0.98) \times (-20) = 19.2$

The next values we choose for O_x are zero and −\$10m, which yield lower values of p_x, say 0.85 and 0.6, respectively. Any number of additional intermediate values can be used. The resulting relation between utility and dollar value (in millions of dollars) is shown in Figure 12.3. From Figure 12.3 we can read off the utility values:

$U(O_{11}) = U(+\$14.5m) = +20$
$U(O_{12}) = U(+\$10.0m) = +19$
$U(O_{13}) = U(-\$20.0m) = -20$
$U(O_{21}) = U(+\$5.0m) = +17$
$U(O_{22}) = U(+\$8.0m) = +18$

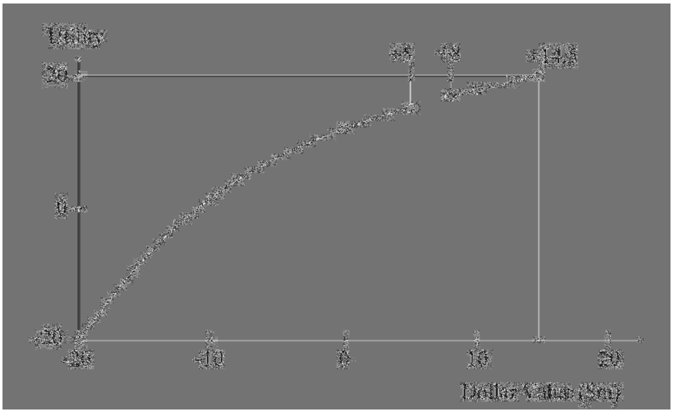


Figure 12.3 Utility function for example.

From these utility values we calculate the expected utilities for a_1 and a_2:

$U(a_1) =$ 0.6 × 20 + 0.25 × 19 + 0.15 × (- 20.0) = + 13. 75
$U(a_2) =$ 0.4 × 17 + 0.6 × 18 = + 17.6

The preferred decision now, allowing for risk aversion, is clearly a_2, which is, intuitively, the better one. A decision tree representation of the problem, with utilities in lieu of dollar values, is shown in Figure 12.4.

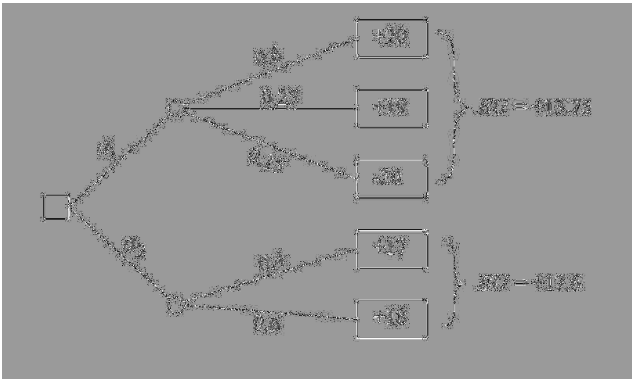


Figure 12.4 Decision tree with utilities instead of monetary value.

To summarise: It is appropriate to use the *expected value criterion* for many-off decisions with risk of outcome. It can also be used when many different but similar decisions are to be made, in which the ranges of outcomes are comparable. In one-off decisions, the *expected utility criterion* should be used whenever the decision maker shows risk aversion or risk acceptance.

12.3 MULTI-STAGE DECISIONS WITH RISK OF OUTCOME

In each step of a multi-stage decision sequence, the decision can depend on the outcomes of previous decisions. The procedure for analysing a multi-stage decision process follows from the observation that, irrespective of any prior decisions and outcomes, the best course of action to follow at any particular stage is to maximise the total payoff for the future steps. This idea leads to a *last-first strategy* which is also used to solve sequential optimisation problems, as will be seen in Chapter 13.

The strategy is illustrated with the following example. In planning for an upgrade to the water supply for a developing outer city region, the changes that occur over two time periods are being evaluated. In the first time period, there are two options: to install either a small capacity pipeline (action a_{11}) or a large capacity pipeline (action a_{12}). In modelling the problem, two possible outcomes (or states of nature) are considered for the first time period: either there is low demand or high demand for water. In the second time period, action a_{11} is followed up with a second-stage decision, namely to do nothing (a_{21}) or to construct a second parallel pipeline to increase capacity (a_{22}). This second decision is again followed by two possible states of nature, or outcomes: low or high demand. The second stage of the process involving a_{21} could well be followed by further stages, but to keep the example simple we consider just a two-stage process. There is no second-stage action if the first-stage decision is to install a large-capacity pipeline (a_{12}).

Figure 12.5 shows the decision tree for this example, where, as previously, a small square box represents a decision node, with possible decisions represented by lines emanating from the box and ending in small circular outcome nodes. The possible outcomes following a decision are represented by lines emanating from the circular node at the end of the action line. The dollar costs, shown in the boxes at the tips of the branches on the right side of Figure 12.5, take account of lost opportunity, as for example when the decision to install a large pipeline is followed by low usage.

In the first stage of the process the probabilities of the states of nature are 0.4 and 0.6 for low demand and high demand, respectively. However, in the second stage, the probabilities are 0.2 and 0.8 for low demand and high demand, in the case where high demand has previously been experienced in the first stage. In the case of low demand in the first stage, the second-stage probabilities are 0.7 for low demand and 0.3 for high demand.

The calculations are carried out in terms of costs, the negative signs are dropped, so that the criterion is now to minimise cost. The dollar signs are also omitted. To apply the last-first strategy, we consider the final stage of the process and examine the outcomes at the boxes in turn. The expected value of each possible outcome is determined from the value entered in its box and the associated probability. Working our way from right to left, i.e., backward in time, we start at node 3, where the expected cost is 0.7 x (70) + 0.3 x (325) = 146.5. The expected value at node 4 is 0.7 x (365) + 0.3 x (325) = 353.

Moving back to decision node B, where the expected value of each course of action is obtained by summing the expected values of the possible outcomes, we see that action a_{21} is preferred. We write the expected value 146.5 at the node and use an arrow on the a_{21} branch to indicate that this is the preferred action at this point.

The expected costs at outcome nodes 3 to 8 are shown in Figure 12.5. At decision node C, the expected cost is equal to that for the preferred action a_{21}, i.e. 316 from action node 5. Working our way from right to left, we now reach action node 1, where the expected cost is obtained from those at B and C as 0.4 x (146.5) + 0.6 x (316) = 248.2. This is the expected value of action a_{11}.

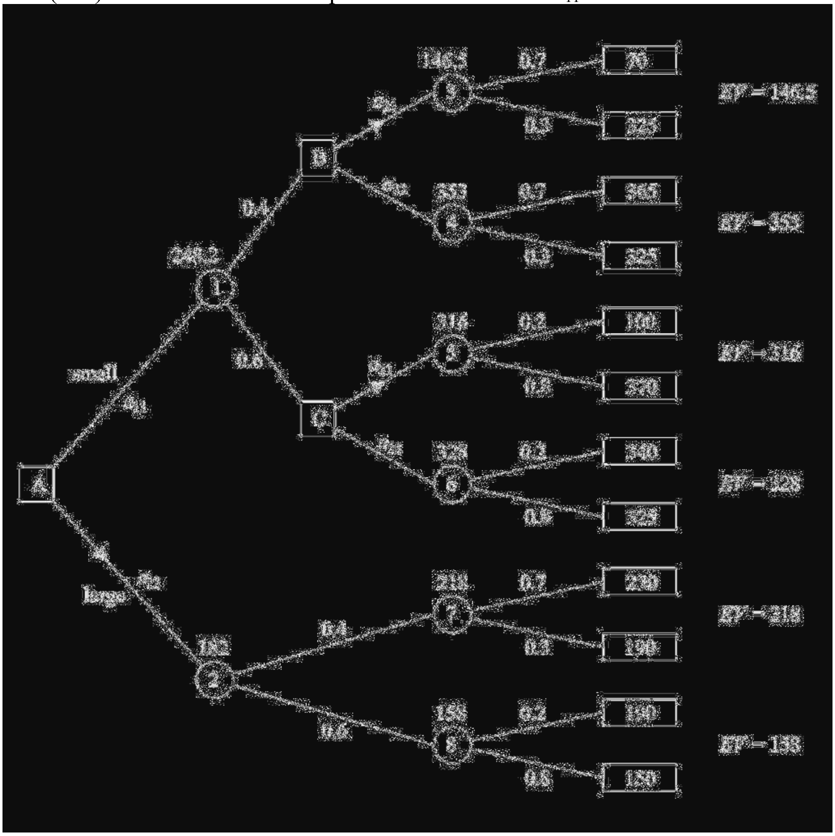


Figure 12.5 A multi-stage decision for a water supply expansion project.

In this way we can move back, stage by stage and repeat the evaluations. As there is no second action to follow a_{12}, the expected cost at node 2 is the weighted sum of the values at nodes 7 and 8: 0.4 x (218) + 0.6 x (158) = 182, which is the lower cost for the two possible actions, so that a_{12} is preferred. The preferred action, a_{12}, is to install the large pipe.

When the decision process has been analysed by this last-step-first procedure, the appropriate courses of action are obtained at every stage of the process, irrespective of the previous decisions and previous outcomes. If for example action a_{11} had initially been chosen, despite the outcome of the analysis, the appropriate course of action in the second stage is shown in the decision tree. The best actions at every stage of the process are indicated by the arrows leaving the (square) decision nodes.

In Figure 12.5 it will be noted that two different actions have been denoted as a_{21}, and another two actions have likewise been denoted as a_{22}. Here there is little chance of ambiguity as the like-named actions emanate from different decision nodes (B and C). In a more complex, multi-stage problem the actions can be given a marker to identify the decision node, as for example a^B_{21} and a^C_{21}. Multi-stage decision processes can clearly arise for which the decision tree becomes a very complex network.

If the multi-stage process has certainty of outcome, there are network optimisation procedures, such as linear programming and dynamic programming, that can be used to obtain a solution. They use the last-first approach as described here. Chapter 13 deals with such procedures, although not specifically in relation to decision analysis.

We have used expected value for our decision criterion in this example. This type of problem would be handled by a government authority or a large organisation. Multi-stage decision problems may also involve possible significant adverse financial results. It would then be appropriate to set up a utility function as explained previously in Section 12.3 and apply the expected utility criterion.

12.4 DECISION MAKING WITH CERTAINTY OF OUTCOME

Some engineering decisions are carried out in situations where a great deal of relevant background information is available, and the decisions concern a deterministic system. In planning a space mission, NASA engineers need to choose a launch time such that the planets and moons are suitably positioned to optimise travel time and minimise fuel loads. In the case of a mission to Mars, a suitable launch window of a month or two comes up every two years (Squyres, 2005). The choice of launch time is based on full knowledge of the relative locations of the celestial bodies (which are calculated accurately using Newton's laws of motion). Provided the appropriate decision criterion has been clearly stated, the decision follows from the necessary computations.

Even when there is full knowledge of the consequences of each possible course of action, engineering decisions must be based on suitable criteria that allow the alternative possible actions to be evaluated, compared and ranked. A complicating factor will often be that there is not just one objective to be achieved or optimised, but several or even many. The case of multi-criteria decision making is discussed below in Section 12.7.

12.5 DECISION MAKING WITH UNCERTAINTY OF OUTCOME

In some situations there may not be sufficient information available for the decision maker to determine the probability of occurrence of the outcomes. Various methods have been proposed for dealing with this uncertainty situation.

To discuss these, we will use an example in which engineers must choose the water level to keep in a dam. The problem is simplified to the following. There are two options to be evaluated: one (action a_1) is to keep the level deliberately low in the hope that good rains will come and fill the dam; the other is to choose a more conservative action (a_2), and maintain a medium-high level. From previous records, it is estimated that keeping the level low would lead to a good payoff, of 100 units, if good rains come, but a poor outcome of 20 units if there is little rain. The conservative approach a_2 has a lower payoff in the event of good rains (50 utility units) but a reasonable payoff with poor rain (40 units).

We will consider this problem with unknown probabilities, but for comparison purposes we first look at the decision process if the probabilities are known, and are as given in Figure 12.6. With a 20 percent chance of good rains and an 80 percent chance of poor rains, the expected utilities for actions a_1 and a_2 are: $EU(a_1) = 36$ and $EU(a_2) = 42$, so that the decision to be taken would be a_2.

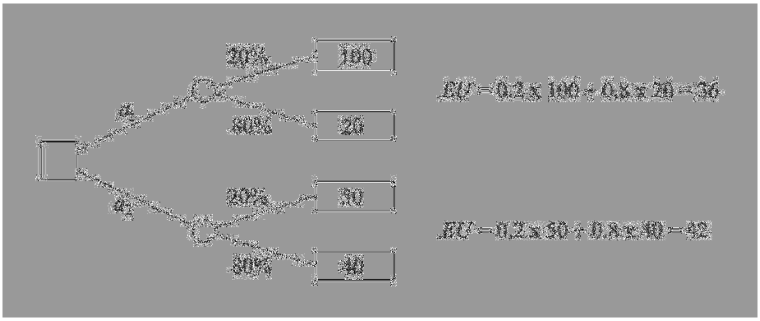


Figure 12.6 A decision tree with utilities and two possible courses of action, a_1 and a_2.

Laplace's principle

According to Laplace's principle (actually first described by Bernoulli), all outcomes are assumed to be equally likely in the case that the probabilities are unknown. The decision tree then becomes as in Figure 12.7. The preferred decision is now a_1 since it has the higher expected outcome (60 compared with 45). Laplace's principle provides a means for making decisions, but it can lead to poor decisions because it gives equal weighting to all possible outcomes, even ones that are considered very unlikely.

Pessimistic criterion

If the probabilities are unknown, one could assume the worst and minimise the potential losses. The choice will then be determined by the course of action that has the least-worst outcome. From Figure 12.7 we see that action a_2 would be selected because its worst outcome is better than for action a_1 (40 compared to 20).

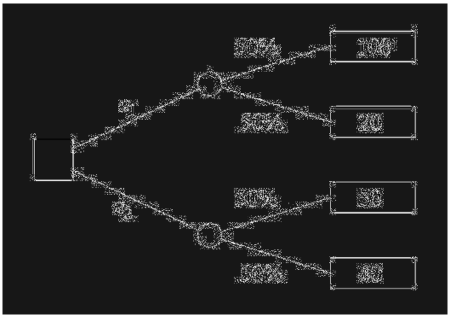


Figure 12.7 The decision tree for Laplace's method where all probabilities are assumed equal.

Bennett (2001) has pointed out that the precautionary principle in environmental theory is equivalent to decision making under uncertainty, and using the pessimistic principle. When decisions are made under competition (to be treated below in Section 12.7), a pessimistic approach is quite reasonable, because your opponent is malevolent, and this "minimax" strategy minimises the maximum damage your opponent can do (Holland, 1998). Looking from the point of view of making profits rather than suffering losses, ReVelle et al. (2004) refer to this as a "maxi-min" strategy where the aim is to maximise the minimum payout.

Optimistic criterion

In contrast to the pessimistic criterion, an optimistic approach would seek the best possible outcome, and maximise the maximum payout. According to Manhart (2005), humans in general have an optimistic bias in their decision making, particularly when it comes to personal preferences, but this will not assist them in a situation where an engineering decision is being made. In the example, a_1 would be preferred as it has the better outcome (100 compared to 50). This approach can lead to poor outcomes, and cannot be justified for engineering decisions.

Minimised regret criterion

Another approach is to minimise the *regret* that a decision can bring. This requires slightly more sophisticated thinking, and is a valid way to make decisions. Regret is taken to be the difference between the payoff for the chosen decision and the payoff for the best possible decision associated with the revealed state of nature. The decision tree for our example has been redrawn in Figure 12.8, where the consequences of two possible natural events, n_1 and n_2, are shown.

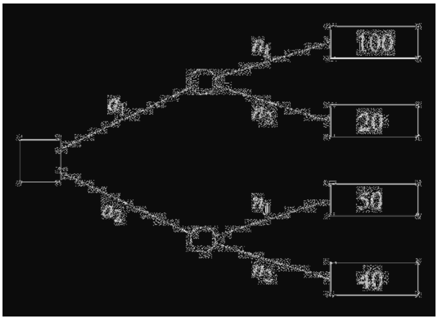


Figure 12.8 The decision tree redrawn with actions a_1 and a_2 and two possible states of nature n_1 and n_2 at unknown probabilities. The outcomes are unchanged.

We first consider the situation if n_1 occurs. Had action a_1 been selected, the decision maker would have no regrets since the chosen action gives the best result for n_1 (100 compared to 50). However, had action a_2 been chosen, the regret would be 50 units of utility because the best possible result with natural event n_1 is 100. For the second outcome, n_2, the regret would be 20 if action a_1 had been chosen, and no regret if a_2 had been chosen. These results are summarised in Table 12.2, where the regret, in utility units, is given for each action and outcome. To minimise maximum regret, action a_1 would be chosen.

Table 12.2 Values of regret for the problem in Figure 12.8.

Decision	Event n_1	Event n_2	Maximum regret
a_1	0	20	20
a_2	50	0	50

Subjective probabilities

If the probabilities are not known, the decision maker may wish to estimate them, using intuition and experience. This is a relatively simple procedure, but the results will clearly depend on the judgement and previous experience of the decision maker, and on how often similar situations have arisen in the past.

However, caution needs to be exercised if data are evaluated prior to choosing the probabilities. Winman et al. (2004) have suggested that answers to questions on probability values can depend on the way the question is framed. The decision maker might be asked to "estimate the lower and upper range of values for the probability of such-and-such occurring". Alternatively, the question might be "what are the chances that the probability of such-and-such occurring will be between x and y?" Both questions seem to be asking much the same question, but, by asking it in different ways, not only are different estimates usually given, but the confidence associated with the answer is quite different. As a matter of interest, the latter question was found to yield better answers than the former.

Unstructured uncertainty

When the decision maker does not even know the possible outcomes of the possible actions, the only responsible course of action is to defer the decision and obtain additional information. Buying information is discussed in Section 12.8 below.

12.6 GAME THEORY AND DECISION MAKING WITH COMPETITION

In August 1955, a decision was taken in Moscow to put a satellite into orbit around the earth before July 1, 1957. According to Schefter (2000) "[the] timing had only one purpose: to beat the United States of America into space". In the end, delays meant that it was not until October 1957 that the task was accomplished, but the aim had been achieved: the Americans had been beaten.

Engineering is often undertaken in a competitive environment. Jobs can be won or lost in a tendering system where an organisation describes the work it wants done in a call for tenders, and competing firms then bid for the opportunity to undertake the work. The bidding process is confidential, and each firm bids without information about the potential rival bids. The aim of each firm is to put in the most attractive bid and to be selected on the basis of cost. The organisation must of course be confident that the successful bidding firm is competent to undertake the work and that the bid is realistic.

When an engineering firm prepares a tender, it knows it is in competition with other firms, and that the chance of success depends in part on how their competitors act. This complicates the decision-making process. In preparing the bid, the firm will want to choose a dollar value as high as is reasonable to maximise profit; however, it knows that too high a bid will mean that another firm will put in a lower offer and win the tender, in which case there is no profit at all. On the other hand, if the successful bid is too low, the firm will lose money on the contract.

In effect, each firm goes through a two-stage decision process. In the first stage, there are two options: either to bid, or not to bid. If the decision is taken to bid, then the second stage becomes, in a simplified model, a choice among three options: bid high, bid low at a level that will do little more than cover costs, or bid in the intermediate range. In this situation, the outcome will depend on the actions of all the competitors and is therefore quite different to the situations treated previously using a decision tree.

Game theory is a branch of applied mathematics that deals with competitive decision making. In principle, the concepts apply to many competitive human situations such as that occur in games of strategy, business, economic behaviour, and warfare. Although real-world situations are often too complex to allow direct application of gaming models, a good insight into decision making in the face of competition can be gained from a study of the elements of game theory. We will look at some elementary aspects of the theory.

It is useful to distinguish between *strategic* and *non-strategic games*. Non-strategic games do not depend on the skills of the players, and do not require intelligent decisions to be made. The outcomes depend purely on chance and can be analysed using probability and statistics. Examples of non-strategic games include board games such as snakes and ladders (where the outcomes depend on the chance throw of dice), and betting on the outcome of a tossed coin or die.

John von Neumann and Oskar Morgenstern: the zero sum game

The foundations for game theory were set out in the book "The Theory of Games and Economic Behavior" by John von Neumann and Oskar Morgenstern in 1944. There was a fundamental breakthrough when von Neumann showed that, for a specific game, the two-person zero-sum game, a strategy was possible to maximise rewards. That strategy was based on each player maximising his or her minimum payoff.

In a zero-sum game, one person's loss is another person's gain, and all payoffs sum to zero for all possible strategies. Poker, for example, when played by two players is a zero-sum game. Add another player, or a banker who takes a cut of the pot, and it is no longer a two-person zero-sum game. The tendering process, even when undertaken with only two firms tendering, cannot be treated as a zero-sum game, since the loser must cover the cost of putting together the tender bid, whereas the winner will presumably profit from undertaking the job. These two sums are quite different.

A strategic game involves two or more players who act rationally to promote their own interests. Provided there is no cooperation among the players, it is a *non-cooperative* game. The outcome of the game depends on the decisions of the players, but chance might also play a part in determining the outcome, which involves some players "winning' and, usually, the other players "losing". A *zero-sum game* is one in which the sum of the winnings for all players is equal to the sum of their losses.

A simple one-stage decision with risk of outcome, such as the one represented by Table 12.1 in Section 12.2, can be thought of as a primitive example of a two-person game. Here, the opponent of the decision maker is Nature. In a "play", the decision maker chooses one of the available courses of action, and the response is one of two possible states of nature. The two moves together determine the outcome of the play, which will be a win or loss for the decision maker (or "player"). Of course, Nature here tends to be a disinterested player, rather than a competitive or malevolent one.

Two-person, zero-sum games

In a two-person, zero-sum game, players A and B each have a range of possible moves, a_1, .. a_i, .. a_m, and b_1, .. b_j, .. b_n. In a "play" A and B each choose a move, not knowing at the time what the opponent's move will be. Each possible pair of moves that make up a play, a_i and b_j, has an outcome O_{ij}, which by convention involves a payoff of value u_{ij} from B to A. A negative payoff means that B wins from A. The game can be represented by a payoff matrix that covers all possible outcomes.

By way of example, Table 12.3(a) contains the payoff matrix for a simple two-person, zero-sum game. In this case, A and B both have three possible moves. It will be clear from Table 12.3(a) that the game is unfair to B, but gaming situations are often not "fair". We wish to find out how to determine the best possible strategies for each player, assuming they do not cooperate and that they will make rational choices.

Table 12.3 Two-person, zero-sum game.

(a) Payoff matrix

	b_1	b_2	b_3
a_1	-1	+4	-3
a_2	+1	+2	+2
a_3	-2	+13	+1

(b) Strategy for A

Payoff matrix		B			$\min_j(u_{ij})$	$\max_i[\min_j(u_{ij})]$
		b_1	b_2	b_3		
	a_1	-1	+4	-3	-3	
A	a_2	+1	+2	+2	+1	+1 (i = 2; j = 1)
	a_3	-2	+13	+1	-2	

(c) Strategy for B

		B			
		b_1	b_2	b_3	
	a_1	-1	+4	-3	
A	a_2	+1	+2	+2	
	a_3	-2	+13	+1	
	$\max_i(u_{ij})$	+1	+13	+2	
$\min_j[\max_i(u_{ij})]$		+1			(j = 1; i = 2)

We first consider how A might reason. A's first thought could be to maximise the payoff, and hence choose action a_3 with the result +13. This strategy of maximising the possible outcome is expressed as:

$$\max_{ij}(u_{ij}) = u_{32} = +13$$

However, recognising that B is intelligent and is unlikely to choose action b_2, A will conclude that it is not feasible to achieve the maximum possible payoff. A reasonable alternative for A would be to determine the minimum payoff from the options a_1, a_2 and a_3, and then choose the move (here a_2) which maximises the minimum payoff. This is shown in Figure 12.3(b). The strategy of maximising the minimum payoff is expressed as:

$$\max_i\left[\min_j\left(u_{ij}\right)\right] = +1 \ \ (i = 2; j = 1)$$

From B's viewpoint, the initial approach might be to minimise the payoff to A, and choose action b_3 with the strategy

$$\min_{ij}\left(u_{ij}\right) = -3$$

But A is intelligent and will avoid choosing a_1, and B must be satisfied with less than the best possible outcome and seek an alternative approach. As with A, B is left with the strategy of identifying the maximum payoff from each move, b_1, b_2 and b_3, but minimising the maximum payoff, as in Figure 12.3(c). This gives

$$\min_j\left[\max_i\left(u_{ij}\right)\right] = +1 \ \ (j = 1; i = 2)$$

In this game, the outcome O_{21} is the most satisfactory result for both A and B. This is said to be a *saddle point*, for which

$$\min_j\left[\max_i\left(u_{ij}\right)\right] = \max_i\left[\min_j\left(u_{ij}\right)\right]$$

and the value of the game is, u_{21} = +1. Even if the game were to be played repeatedly, there would be no advantage for A or B to use moves other than those already identified. A sudden change in strategy would lead to a worse outcome for both players. The decision criteria are therefore:

For A: maximise the minimum payoff
For B: minimise the maximum payoff

Here, the criteria lead to a single best move for both players, which are, for A always to choose action a_2, and for B always to choose action b_1.

This first example was very simple. Not all games will have a saddle point. The payoff matrix for another game is given in Table 12.4 below. Here we obtain:

$$\max_i\left[\min_j\left(u_{ij}\right)\right] = +2 \ \ \ (i = 1, j = 1)$$
$$\min_j\left[\max_i\left(u_{ij}\right)\right] = +3 \ \ \ (j = 1, i = 2)$$

If this game is to be played just once, it is not clear which are the best moves for A and B. To maximise the minimum payoff, A would play a_1, while to minimise the maximum payoff, B would play b_1. However, in the expectation that B will play b_1, A might be tempted to play a_2, and hence increase the payoff to +3. But this would run the risk of the payoff falling to – 5, depending on what B chooses to do. Likewise, in trying to double-guess A, B might be tempted to play b_2, with a possible payoff of – 5 (a win for B), but also with the risk of the payoff rising to +5.

If the game is to be played many times, A can use a mixed strategy by playing a_2 occasionally to improve the overall payoff. B can also play b_2 occasionally to reduce the payoff. A mixed strategy can be found for each player, which involves playing each move a definite proportion of the time, but in a random and, from the opponent's viewpoint, unpredictable manner. The strategy for A is represented by probabilities p and (1 - p), which are the relative frequencies of choosing a_1 and a_2. For B, the strategy is q and (1 - q). It is possible to evaluate optimal mixed strategies for such games, based on the criterion for A of maximising the minimum expected payoff, and for B of minimising the maximum expected payoff.

Table 12.4 Payoff matrix for a problem without a saddle point.
Note: The strategies shown under the table are written as determinants: (1x2) for B, and (2x1) for A.

		B		$\min_j(u_{ij})$	$\max_i[\min_j(u_{ij})]$
		b_1	b_2		
A	a_1	+2	+5	+2	+2
	a_2	+3	-5	-5	
$\max_i(u_{ij})$		+3	+5		
$\min_j[\max_i(u_{ij})]$		+3			
Strategy for B		$[q, (1-q)]$		$0 \leq q \leq 1.0$	
Strategy for A		$\begin{bmatrix} p \\ (1-p) \end{bmatrix}$ $0 \leq p \leq 1.0$			

If there were contact between the two players, bluffing could be attempted by either or both, to mislead the opponent about upcoming plays. Cooperation could also be considered. This adds further dimensions to the game. Although we will not go into the details of evaluating strategies here, it should be noted that analyses can also be carried out for large $m \times n$ games and that the analyses have very close parallels with linear programming (Taha, 2017).

Other types of games

So far, we have only looked at zero-sum games, which are not applicable to many real engineering decision situations. Other forms of games have been studied and developments continue to be made in this relatively new branch of applied mathematics.

Extension of the two-person zero-sum game can be to a three-person game and to a multi-person game. With more players taking part in a game that is repeated, it becomes possible for two or more players to form a *coalition*, and act jointly against the other players. In a three-person game, A and B might join forces, temporarily, to win from C. A strategy for C is then to try to break the coalition and form a different coalition, say with A. This is possible because the best strategy for A will not necessarily be to remain in a coalition with B. If the coalition has been successful so far and C has incurred losses, there may well be a good opportunity for A for further winnings by breaking the alliance with B and taking C as an ally.

In non-zero-sum games, the losses are not necessarily equal to the gains in a play. In Table 12.5 the payoff table for a simple non-zero sum, two-person game is shown. The double entry for each outcome shows the payoffs to A and to B, respectively.

Table 12.5 Payoff table for a two-person game
Note convention for payoffs: (payoff to A, payoff to B).

		Player B		
		b_1	b_2	b_3
	a_1	(– 1, + 1)	(+ 4, – 4)	(+ 3, + 3)
Player A	a_2	(+ 1, + 1)	(+ 2, + 2)	(– 2, – 2)
	a_3	(– 2, + 2)	(+ 7, – 7)	(– 1, – 1)

Some of the outcomes in the table have zero sums. For example, if A plays a_1 and B plays b_1, the outcome is a loss of one unit by A and a gain of one unit by B. However, some outcomes result in wins for both players, with joint plays (a_1, b_3) and (a_2, b_2) being examples. These are win-win situations. On the other hand, (a_3, b_3) and (a_2, b_3) lead to losses for both players. This form of game can be difficult to deal with mathematically.

Further strategies

Rational strategies are available for various types of games, but we will consider just several. In some games, one course of action for a player will be better than all the others, no matter what the competitor chooses to do. This is called a *strictly dominant strategy*. If a course of action is equal to, or better than, any other, it is a *dominant strategy*. In game theory, a rational player will always choose a strictly dominant or dominant strategy, if one exists.

As an example, we consider the game with the payoff matrix given in Table 12.6. Although the double entry convention for non-zero-sum games has been used, this is a zero-sum game because the sum of the payoffs to the players happens to be zero for every entry.

Table 12.6 Payoff table for a two-person zero-sum game.

		Player B		
		b_1	b_2	b_3
	a_1	(– 1, + 1)	(+ 4, – 4)	(– 3, + 3)
Player A	a_2	(+ 1, – 1)	(+ 2, – 2)	(+ 2, – 2)
	a_3	(– 2, + 2)	(+ 9, – 9)	(+ 1, – 1)

Looking first at B's options, action b_2 will not be chosen by a rational player because it is *dominated* by both b_1 and b_3. No matter what Player A does, Player B will be worse off choosing b_2, with payoffs for B of $-4, -2$ or -13. By choosing b_1, Player B, would improve the payoffs to $+1, -1$ or $+2$. Option b_2 can therefore be removed from the game, and the payoff table is now as in Table 12.7.

Table 12.7 Payoff table with b_2 removed.

		Player B	
		b_1	b_3
	a_1	(– 1, + 1)	(– 3, + 3)
Player A	a_2	(+ 1, – 1)	(+ 2, – 2)
	a_3	(– 2, + 2)	(+ 1, – 1)

Turning now to Player A, it is evident that option a_1 is strictly dominated by a_2 and should be removed. This leaves the payoff table shown in Table 12.8. For Player B, option b_1 is seen to be strictly dominant: $(-1 > -2$ and $2 > -1)$. Knowing this, Player A will choose a_2. The solution for this game is therefore (a_2, b_1). The procedure of sequentially removing lesser options has led to the name *iterated dominant strategy*.

Table 12.8 Payoff table with b_2 and a_1 removed.

		Player B	
		b_1	b_3
	a_2	(+ 1, – 1)	(+ 2, – 2)
Player A	a_3	(– 2, + 2)	(+ 1, – 1)

Nash Equilibrium

The payoff matrix for another *non-cooperative, two-person game* is shown in Table 12.9, below. There are no dominant or dominated strategies for either player. This is an example of the type of game studied by John Nash, who shared the Nobel Prize for economics in 1994 (actually, the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel).

Table 12.9 Payoff table for a two-person game.

		Player B	
		b_1	b_2
	a_1	(+ 1, + 1)	(+ 2, 0)
Player A	a_2	(0, + 2)	(+ 3, + 3)

In studying such games, Nash realised that a player cannot consider his or her options in isolation. To take account of competition, each player needs to think through all the options in turn, including those of the opponent. A player might then be able to identify his or her best strategy for each possible play of the opponent.

In our example: If Player B(etty) assumes Player A(ndrew) will select $a_{1,}$ then she should choose b_1 (as this has the higher payoff); on the other hand, if A assumes B will play $b_{1,}$ then he should choose a_1. In this particular instance, there is agreement between the players, and the joint strategy (a_1, b_1) is denoted as a *Nash equilibrium* point. A similar analysis shows that there is a second equilibrium point at (a_2, b_2).

Prisoner's dilemma

A classic problem in the early theory of games is a situation described as the prisoner's dilemma. The scenario is that two long-time thieves have been caught near the scene of a burglary. They have been taken to the local police station and are being interviewed separately. Each is told that if he confesses to the pair's guilt and his partner does not, then he will be given protection and escape jail, while his partner will get 20 years in jail. If both confess, they have been told that they will be treated leniently and get 10 years of jail. The thieves know that if neither confesses, then the worst that is likely to happen is that they will get 1 year in jail because there is little evidence on which to convict them. What should each man do?

There is certainly an argument for not confessing because that gives a good outcome, but only if the other does not too. In developing a solution, the first task is to represent the range of outcomes as a payoff matrix, like the one in Table 12.8. Note that it is not a zero-sum game since the sum of each of the outcomes at each of the four possible strategies does not sum to zero. Indeed, none of them sum to zero. Table 12.10 presents all the options and the payoffs in terms of years in jail. In this case, therefore, the smaller the payoff the better the option.

Table 12.10 Payoff table for the prisoner's dilemma. Note: payoff is (Payoff to A, Payoff to B)

		Prisoner B	
		confess	silent
Prisoner A	confess	(10,10)	(0,20)
	silent	(20, 0)	(1, 1)

To determine the most likely outcome, it is necessary to investigate the various options. Consider first Prisoner A. He considers his options in the light of the behaviour of Prisoner B. If Prisoner B confesses, then A will either get 10 years if he also confesses or 20 years if he does not. Therefore, in that situation he should confess. On the other hand, if Prisoner B stays silent, A will either get 0 years if he confesses or 1 year if he stays silent too. In that case he should confess. Therefore, no matter what B does A should confess. Prisoner B can go through the same process and arrive at the same conclusion. Therefore, rational prisoners would both confess and get 10 years in jail. This is not the best outcome for them, but if they are acting rationally and are unable to confer, this is the action they would choose. There are a range of extensions to the sorts of games that can be contemplated. The concept of cooperative and non-cooperative games adds another level of complexity. For the treatment here, cooperative games will not be considered.

The argument used to come to a final rational decision with the prisoner's dilemma can be formalised by the definition of a number of strategies: of dominant, strictly dominant, dominated and strictly dominated, which are described briefly in the main text. These allow a formal approach to be taken to the problem of making decisions with competition.

In general, there may not be a Nash equilibrium point in a game, alternatively there may be one or more. Strategically, it is not in any player's interest to unilaterally change away from a Nash equilibrium point, and so strategies for non-cooperative games can be developed from the Nash equilibrium concept. Since Nash's work, there have been significant further developments in game theory.

In this brief review, we have only discussed simple, basic concepts. From a practical viewpoint, it is also possible to find near-optimal strategies by brute-force computations using search, trial-and-error, and simulation, as will be explained in Chapter 13.

12.7 MULTIPLE-CRITERIA DECISIONS WITH MANY PARTICIPANTS

The decision models so far considered have used a single decision criterion, which was either dollar payoff or utility. Many real engineering decision problems are more complex, in that there are many separate criteria to be considered simultaneously. Personal decision problems can also be of this type. Consider for example the purchase of an automobile. The prime consideration might be initial cost, but there are other relevant matters to consider, including the ongoing costs for service and repair, the safety and performance of the vehicle on the road, fuel economy, emissions levels, driver and passenger comfort, electronic aids (radio, music player, navigation aids and parking assistance), internal and external styling and possibly even colour.

Also, multi-criteria problems can be classified according to the amount of information available to the decision maker, and so the range of possibilities includes certainty, risk, uncertainty and ignorance of outcome. These may be in one-off or many-off situations, and in competitive, game-like situations.

A further complication occurs if the participants in the decision process are numerous. Large engineering projects involve decisions that affect the lives of many people and organisations. Some at least of those affected will consider themselves to be stakeholders and will want to take part in the decision-making.

Multi-criteria and multi-participant decision problems arise in large engineering, public-benefit projects. In the management of water resources, some of the important criteria include maximising flood prevention (or at least providing flood control), maintaining a continuous supply of water for irrigation and household use, and accommodating water sports and other recreational activities.

It is rarely possible to find optimal solutions to such complex decision problems. However, a pragmatic approach is to find a "good" result, rather than the best. Researchers have studied multi-criteria and multi-participant decision problems since (at least) the 1970s, and various approaches have been proposed. We will shortly consider several that have been used successfully to effect community decision making. It needs to be said, however, that multiple-participant problems can in some circumstances become very difficult to deal with, particularly when the participants have differing and strongly held, entrenched views and agendas. The problem can then become "wicked", and in extreme circumstances, a "mess", as discussed previously in Chapter 3. If it is not possible to find an engineering solution, then a political solution can be attempted.

Satisficing with multiple criteria

In some situations, one decision criterion will be more important than the others. A simple and practical approach is then to use the main decision criterion to compare and rank the alternatives, but to set minimum acceptable levels for the other criteria. This approach has been called *satisficing*. Any contemplated course of action is then eliminated or modified if it does not meet all of the minimum acceptance levels. The remaining feasible actions can then be evaluated and ranked according to the main criterion.

If a new computer facility is to be chosen by design consultants, there will be numerous criteria to consider. Initial cost will be important, but ongoing costs, reliability, availability and compatibility of software, expertise and practical help in the event of system failure, average downtime for the different units such as printers and work stations, response time for callouts, are all important factors. Information on these aspects should be obtainable from the distributors of the competing systems in the market, and it would then be possible to set realistic minimum requirements for each criterion. A satisficing approach can now be used, with present dollar value as the prime criterion, but with minimum acceptance standards for other factors, such as:

service time to be within three hours of call during office hours for any piece of equipment; and

total downtime to be not more than five percent.

Traditional engineering design procedures are often based on what is, in effect, a satisficing approach. In routine structural design, various criteria must be taken into account when, say, reinforced concrete beams and slabs are proportioned. The design criteria include adequate serviceability, adequate safety against failure, ease of construction, robustness, initial cost and ongoing maintenance costs. National design standards impose strict legal minimum requirements on many of the structural factors related to these criteria, for example by placing upper allowable limits on deflections and crack widths to ensure serviceability, and minimum acceptable strength levels for different possible modes of failure to ensure safety. With such limits, the designer can optimise the design by minimising cost over the full life cycle, while confirming that all relevant limits on structural behaviour are met.

An advantage of the satisficing approach, if it can be used, is that it is simple to apply. On the other hand, it will not necessarily deliver the best possible solution. In fact, the method avoids addressing the question of what constitutes the best solution. Nevertheless, it can deliver a range of good, workable solutions for relatively complex decision problems.

Weighting and scoring multiple criteria

As an alternative to the satisficing approach, a weighting-and-scoring technique might be used. Each course of action, a_i, is studied in turn, together with the possible outcomes associated with it. A score, s_{ij}, (for example out of 10) is allocated for each criterion, c_j. Weights, w_j, are also given to each criterion to reflect its relative importance. An overall score can then be calculated for each possible course of action, taking all criteria and possible outcomes into account. This allows a ranking to be made. With n criteria to be considered, the score for the i^{th} course of action is

$$S_i = \sum_{j=1}^{n} w_j s_{ij} \qquad (12.4)$$

This technique is suitable for evaluating and ranking bids for tenders. An example of a bid evaluation is summarised in Table 12.11.

Table 12.11 Evaluation of competing bids for a commercial tender.

	Weight	Company A		Company B		Company C		Company D	
Criteria	%	score 0-10	weight x score	score 0-10	weight x score	score 0-10	weight x score	score 0-10	weight x score
Fee	30	10	30	7	21	8	24	5	15
Methodology	20	9	18	6	12	1	20	10	20
Resources	25	8	20	1	25	8	20	8	20
Consultants	15	9	13.5	8	12	6	9	9	13.5
Experience	5	8	4	8	4	8	4	8	4
Management	5	10	5	1	5	6	3	10	5
TOTAL	**100**		**91**		**79**		**80**		**78**
Rank			**1**		**3**		**2**		**4**

The tender was for the architectural design and documentation of a university building. The tenderers had been advised of the criteria to be used for evaluation, which were: fixed lump *fee*; design *methodology*; *resources* available to undertake the work; availability of *consultants*; previous project *experience* and educational qualifications; and previous experience in the *management* of building contractors.

The evaluations of the submissions from four companies, listed as A, B, C and D, are presented in Table 12.11, using two columns for each company. The first column shows the raw scores given for each criterion; the second column gives the weighted scores. The total sums of the weighted scores are entered in the second to last row. The rankings of the companies, based on the weighted scores, are in the last row of the table. The weightings given to each of the criteria are shown in the second column in the table. The range of scores were: 0: not submitted; 1 – 2: poor; 3: reasonable; 4 – 5: good; 6 – 7: above average; 8 – 9: excellent; 10: exceeds request.

The weighting and scoring technique has various advantages. Firstly, it is useful to have a method that gives a quantitative answer when the alternatives are compared. Secondly, in the planning stage it forces the user to consider the performance of the options and this assists in that process. Thirdly, the method can be applied at different stages of the design and planning processes.

However, Hobbs (1980) has suggested that a proper set of weights must have certain characteristics if it is to take proper account of trade-offs among the criteria. The weights need to be proportional to the *relative* value of unit changes in their objective functions. This means that weights should not be given simple ordinal values such as 1 for low importance, 2 for average importance and 3 for high importance. This does not allow for trade-offs among the criteria. For example, if $w_1 = 2$ and $w_2 = 4$, then an increase in the score of objective 1 by 1 unit should be viewed as identical in overall benefit to a change in value of objective 2 by 0.5 units. The weights are much more about relative trade-off values than importance.

The process of determining weights may take some time but as Hobbs notes: “As a typical siting study [for power plants in the U.S.] costs tens of millions of dollars or more, it seems reasonable to require that an extra afternoon be taken to ensure that weights are theoretically valid."

Power supply for Baden-Wűrttemberg

Renn (2003) shows the application of a weighting and scoring technique in a survey that was carried out in Germany to provide power supply planning data for the medium to long term. Four scenarios were developed:

1. Utilization of technologies (nuclear power);
2. Protection of resources (sustainable, conservation);
3. New lifestyle (green, change in lifestyle); and
4. Business as usual.

Two groups of people (a group of engineers and a group of church representatives) were then surveyed to come up with a point value (weighting) for a total of 74 criteria and an expected utility (EU). The final value of each scenario was summed as the product:

$$EU = \sum p_i u_i$$

where p_i is the point value and u_i the utility of each of the n components.

The criteria covered, among other things, economic aspects (e.g., technical efficiency, security of supply, profitability, macro-economic effects), protection of environment and health (e.g. using environment as a waste sink, conservation of nature and ecosystems, health risks) and social and political aspects (e.g., protecting dignity of humans, political stability, social justice). Each criterion was allocated a score from 0 to 100 points based on the groups assessment of its likely outcome. Unfortunately, few details were available on the system used to allocate the weights, although it was noted that the final decision was sensitive to these. The final scores showed both the engineers and church representatives preferred Scenario 3 (new lifestyle) over Scenario 2 (protection of resources).

The results are interesting for a number of reasons. The engineers were actually surprised to find that their deliberations had led to them selecting Scenario 3. This was not what they had expected and the author suggests they had probably set some of the weights too low. Also the scoring of the various options indicated that people often employ a highly non-linear scale. One of the options, for example, had energy savings of 40 percent yet led to a much smaller difference (20 percent) in the scores given to that scenario. It was suggested that this may have been partly due to people not believing the efficiency figures and discounting the benefits to a degree.

The choice of weights usually involves subjective judgements about the relative importance of the criteria. As Martland (2012) has cautioned: .. "avoid thinking that it will be easy to agree on criteria or weights or to think that the public will agree with whatever criteria or weights are proposed by any of the parties". As with other complex engineering decision problems, even the weighting and scoring procedure may in practice require political input.

Multi-objective decisions with certainty and uncertainty of outcome

Decision making becomes problematic if some of the decision criteria are not quantifiable. In such circumstances, Hardy (1995) has suggested the use of fuzzy logic. In an example, a site is to be chosen for storing hazardous waste material. Some of the decision criteria are quantifiable (such as transportation costs, surface water quality and ground water quality), but others are qualitative and involve linguistic descriptions such as “good” or “poor” (as are used in aesthetic evaluations).

He suggests that fuzzy logic concepts can be employed to treat both quantitative and non-quantitative criteria, and discusses two problems: the selection of the most appropriate propellant fuel for use in a lunar lander module, and the choice of an aeration system to improve water quality in a river. He also refers to *decision software* that can be used in such engineering decision making.

Situations where there is complete knowledge of the system are of course much easier to deal with, and in this case it is realistic to search for optimal solutions. Such problems are discussed in some detail in Chapter 13 in relation to optimisation.

Tendering to repair a dam wall

Tabatabaei and Zoppou (2000) report on an engineering project where faulty construction work on a concrete dam wall was discovered during routine maintenance work. The concrete poured in 1912 near the base of the dam in the Australian Capital Territory was found to be of very poor quality – a situation with no easy or obvious solution, given the dam’s location and role as a water storage facility for Canberra, the national capital. Tenders were called and this led to three submissions, each proposing a quite different solution, particularly in relation to handling possible high flow events that might occur while the repairs were being carried out.

The tenders were assessed using a series of objectives (the total cost, the flood risk, the demolition technique and relevant experience) which were weighted in the ratios of 100:50:30:10, respectively. Each of the tenders was given a score for each objective and the product of the score and weight summed. This led to the decision to award the tender to Tender T1.

The decision on whether to require a coffer dam during repair was based on a decision tree analysis where the cost of the coffer dam was compared to the expected cost of having no coffer dam based on the assumption of a 70 percent chance of a rain (potentially leading to a flood) occurring during the repairs, and where the probable number of rain days was determined using a statistical analysis that summed over all possible occurrences.

Decisions that involve multiple criteria and multiple stakeholders

Important engineering decisions usually affect the lives of people in the community. In the field of water resource management, for example, decisions on flood control are vitally important to those people who live in regions that are prone to flooding.

Likewise, all people living in a city will be seriously affected by an inadequate water supply. In such circumstances, the entire population may become interested observers of, and possibly stakeholders in, the engineering decisions. Community input is needed in engineering decision such as these, and it is not unusual for significant disagreement to exist among individuals, even on underlying principles. When such differences cannot be resolved, the decisions become political and a political solution might become necessary.

On a much smaller scale, it is still usual for decision problems with multiple criteria to involve more than one stakeholder. Keller et al. (2009) have explained how even in personal decisions there can be many stakeholders, each with different objectives and criteria. One of the cases they describe concerns the choice of treatment for a man diagnosed with prostate cancer.

Akter and Simonovic (2002) have dealt with a wide range of multiple-criteria, multiple-participant decision problems. They discuss practical problems, such as decision making for flood control, that involve both multiple criteria and multiple participants. They consider deterministic decisions (with certainty of outcome) and decisions with uncertainty of outcome. The methods they suggest for dealing with these different cases include structuring and measuring the alternative courses of action, applying utility theory, and prior and progressive articulation of preferences.

12.8 USING ADDITIONAL INFORMATION

In previous discussions we have classified decision problems according to the amount of information that is available to the decision maker.

The acquisition of additional information is a feasible option to consider in most decision situations. Although additional information will have a cost, it may make it possible to move up one or two rungs on the information scale, for example from decision making under uncertainty to decision making under risk. Even if the decision situation is one of risk, additional information could benefit the decision process by setting some of the probability estimates to unity and zero, thus reducing the range of possibilities. In complex situations, where there are multiple objectives and multiple stakeholders, acquiring additional information can lead to clearer thinking and a clearer problem statement, with better decisions and outcomes.

Additional information for engineering decisions can be acquired in various ways, with the choice depending on the specific field in which the decision is to be made. Options include laboratory testing, pilot studies, theoretical analyses, field trials, and engaging an expert or a specialist research agency. In a study of the tendering behaviour of construction companies Mochtar and Arditi (2001) found that around 50 percent of firms purchased information (sometimes, often or always) from research agencies to strengthen their bids, even though the authors considered that this was not the preferred method, given its cost.

For the specific case of decisions that involve multiple stakeholders with differing viewpoints, role-playing has been proposed as a way to improve outcomes (Keller et al., 2009).

In decision situations that lend themselves to modelling, it may be possible to carry out sensitivity analyses using possible and likely ranges of values, or probabilities if available, for the main variables. In the work of Mankuta et al. (2003) probabilities were reported as a base value, but low and high options were also given. In this way, it was possible to investigate how sensitive the decisions were to the actual probabilities and the results of decisions could be viewed in this light. The idea of using a range of values is one of the features of fuzzy logic and fuzzy sets, where, for example, variables are given a range of values in the form of a triangular distribution (Chan et al., 2000). Computer modelling can also be combined with brute-force computations using search techniques, trial-and-error search, simulation and Monte Carlo simulation (Warner and Kabaila, 1968).

12.9 SUMMARY

Decision making, like problem solving, is central to almost all aspects of engineering work. We commenced this chapter with a review of the types of decisions that are studied in decision theory and game theory. These often concern decisions where the outcomes are measured in monetary terms. However, the concept of utility allows risk aversion and risk acceptance to be taken into account in decision making.

In decision situations that involve competition, rational approaches can be found, although the logic is rather different to that in simple decision theory. To take account of competition, the decision maker needs to think through all the options of *all* the participants, including potential friends and foes, and not just his or her own.

While the problems that can be dealt with theoretically are idealised, and often not directly applicable to engineering decisions, they provide a useful insight into decision making processes and into how to structure decision problems.

Most engineering decision problems are quite complex because there is not a single decision criterion to consider, but many. A further complication is that many persons and organisations will be affected by the outcomes of a large engineering project. The decisions then involve many participants and stakeholders and it is not usually feasible to find optimal solutions. A pragmatic approach is to look at a range of practical and realistic solutions and to choose the best. Some techniques that have been used in complex engineering decision problems have been reviewed.

The theoretical approaches discussed in this chapter provide a useful insight into decision-making processes and into how to structure decision problems. This is important because engineers have to be able to articulate and explain the rationale for their decisions.

PROBLEMS

12.1 A computer manufacturer has the possibility of expanding the company's operations into New Zealand with the potential of greatly increased profits. The managing director considers that the move would lead to a 20 percent chance of making the company $10m in profits, but that there would be an 80 percent chance of losses of $2m if the move were unsuccessful. If the expansion is not attempted, the company would continue to trade with an 80 percent chance of making a profit of $2m and a 20 percent chance of breaking even. Based on expected monetary value, what would you suggest? Under what circumstances would it *not* be appropriate to use this criterion?

12.2 A government department associated with the Cooperative Research Centre for Low Grade Coal has come up with a new process to pre-treat raw coal and has called for tenders to implement this in a new factory. As Research and Development Manager of a consulting firm, you are familiar with the theory behind the process, but you are also aware of new developments overseas that may mean the technology will be out of date in the near future. The three courses of action available are

(a) do not tender;
(b) prepare and submit a tender, based on the designated process;
(c) work on the new overseas development and prepare a tender based on this new method.

You estimate that, if you get the contract for the CRC process there is a 70 percent chance of making a $600,000 profit, a 15 percent chance of breaking even, and a 15 percent chance of making a loss of $3m. The cost of preparing the tender has been taken into account in the above profit-and-loss and break-even estimates. The chances of winning the contract for the CRC process are good: about 70 percent.

Although the chances of winning the contract on the basis of the new overseas developments are much lower, about 40 percent, you stand to make a much larger profit. You estimate that, if you get the contract, you would have a 70 percent chance of making a profit of $15m. There is a 30 percent chance of losing $3m. The costs of preparing the tender and bringing the overseas method to the tendering stage have been included in the profit and loss estimates given.

Make any calculations you consider appropriate, state any assumptions you decide to make and hence indicate your decision. What is the expected value of your decision?

12.3 Assuming there is a 1 in a 600,000* chance of winning $10,000 in a $1 lottery, derive the expected value of the return of buying a ticket. (*This figure came from a television advertisement running in South Australia in April 2005.)

12.4 Re-analyse Question 12.1 using the utility information given in Table 12.12 which you have obtained from the managing director about her feeling for the utility value of the company's profits and losses. Does this change the earlier decision? Is the managing director showing risk aversion or risk acceptance in this case?

Table 12.12 Utility information for the manager.

Profit/Loss	Utility
+$10m	100
+$5m	30
break even	0
-$5m	-20
-$10m	-40

12.5 In the design of a pulp mill, annual losses due to interruption of the electricity supply may be one of the following values depending on the time that the interruption lasts: $0, $2m, $6m. The losses in any one year will depend on the severity of the interruptions which actually occur.

Table 12.13 Probabilities of annual losses.

Action	Annual Cost of Action ($)	Probability of Incurring Annual Loss		
		$0	$2m	$6m
A	0	0.60	0.25	0.15
B	200,000	0.75	0.20	0.05
C	650,000	0.90	0.09	0.01

The three alternative actions being investigated have annual costs associated with them as follows:

A: do nothing: $0;
B: provide electricity from diesel-powered generators with manual operation for critical areas of the mill when power blackout occurs: $200,000;
C: provide diesel-powered generators for all mill operations with automatic cut in control system: $650,000.

The probabilities of the annual losses will depend on the action taken by the design engineer. Estimated values of the probabilities are given in Table 12.13. What is the best course of action, and why?

12.6 The payoff matrix for a two-person game is contained in Table 12.14. As player A, what strategy would you choose if (a) the game is to be played only once, and (b) it is to be played many times?

Table 12.14 Payoff table for a two-person game.

		Player B		
		b_1	b_2	b_3
Player A	a_1	(– 1, + 1)	(+ 4, – 4)	(+ 3, + 3)
	a_2	(+ 1, + 1)	(+ 2, + 2)	(– 2, – 2)
	a_3	(– 2, + 2)	(+ 7, – 7)	(– 1, – 1)

Table 12.15 Payoff matrix.

		B		
		short	long	Mid
A	short	(2,-2)	(2,-2)	(2.5,-2.5)
	long	(1,-1)	(3,-3)	(2.5,-2.5)
	mid	(2,-2)	(2,-2)	(2.5,-2.5)

12.7 In the situation represented by the payoff matrix given in Table 12.15 above, determine for each player: (a) strictly dominant strategies, (b) strictly dominated strategies, (c) weakly dominant strategies and (d) weakly dominated strategies. Find also any Nash equilibrium points.

12.8 Assume it had not been possible to estimate the probability of the various outcomes occurring in Question 12.1. What decision would be made if the Laplace Criterion were used? What decision would be made if the Pessimistic Criterion were used? Base any calculations on monetary values.

12.9 Create your own personal utility functions that take account of:
(a) your aversion to risk (for example to evaluate your attitude to insurance); and
(b) your acceptance of risk when large monetary gains are possible but unlikely, and small losses are very likely. Choose the ranges of monetary values to suit your current personal situation.

12.10 You are working on a feasibility study where it is necessary to assess five alternative projects. Your company has worked with stakeholders to weight 6 performance criteria to assess how well each project alternative meets the criteria. It is now your responsibility to rank the different alternatives. Using the weighted score technique, and the information provided in Table 12.16, which alternative would you recommend?

Table 12.16 Performance criteria weights and assessment scores.

Criteria:	1	2	3	4	5	6
Weighting:	0.1	0.4	0.2	0.1	0.1	0.1
Project A	3	5	3	1	3	2
Project B	3	3	3	2	1	1
Project C	2	3	3	5	2	3
Project D	1	4	3	4	3	4
Project E	4	1	3	3	1	5

REFERENCES

Akter, T. and Simonovic, S. P., 2002, *A general overview of multi-objective multiple-participant decision making for flood management*, Report No. 041, Department of Civil and Environmental Engineering, The University of Western Ontario.

Bazerman, M., 1998, *Judgement in Managerial Decision-making*. (John Wiley & Sons), 200pp.

Bennett, D., 2001, Development of Sustainability Concepts in Australia. In: *Towards Sustainability. Emerging Systems for Informing Sustainable Development* (Eds. Venning, J. and Higgins, J.) (University of New South Wales Press), 22-47.

Bierman, H.S. and Fernandez, L., 1998, *Game Theory with Economic Applications*, 2nd Edition, (Addison-Wesley), 452pp.

Chan, F.T.S., Chan, M.H. and Tang, N.K.H., 2000, Evaluation Methodologies for Technology Selection. *Journal of Materials Processing Technology*, 107, 330-337.

Dörner, D., 1989, *The Logic of Failure*. (Basic Books), 222pp.

Dreizin, Y., 2006, Ashkelon Seawater Desalination Project – Off-Taker's Self Costs, Supplied Water Costs, Total Costs and Benefits. *Desalination*, 190, 104-116.

Feynman, R., 1999, The Best Works of Richard P. Feynman. *The Pleasure of Finding Things Out.* (Ed. J. Robbins), (Penguin), 270pp.

Greening, L.A. and Bernow, S., 2004, Design of Coordinated Energy and Environmental Policies: Use of Multi-Criteria Decision-making. *Energy Policy*, 32, 721-735.

Hardy, T. L., 1995, Multi-objective decision-making under uncertainty: Fuzzy logic methods, NASA Technical Memorandum 106796.

Hobbs, B.F., 1980, A Comparison of Weighting Methods in Power Plant Siting. *Decision Sciences*, 11, 725-737.

Holland, J.H., 1998, *Emergence from Chaos to Order*. (Oxford University Press), 258pp.

Keller, L.R., Simon, J. And Wang, Y., 2009, Multiple-Objective Decision Analysis Involving Multiple Stake Goldens, *Tutorial in Operations Research, Informs 2009, http://journal.informs.org/.*

Kmietowicz, A.W. and Pearman, A.D., 1981, *Decision Theory and Incomplete Knowledge*, (Gower), 121pp.

Knutson, B. and Peterson, R., 2005, Neurally Reconstructing Expected Utility. *Games and Economic Behaviour* (in press).

Manhart, K., 2005, Lust for Danger. *Scientific American Mind*, 16(3), 24-31.

Martland, C. D., 2012, Toward more Sustainable Infrastructure: Project Evaluation for Planners and Engineers, John Wiley & Sons Inc.

Mankuta, D.D., Leshno, M.M., Menasche, M.M. and Brezis, M.M., 2003, Vaginal Birth after Cesarean Section: Trial of Labor or Repeat Cesarean Section? A Decision Analysis. *American Journal of Obstetrics and Gynaecology*, 189(3), 714-719.

Mochtar, K. and Arditi, D., 2001, Role of Marketing Intelligence in Making Pricing Policy in Construction. *Journal of Management in Engineering*, 17(3), 140-148.

Neumann, John von; Morgenstern, Oskar, 1953, [1944]. *Theory of Games and Economic Behavior (*Third ed.). Princeton, NJ: Princeton University Press.

Paulos, J.A., 1988, *Innumeracy*. (Penguin), 135pp.

Peterson, I., 2002, Tricky Dice Revisited. *Science News Online*, 161(15) http://www.sciencenews.org/articles/20020413/mathtrek.asp (downloaded 21st June, 2006)

Renn, O., 2003, Social Assessment of Waste Energy Utilization Scenarios. *Energy*, 28, 1345-1357.

ReVelle, C.S., Whitlatch, E.E. Jnr. and Wright, F.R., 2004, *Civil and Environmental Systems Engineering*, (Pearson Prentice Hall), 552pp.

Rorie, A.E. and Newsome, W.T., 2005, A General Mechanism for Decision-making in the Human Brain? *Trends in Cognitive Science*, 9(2), 41-43.

Sarvary, M. and Parker, P.M., 1997, Marketing information: A competitive analysis. *Marketing Science*, 16 (1): 24-38.

Schefter, J., 2000, The Race. The Definitive Story of America's Battle to Beat Russia to the Moon, (Arrow Books), 303pp.

Schwartz, J.M. and Begley, S., 2002, *The Mind and The Brain*. (Regan Books), 420pp.

Shiv, B., Loewenstein, G. and Bechara, A., 2005(a), The Dark Side of Emotion in Decision-making: When Individuals with Decreased Emotional Reactions Make More Advantageous Decisions. *Cognitive Brain Research*, 23, 85-92.

Shiv, B., Loewenstein, G., Bechara, A., Damasio, H. and Damasio, A.R., 2005(b), Investment Behaviour and the Negative Side of Emotion. *Psychological Science*, 16(6), 435-439.

Squyres, S., 2005, *Roving Mars*. (Hyperion), 422pp.

Tabatabaei, J. and Zoppou, C., 2000, Remedial Works on Cotter Dam and the Risk to These from Floods. *ANCOLD Bulletin*, 114, 97-111.

Taha, H. A., 2017, *Operations Research: An Introduction,* 10th Edition, (Upper Saddle River, NJ, Pearson Education).

Warner, R. F. and Kabaila, A. P., 1968, Monte Carlo study of structural safety, *Journal of the Structural Division, ASCE*, (*cedb.**asce**.org/CEDBsearch*).

Winman, A.; Hansson, P. and Juslin, P., 2004, Subjective Probability Intervals: How to Reduce Overconfidence by Interval Evaluation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30(6), 1167-1175.

CHAPTER THIRTEEN

Optimisation

This chapter introduces general concepts of optimisation. Optimisation is a process aimed at identifying the "best" solution to a planning or design problem. Optimisation may be carried out using judgement, by comparing a large number of options or by using analytical techniques. Analytical optimisation techniques may be applied to problems that can be represented by one (or more) objective functions and a number of constraints. Specific optimisation techniques including linear programming, differential calculus, separable programming, dynamic programming and genetic algorithms are considered in detail in this chapter.

13.1 INTRODUCTION

As outlined in Chapter three, engineering problems are often ill-defined and open-ended. As such, there is usually no single "correct" answer to an engineering problem. The challenge for the engineer is to find the "best" solution to a particular problem.

The engineering methodology outlined in Chapter 3 emphasizes the identification of objectives to be achieved by a particular plan or design and measures of effectiveness for ranking alternative solutions. Furthermore, most engineering problems have a number of constraints that limit the available solutions to the problem.

Optimisation is a process of identifying the "best" solution to a problem i.e. that solution that achieves the maximum (or minimum) level of the objectives while satisfying all of the constraints. Usually there are several objectives to be achieved. The choice of a final solution may involve a compromise between these. For example, a reservoir for a town water supply is to be designed to achieve the following objectives:

- maximise reliability of the water supply;
- minimise cost; and
- minimise environmental impact.

Clearly, no single size of reservoir will satisfy all three of these objectives. A large reservoir would be needed to maximise reliability. This is unlikely to have the minimum cost or the minimum environmental impact. The reservoir that has minimum cost is small in size (or, in fact, no reservoir at all). This will provide a water supply with low reliability, although the environmental impact could be low. Ultimately, the selection of a final design involves compromise between the various objectives. This usually requires the input of community values and a judgement to be made about what is in the best interest of the community as a whole.

Optimisation involves the selection of values for a set of *decision variables*. In the above example, the primary decision variables are the location and capacity of the reservoir. In general, the decision variables may be *continuous* or *discrete*. The capacity of a reservoir can take on any positive value within a defined range and so is a continuous decision variable. The selection of the location for the reservoir is an example of a discrete decision variable, as there will be a finite set of suitable locations available and only one of them will be chosen.

A *solution* to an optimisation problem is any combination of values of the decision variables. Any set of values of the decision variables that satisfies all of the constraints is called a *feasible solution*. For a single objective maximisation problem, any feasible solution that produces a value of the objective function not less than that produced by any other feasible solution is called an *optimum solution*. Similarly for a single objective minimisation problem, any feasible solution that produces a value of the objective function not more than that produced by any other feasible solution is called an *optimum solution*. The concept of Pareto optimal solutions for multi-objective optimisation problems is discussed in Section 9.5.

The following is an example of a simple optimisation problem with continuous variables that can be solved using calculus.

Example 13.1: A tank of volume 200 m^3 is to be made from stainless steel. It will have a cylindrical body and hemispherical ends as shown in Figure 13.1. If the material for the ends costs twice as much per square metre as that for the body, find the dimensions of the tank that minimise its cost.

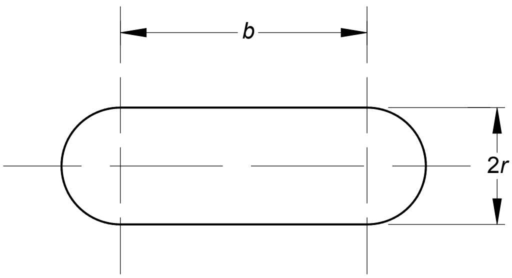


Figure 13.1 A stainless steel tank.

Solution: In this example, the decision variables are the length of the body of the tank and its radius (both in metres). These will be denoted by b and r (respectively). The objective is to minimise the cost of the design. If the cost per unit area of the body is C \$/m^2, then the total cost of the tank is given by:

$$Y = 2C\pi rb + 2C(4\pi r^2)$$

$$Y = 2C\pi rb + 8C\pi r^2 \tag{13.1}$$

Note that the first term in Equation (13.1) represents the cost of the tank walls and the second term represents the cost of the ends. As C is constant, it can be ignored in the optimisation and the problem becomes:

$$\text{Minimise} \quad Z = 2\pi rb + 8\pi r^2 \qquad (13.2)$$

Having identified the objective function, we need to consider what constraints exist. In this case, we require the volume of the tank to equal 200 m^3. This can be expressed by the following equation:

$$\pi r^2 b + 4\pi r^3/3 = 200 \qquad (13.3)$$

The first term on the left hand side represents the volume of the cylindrical body of the tank and the second term represents the volume of its hemispherical ends. Further constraints are required to represent the physical reality that both dimensions of the tank must be non-negative, i.e.

$$b \geq 0 \qquad (13.4)$$
$$r \geq 0 \qquad (13.5)$$

Expressions (13.2) through (13.5) represent a simple optimisation model that can be solved to find the dimensions of the tank that will minimise cost while satisfying the relevant constraints. The solution of this particular problem is discussed in Sections 13.11 and 13.12 of this chapter.

The above example involved only two decision variables and one objective function. Real engineering optimisation problems may have hundreds or even thousands of decision variables, several objectives and hundreds of constraints. A large number of techniques have been developed for the solution of engineering optimisation problems. Several of these techniques will be discussed in this chapter including linear programming, separable programming, dynamic programming and genetic algorithms. The optimisation techniques considered in this chapter deal only with problems that have a single objective function. Interested readers are referred to Goicoechea *et al.* (1982) for a discussion of optimisation techniques for multiple objectives.

13.2 APPROACHES TO OPTIMISATION

In general the following three approaches to optimisation can be identified (Meredith et al., 1985):

- subjective optimisation;
- combinatorial optimisation; and
- analytical optimisation.

These are described in more detail below.

Subjective Optimisation

The simplest type of optimisation is based on engineering experience and judgement. This often takes the form of comparing a number of alternatives and selecting the best. In these circumstances, it may be claimed that the design has been "optimised". In other cases, "rules of thumb" based on experience are used to identify an efficient design. For example, for the design of water supply pipelines a number of water authorities try to choose pipe sizes such that the pressure loss per kilometre of length falls within defined bounds. A typical value is 3m head loss per kilometre length of pipe.

Subjective optimisation is used in cases where the objective(s) and/or the constraints cannot be formulated in mathematical terms. In other cases, there may be so many constraints that the principal task is to identify a design that satisfies all of them (i.e. a feasible design). In any case, the application of judgement by experienced engineers is an essential part of the design process and should not be undervalued.

Combinatorial Optimisation

A number of engineering planning and design problems involve the comparison of a finite number of discrete alternatives relative to a set of objectives (or selection criteria). For problems of this type, combinatorial optimisation techniques are appropriate. The simplest combinatorial technique is called complete enumeration and involves the evaluation of all of the alternatives according to the criteria and the selection of one alternative using appropriate multi criteria techniques (as discussed in Section 12.7). This approach may be used where the number of combinations is relatively small.

An example of where complete enumeration of alternatives has been applied successfully is Emerson Crossing in adelaide, Australia. Emerson Crossing is an intersection of two roads in the city of Adelaide that has the unusual feature of having a dual railway line crossing diagonally through it (Figure 13.2).

This came about from a desire by the early transportation engineers to reduce the number of level crossings where cars and trains were in potential conflict. It made excellent sense when traffic volumes were quite low, but by the 1960s the traffic volumes on Cross Road and South Road had built up to such levels that the delays at this intersection at peak hour were extremely long. This was aggravated by the fact that the intersection was closed to traffic in both directions for 15 minutes of each peak hour due to the passage of trains. A study of ways to alleviate the congestion was carried out by the South Australian Highways Department (the responsible authority at the time). A large number of options were considered to reduce the traffic delays. One class of these options involved the use of grade separation at the intersection to provide, spatial separation of road and railway traffic. Other options considered included diverting traffic away from the intersection, the use of roundabouts, and moving the railway line so that it had separate intersections with South Road and Cross Road. For various reasons, the grade separation options were considered to be better than the other options. In fact, a number of grade separation options were available. These included the following:

- South Road on an overpass or underpass;
- Cross Road on an overpass or underpass; and
- the railway on an overpass or underpass.

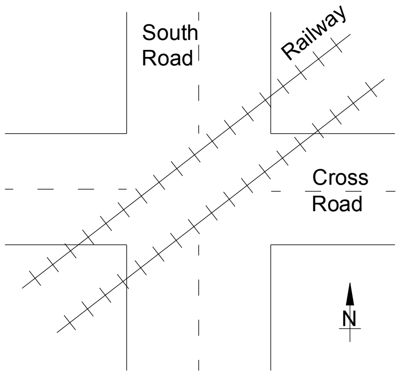


Figure 13.2 Layout of Emerson Crossing before the construction of the overpass (Not to scale)

These six basic solutions can be combined with each other (e.g. South Road on an underpass and the railway on an overpass) to give an additional 6 combinations. To these can be added the "do nothing" alternative to give a total of 13 alternatives. This is a small enough number for the application of complete enumeration to be carried out.

All 13 alternatives were considered in a feasibility study (Section 3.5) in order to eliminate the less competitive options. Using appropriate mathematical models the options were compared in terms of construction and operational costs, reduction in delays to vehicular traffic, likely reduction in accidents and their environmental and social impacts. Eventually, the option of putting South Road on an overpass was identified as the best option. This was implemented in 1982, and has proven to be quite successful.

In many real engineering problems, the number of options can be extremely large. This is particularly the case in systems where discrete sizes need to be chosen for a large number of elements. For example, the design of engineering networks such as road networks, water supply networks, electricity networks and communication networks require sizes to be chosen for a large number of elements such as roads, pipes, wires or communication channels. As each element is usually only available in discrete sizes e.g. number of lanes in a road, pipe sizes and wire sizes the problem becomes a large combinatorial optimisation problem. Techniques for optimisation of large combinatorial problems are discussed in Section 13.15.

Analytical Optimisation

Analytical optimisation can be applied when the objective function and constraints can be expressed in mathematical form. It involves the application of calculus or advanced mathematical optimisation techniques, some of which are discussed in detail in Sections 13.3–13.15 of this chapter.

13.3 LINEAR PROGRAMMING

The simplest case of the general mathematical optimisation problem consists of a single *linear objective function* and one or more *linear constraints*. This is called the *linear programming* (LP) problem. It may be written as follows:

$$\text{Max (or Min) } Z = \sum_{j=1}^{n} c_j x_j \qquad (13.6)$$

subject to

$$\sum_{j=1}^{n} a_{ij} x_j \{\leq, \ =, \text{ or } \geq\} b_i \quad (i = 1, \ldots, m) \qquad (13.7)$$

$$x_j \geq 0 \qquad (j = 1, \ldots, n) \qquad (13.8)$$

where x_j are the decision variables, Z is the objective function, c_j are the cost coefficients, a_{ij} are the constraint coefficients and b_i are the right hand sides. c_j, a_{ij} and b_i are assumed to be constant and known. Note that the non-negativity conditions (13.8) are assumed in all LP problems.

Linear programming has been used to solve many large scale problems in engineering, economics, and commerce. Examples include optimising the mix of products from a refinery, optimising the assignment of crews to vehicles for a public transport system, and optimising the allocation of water among conflicting uses.

Example 13.2 The engineer for the Bullbar City Council must decide how to allocate the Council's workforce and machinery among various activities. For simplicity it will be assumed that only two activities are possible, constructing roads and constructing drains. The engineer has estimated that each kilometre of new road brings a net benefit of $30,000 per year to the community, whereas each kilometre of drain brings a net benefit of $20,000 per year. A kilometre of road requires 250 person-days of labour and 640 machine-hours to construct. On the other hand, a kilometre of drain requires 500 person-days of labour and 320 machine-hours to construct. The Council has a workforce of 50 people and 10 machines. Assuming 200 effective working days of 8 hours each per year, write an LP formulation to determine the mix of roads and drains which the Council should undertake each year in order to maximise net benefits.

Solution Let x_1 be the number of km of roads to be constructed in a year, and x_2 be the number of km of drains to be constructed in a year. The objective function to be maximised is the total net benefits per year (in thousands of dollars), i.e.

$$\text{Max } Z = 30x_1 + 20x_2$$

The available workforce and machinery impose constraints. Considering the workforce first, the available resource is 50 people for 200 days, a total of 10,000 person-days. The required resource is $(250x_1 + 500x_2)$. Therefore:

$$250x_1 + 500x_2 \leq 10{,}000$$

Similarly for machinery, the available resource is 10 machines x 200 days x 8 hours/day, i.e. 16 000 machine-hours. Therefore:

$$640x_1 + 320x_2 \leq 16{,}000$$

In addition, the lengths of roads and drains constructed cannot be negative, and so

$$x_1 \geq 0 \qquad x_2 \geq 0$$

Therefore the problem formulation becomes:

$$\text{Max } Z = 30x_1 + 20x_2 \qquad (13.9)$$

subject to:

$$250x_1 + 500x_2 \leq 10{,}000 \qquad (13.10)$$
$$640x_1 + 320x_2 \leq 16{,}000 \qquad (13.11)$$
$$x_1 \geq 0 \qquad x_2 \geq 0 \qquad (13.12)$$

13.4 GRAPHICAL SOLUTION OF LP PROBLEMS

A graphical solution technique will be illustrated for Example 13.2 given above. Firstly, consider constraint (13.10). If this were an equality it could be plotted as a straight line in (x_1, x_2) space as shown in Figure 13.3. All combinations of x_1 and x_2 which satisfy this constraint lie below this line. Clearly any solution above this line is infeasible. This is indicated by shading on the infeasible side of the constraint as shown.

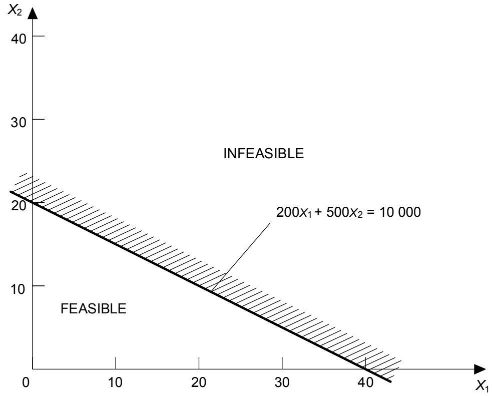


Figure 13.3 Workforce constraint for the Bullbar City Council problem.

In a similar fashion by plotting the other constraints (including the non-negativity constraints) the feasible solution space (or feasible region) can be identified. This is the region ABCD in Figure 13.4. The *feasible solution space* is the set of all combinations of the decision variables that satisfy all of the constraints of the problem.

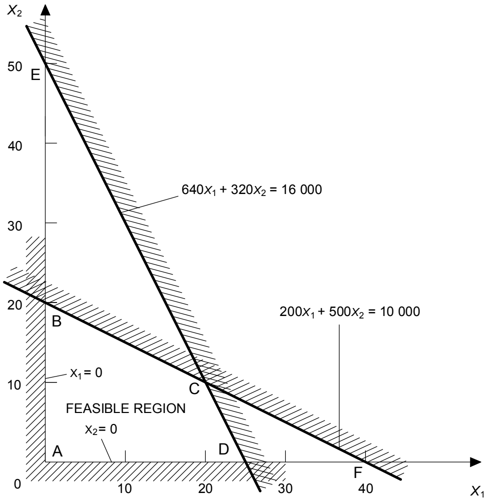


Figure 13.4 Feasible region for the Bullbar City Council problem.

The optimum solution may be determined by plotting contours of the objective function. As shown in Figure 13.5 these are parallel straight lines for any two-variable LP problem. It should be apparent that the optimum solution for this problem occurs at point C, i.e. at the intersection of the constraints (13.10) and (13.11). This corresponds to the highest value of the objective function Z that can be achieved within the feasible region.

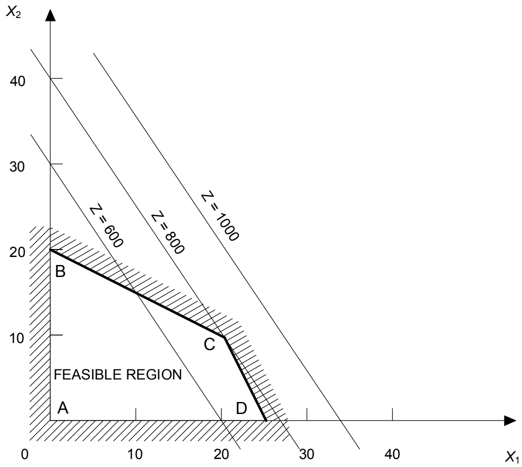


Figure 13.5 Optimum solution to the Bullbar City Council problem.

The optimum solution is:

$x_1 = 20$ km, $x_2 = 10$ km, $Z = \$800,000$ per year

If the net benefits of drains relative to roads were changed by a sufficient amount, point B or D would become the optimum solution.

It should be noted that for this example, and indeed all LP problems, the feasible solution space is a *convex region*. A convex region is one that satisfies the property that a straight line joining any two points in the region only contains points in the region. Any region that does not satisfy this property is called a *non-convex region*.

Examples of convex and non-convex regions in two dimensions are illustrated in Figure 13.6. It should also be apparent from Figure 13.5 that the optimal solution(s) to any two-dimensional LP problem always occurs at one (or two) *corner points*. A corner point is a feasible solution defined by the intersection of two constraints. It can be seen that if the contours of Z were parallel to BC, both B and C and the points in between would be optimum solutions.

In m dimensions, an *extreme point* is a feasible solution that occurs at the intersection of m constraints. It can be shown that the optimum solution to a general LP problem always occurs at one (or more) extreme points.

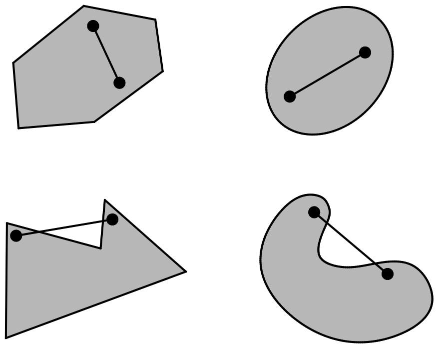


Figure 13.6 Examples of convex (top two) and non-convex (bottom two) regions.

13.5 CANONICAL AND STANDARD FORM OF AN LP PROBLEM

It is often useful in the solution of LP problems to reformulate them into either canonical form or standard form. The *canonical form* is

$$\text{Max } Z = \sum_{j=1}^{n} c_j x_j \tag{13.13}$$

subject to:

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i \qquad (i = 1, \ldots, m) \tag{13.14}$$

$$x_j \geq 0 \qquad (j = 1, \ldots, n) \tag{13.15}$$

It has the following features:

- the objective function is of the maximisation type;
- all constraints are of the 'less than or equal to' type;
- the right-hand side constants, b_i, may be positive or negative; and
- all decision variables are non-negative.

The canonical form of LP problems is used in Section 13.9 of this Chapter. Any LP problem may be rearranged into canonical form using one or more of the following operations:

(1) An objective function of the minimisation type can be made into the maximisation type by multiplying it by – 1 (and vice versa), e.g. minimising $Z = 2x_1 - 3x_2$ is the same as maximising $Y = -2x_1 + 3x_2$.

(2) 'Greater than or equal to' constraints can be changed to 'less than or equal to' constraints by multiplying both sides by -1, e.g. $5x_1 - 4x_2 \geq 12$ is equivalent to $-5x_1 + 4x_2 \leq -12$

(3) Equality constraints can be converted into two inequalities, e.g.
$4x_1 + 2x_2 = 12$
is equivalent to
$4x_1 + 2x_2 \leq 12$ and $4x_1 + 2x_2 \geq 12$
i.e.
$4x_1 + 2x_2 \leq 12$ and $-4x_1 - 2x_2 \leq -12$

(4) Variables which are unrestricted in sign may be replaced by two nonnegative variables; for example, consider the problem

Max $Z = 3y_1 + 2y_2$

subject to:

$y_1 + y_2 \leq 6$
$y_1 \leq 8$
$y_1 \geq 0$
y_2 unrestricted in sign

This may be reformulated by using the substitution $y_2 = (y_2^+ - y_2^-)$ where y_2^+ = the positive part of y_2 and y_2^- = the negative part of y_2. Thus, the problem becomes:

Max $Z = 3y_1 + 2y_2^+ - 2y_2^-$

subject to:

$$y_1 + y_2^+ - y_2^- \le 6$$
$$y_1 \le 8$$
$$y_1, y_2^+, y_2^- \ge 0$$

It can be shown that if y_2^+ is positive in the final solution, y_2^- will be zero and vice versa.

The *standard form* of an LP problem may be written as follows:

$$\text{Max (or Min) } Z = \sum_{j=1}^{n} c_j x_j \qquad (13.16)$$

subject to:

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \qquad (i = 1, \ldots, m) \qquad (13.17)$$

$$x_j \ge 0 \qquad (j = 1, \ldots, n) \qquad (13.18)$$

The standard form has the following features:

- the objective function may be of the maximisation or minimisation type;
- all constraints are equations, with the exception of the non-negativity conditions which remain as inequalities of the 'greater than or equal to' type;
- the right-hand side of each constraint must be non-negative; and
- all variables are non-negative.

The standard form is used in Section 13.6 of this Chapter. Any LP problem can be transformed into standard form by using one or more of the following operations (in addition to those listed above):

(1) A negative right-hand side can be made positive by multiplying through by -1 (and reversing the sign of the inequality).

(2) Inequality constraints may be converted to equalities by the inclusion of *slack* or *surplus variables*. For example, the 'less than or equal to' constraint,

$$2x_1 + 4x_2 \le 15 \qquad (13.19)$$

may be written as

$$2x_1 + 4x_2 + S_1 = 15$$

where S_1 is a *slack variable* which must be non-negative in order to satisfy the original constraint.

(3) Similarly, the 'greater than or equal to' constraint

$$3x_1 + 5x_2 \geq 18 \tag{13.20}$$

may be written as

$$3x_1 + 5x_2 - S_2 = 18$$

where S_2 is a *surplus variable* which must be non-negative.

The term slack variable is used because the right-hand side of Equation (13.19) often represents the availability of a particular resource (such as machine-hours in the case of the Bullbar City Council). If S_1 is positive then not all of the available resource is being utilized and there is 'slack' in the solution.

Constraints of the 'greater than or equal to' type often represent minimum requirements. For example, the right-hand side of Equation (13.20) may represent the minimum required output of steel from a particular process. If S_2 is positive the minimum output is being exceeded and there is a 'surplus' output.

13.6 BASIC FEASIBLE SOLUTIONS

By examining the standard form of an LP problem it can be seen that the constraints expressed by Equations (13.17) are a set of m linear equations in n unknowns (including slack and surplus variables).

Clearly we would expect n to be greater than m. If n is equal to m there is a unique solution to the constraint set and optimisation is not required. If n *is* less than m, the problem is over-constrained and there is, in general, no solution that satisfies all of the constraint equations.

A *basic* solution to an LP problem is one in which $(n - m)$ of the variables are equal to zero. A *basic feasible* solution to an LP problem is a basic solution in which all variables have non-negative values. Thus a basic feasible solution satisfies the constraints (13.17) and the non-negativity conditions (13.18).

Let us examine Example 13.2 further. In order to identify the basic solutions, we need to express the problem in standard form. This is achieved by introducing slack variables S_1 and S_2 into constraints (13.10) and (13.11) so that the problem becomes:

$$\text{Max } Z = 30x_1 + 20x_2 \tag{13.21}$$

subject to:

$$250x_1 + 500x_2 + S_1 = 10{,}000 \tag{13.22}$$
$$640x_1 + 320x_2 + S_2 = 16{,}000 \tag{13.23}$$
$$x_1, x_2, S_1, S_2 \geq 0 \tag{13.24}$$

Thus, excluding the non-negativity conditions, there are two constraint equations in four variables (including the slack variables S_1 and S_2). A basic solution may be found by setting any two of the variables equal to zero, and solving Equations (13.22) and (13.23) for the other two variables. In this manner the six basic solutions shown in Table 13.1 can be identified. Each of these basic solutions is identified by a letter in Figure 13.4. Clearly only the first four of these solutions are basic feasible solutions, as the other two do not satisfy the non-negativity conditions (13.21).

Table 13.1 Basic solutions to the Bullbar City Council problem.

Solution	x_1	x_2	S_1	S_2
A	0	0	10,000	16,000
B	0	20	0	9,600
C	20	10	0	0
D	25	0	3,750	0
E	0	50	-15,000	0
F	40	0	0	-9,600

It should be apparent from Figure 13.4 that each basic feasible solution is an *extreme point* of the feasible region. As it has already been shown that the optimum solution to an LP problem is an extreme point, it follows that *the optimum solution to an LP problem is always a basic feasible solution.*

Now consider the basic solutions to this problem a little more carefully. A basic solution occurs when two of the variables x_1, x_2, S_1 and S_2 equal zero. If we set S_1 equal to zero in Equation (13.22) we obtain

$$250x_1 + 500x_2 = 10{,}000 \qquad \text{(i.e. the line BCF in Figure 13.4)}$$

Similarly, by setting S_2 equal to zero in Equation (13.23) we obtain

$$640x_1 + 320x_2 = 16{,}000 \qquad \text{(i.e. the line ECD in Figure 13.4)}$$

The line where $x_1 = 0$ is the x_2 axis, and vice versa. Thus each constraint line (including the x_1 and x_2 axes) in Figure 13.4 corresponds to one variable being equal to zero. The basic solutions (where two variables equal zero) occur at the intersections of two constraint lines. Furthermore each basic feasible solution corresponds to the intersection of two constraints within the feasible region. Therefore, the basic feasible solutions must be the extreme points of the feasible region.

13.7 THE SIMPLEX METHOD

The simplex method is a numerical procedure commonly used for solving LP problems. It is based on the property that the optimal solution to an LP problem is always a basic feasible solution. The simplex method uses the following steps:

Step 1 Identify an initial basic feasible solution (i.e. an extreme point).

Step 2 Consider movement from this extreme point to all adjacent extreme points and determine whether it is possible to improve the value of the objective function.

Step 3 If improvement is possible choose the next extreme point such that the maximum rate of change of objective function is achieved and move to this point. Return to Step 2.

Step 4 If no improvement is possible, an optimum solution has been obtained.

The interested reader is referred to Taha (2017) for the details of the simplex method. Many computer packages are now available for solving very large LP problems using the simplex method.

13.8 SOME DIFFICULTIES IN SOLVING LP PROBLEMS

Occasionally difficulties may occur with LP problems caused by poor problem formulation or for other reasons. These problems and possible remedies are discussed in this Section.

Degenerate solutions

A degenerate solution occurs when one or more of the basic variables is equal to zero. One degenerate solution may be followed by another degenerate solution in which no improvement in the objective function occurs. This does not cause a problem with the simplex method and it is quite possible that several degenerate solutions may be followed by non-degenerate solutions in which an improvement in objective function is obtained. On some occasions, cycling between several degenerate solutions may occur. This rarely happens in practice and most commercial computer packages can identify it.

Unbounded solutions

It is sometimes possible to increase the value of the objective function without limit in an LP problem. This indicates a poorly formulated problem. For example:

$$\text{Max } Z = x_1 + 2x_2$$

subject to:

$$2x_1 + x_2 \geq 12$$
$$x_1 + x_2 \geq 8$$
$$x_1, x_2 \geq 0$$

The feasible region for this problem is shown in Figure 13.7. Clearly the objective function may be increased without limit in this feasible region. If the objective function were of the minimisation rather than maximisation type, an optimum solution could be found (assuming positive coefficients in the objective function).

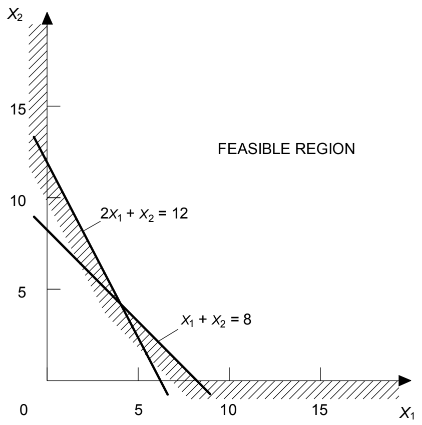


Figure 13.7 Feasible region for an unbounded problem.

No feasible solution

Another problem which may arise due to poor problem formulation is that there may be no solution which satisfies all of the constraints. For example:

$$\text{Max } Z = 2x_1 + x_2$$

subject to:

$$2x_1 + 3x_2 \geq 30$$
$$2x_1 + x_2 \leq 8$$
$$x_1 + 3x_2 \leq 9$$
$$x_1, x_2 \geq 0$$

The constraints to this problem are shown in Figure 13.8. Clearly, there is no feasible region and hence no feasible solution to this problem.

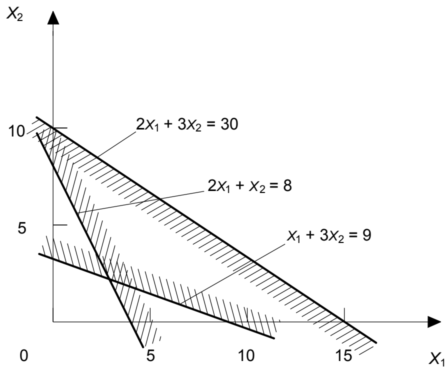


Figure 13.8 No feasible solution.

Alternate optima

Occasionally an LP problem has more than one optimum solution. This occurs when the objective function is parallel to a boundary of the feasible region. For example:

$$\text{Max } Z = 2x_1 + x_2$$

subject to:

$$x_1 + x_2 \leq 8$$
$$2x_1 + x_2 \leq 12$$
$$x_1, x_2 \geq 0$$

The feasible solution space for this problem is shown in Figure 13.9. Also shown is the contour for $Z = 18$. Because contours of the objective function are parallel to the second constraint for this problem, it should be clear that the value of the objective function is the same at points C and D. So both C and D and all points on the boundary between them are alternate optimum solutions to this problem.

The simplex method will identify either C or D as the optimum solution and then stop. The existence of alternative optima may be indicated as part of the output.

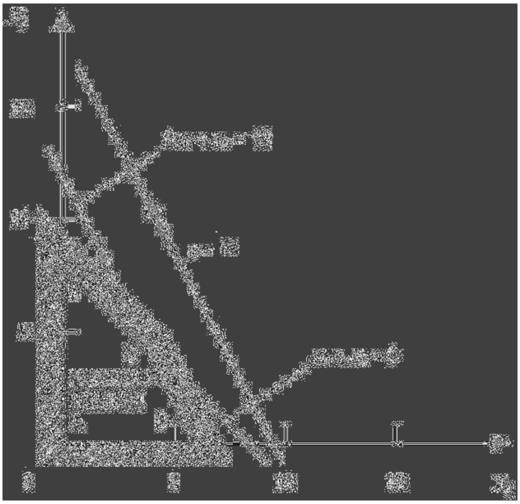


Figure 13.9 Alternate optima.

Redundant constraints

In some LP problems not all of the constraints are boundaries of the feasible region. Constraints that are not boundaries are called **redundant constraints** because they have no effect on the solution. For example:

$$\text{Max } Z = 2x_1 + x_2$$

subject to:

$$x_1 + x_2 \leq 8$$
$$x_1 + 2x_2 \leq 12$$
$$6x_1 + 5x_2 \leq 60$$
$$x_1, x_2 \geq 0$$

The constraints and feasible region for this problem are shown in Figure 13.10. Clearly the constraint $6x_1 + 5x_2 \leq 60$ is a redundant constraint as it has no effect on the feasible region.

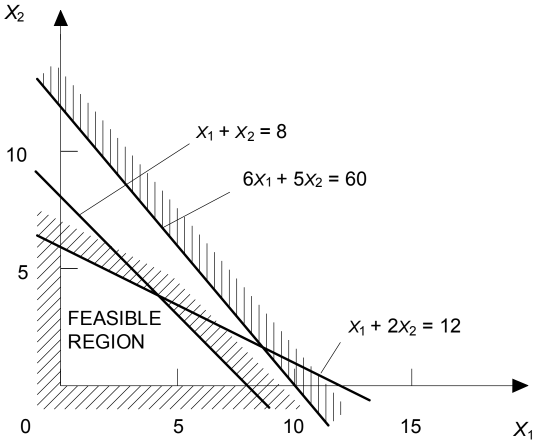


Figure 13.10 Redundant constraint.

The presence of redundant constraints does not in any way inhibit the simplex method from finding an optimum solution. However, because their presence affects the size of the problem and hence the computer time for its solution it is preferable to remove redundant constraints from the formulation if they can be readily identified.

13.9 EXAMPLES OF LP PROBLEMS

LP problems occur in many engineering design and planning problems. Some examples are outlined below.

Reservoir design

A common problem in water supply engineering is to determine the required size of a reservoir to provide a specified set of releases to a town or irrigation area.

Consider the proposed reservoir shown schematically in Figure 13.11. The design variable is the capacity of the proposed reservoir, K. Flow data have been collected at the proposed site for the last T years. Q_t ($t = 1, \ldots, T$) is the known flow in year t. The known annual water demand in the city is R. E_t ($t = 1, \ldots, T$) represents the annual evaporation and other losses from the reservoir in year t and is assumed to be known. S_t ($t = 0, \ldots, T$) represents the volume of storage in the reservoir at the end of year t. O_t ($t = 1, \ldots, T$) is the spill from the reservoir in year t. S_t ($t = 1, \ldots, T$) and O_t ($t = 1, \ldots, T$) are derived variables whose values will be determined in the course of the optimisation.

The reservoir size will be determined on the assumption that the historical series of flows will be repeated. The annual demand of the city must be met in every year without the reservoir becoming empty. The objective is therefore to minimise the required capacity, K, subject to the constraints imposed by:

- continuity; and
- the reservoir capacity not being exceeded.

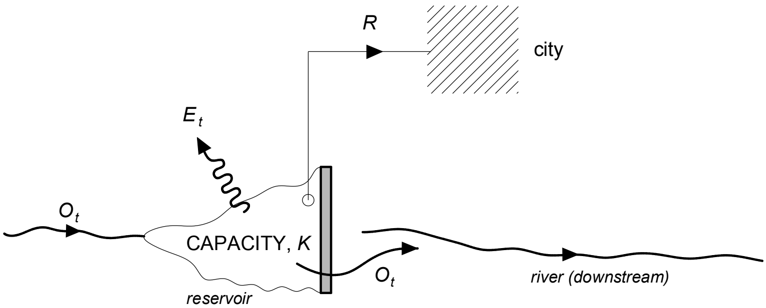


Figure 13.11 Reservoir design problem.

The continuity constraint for year t is as follows:

$$S_t - S_{t-1} = Q_t - E_t - R - O_t$$

i.e. the change in storage in the reservoir in year t equals the total inflow minus the total outflow. The constraint on capacity at the end of year t is simply

$$S_t \leq K$$

It is usual to assume that the reservoir has at least the same storage level at the end of the T years as at the beginning. This allows future demand to be met if the historical sequence repeats itself, i.e.

$$S_T \geq S_0$$

Therefore, the LP problem is:

$$\text{Min } Z = K \tag{13.25}$$

subject to:

$$S_t - S_{t\text{-}1} + O_t = Q_t - E_t - R \quad (t = 1, \ldots, T) \tag{13.26}$$
$$S_t - K \leq 0 \quad (t = 0, \ldots, T) \tag{13.27}$$
$$S_T - S_o \geq 0 \tag{13.28}$$
$$S_t, O_t, K \geq 0 \tag{13.29}$$

where decision variables have been placed on the left-hand side of constraints and known values on the right-hand side.

Road network analysis (adapted from Stark and Nicholls 1972, Example 4-5)

A common problem in transportation planning is the allocation of traffic volumes to a road network. A simple road network is shown in Figure 13.12.

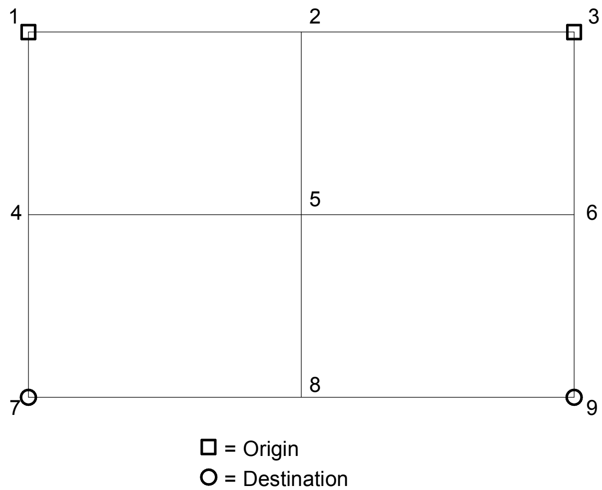


Figure 13.12 Portion of a road network.

The usual demand information for the network will be expressed in terms of an origin-destination matrix, i.e. the volume of traffic wishing to travel from each origin to each designation in a defined time period. Nodes 1 and 3 in Figure 13.12 are origins and nodes 7 and 9 destinations. A typical origin-destination matrix is given in Table 13.2. Other information available is the capacity (vehicles/h) and the travel time in each direction (minutes) for each road. Typical values are given in Table 13.3. In this example the capacity and travel time is the same in both directions for each road, although this is not necessarily always the case. The objective is to minimise the total time of travel for all traffic.

Table 13.2 Origin-Destination matrix (vehicle/h)

Origin node	Destination	Node
	7	9
1	4,000	3,000
3	3,000	2,000

Let $x_{ij}^{(k)}$ be the volume of traffic (vehicles/h) which originates from node k and travels on road ij, and let K be the set of all origin nodes. For all traffic originating at node k the total travel time is given by

$$Z^{(k)} = \sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}t_{ij}x_{ij}^{(k)} \tag{13.30}$$

where

$$\delta_{ij} = \begin{cases} 1 \text{ if there is a road from node } i \text{ to node } j \\ 0 \qquad\qquad \text{otherwise} \end{cases}$$

and t_{ij} is the travel time (minutes) on road ij and n is the total number of nodes.

Table 13.3 Capacities and travel times for road network problem (these are the same in both directions).

Road	Capacity (vehicles/h)	Travel time (minutes)
1-2	4,000	10
2-3	3,000	12
3-4	4,000	15
4-5	6,000	13
5-6	6,000	16
6-7	2,500	8
7-8	4,000	6
8-9	3,500	12
9-10	5,000	10
10-11	6,000	5
11-12	6,000	8
12-13	2,500	7

The constraints represent continuity at each node, i.e.

$$\sum_{i=1}^{n}\delta_{ip}x_{ip}^{(k)} - \sum_{j=1}^{n}\delta_{pj}x_{pj}^{(k)} = D_p^{(k)} \ (p = 1,...,n) \tag{13.31}$$

where

$$D_p^{(k)} = \begin{cases} \text{the volume of traffic which has origin } k \text{ and} \\ \quad \text{destination } p \text{ (for } p \neq k) \\ \\ \text{minus the volume of traffic which originates at node } k \\ \quad \text{(for } p = k) \end{cases}$$

When considering all origin nodes the objective function becomes

$$\text{Min } Z = \sum_{k \in K} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} t_{ij} x_{ij}^{(k)} \qquad (13.32)$$

where the first summation is over all origin nodes. The continuity constraints (13.31) become

$$\sum_{i=1}^{n} \delta_{ip} x_{ip}^{(k)} - \sum_{j=1}^{n} \delta_{pj} x_{pj}^{(k)} = D_p^{(k)} \quad \begin{matrix} (p = 1, \ldots, n) \\ (k \in K) \end{matrix} \qquad (13.33)$$

In addition, capacity constraints for each road must be considered. These are

$$\sum_{k \in K} x_{ij}^{(k)} \leq C_{ij} \quad \text{(for all } i, j \text{ which are connected by a road)} \qquad (13.34)$$

where C_{ij} the capacity of road ij. In addition,

$$x_{ij}^{(k)} \geq 0 \quad (\text{all } i, j, k) \qquad (13.35)$$

The LP problem to be solved is defined by the objective function (13.32) and the constraints (13.33), (13.34), and (13.35).

In a more realistic case the travel time on each road depends on the volume of traffic on it. In addition, each driver tries to minimise his or her own travel time instead of the system being at a global minimum. In these cases, the objective function becomes non-linear but the constraints have the same form.

13.10 DUALITY

Duality is an important concept in mathematical analysis that often gives valuable insights into the physical problem being solved. In LP, every problem has another problem associated with it called the *dual***.** By solving the original or *primal* problem using the simplex method, the solution to the dual is found automatically. The dual variables have an important interpretation that adds considerable information to an LP solution.

Writing the dual problem

In Section 13.5 the canonical form of an LP problem was given as

$$\text{Max } Z = \sum_{j=1}^{n} c_j x_j \qquad (13.13)$$

subject to:

$$\sum_{j=1}^{n} a_{ij} x_j \le b_i \quad (i = 1,\ldots,m) \qquad (13.14)$$

$$x_j \ge 0 \qquad (j = 1, \ldots, n) \qquad (13.15)$$

For an LP problem in canonical form the dual problem may be written using the following rules:

1. There is a dual variable corresponding to each primal constraint and a primal variable corresponding to each dual constraint.
2. If the primal problem is of the maximisation type, the dual problem is of the minimisation type and vice versa.
3. All of the constraints in the maximisation problem are of the 'less than or equal to' type. All of the constraints in the minimisation problem are of the 'greater than or equal to' type.
4. The coefficients of the objective function in the primal problem are the right-hand sides in the dual problem and vice versa.
5. The variables in both problems are non-negative.

To write the dual to the problem given by (13.13) - (13.15), we need to define m dual variables $y_i (i = 1, \ldots, m)$ with one corresponding to each primal constraint. The dual problem is:

$$\text{Min } W = \sum_{i=1}^{m} b_i y_i \qquad (13.36)$$

subject to:

$$\sum_{i=1}^{m} a_{ij} y_i \ge c_j \quad (j = 1,\ldots,n) \qquad (13.37)$$

$$y_i \ge 0 \qquad (i = 1, \ldots, m) \qquad (13.38)$$

Note that if these rules are applied to the problem defined by Conditions (13.36) - (13.38), the 'primal' problem is regained (i.e. Conditions (13.13) - (13.15)). Clearly, symmetry exists between the two problems.

It can be shown that, for any feasible solutions of the primal and dual, $Z \le W$. That is, the objective function value for a feasible solution of the maximisation problem is a lower bound on the objective function value of the minimisation problem. Furthermore, if Z^* is the maximum value of Z and W^* is the minimum value of W then

$$Z^* = W^* \qquad (13.39)$$

i.e. at optimality, both problems have the same objective function value. For example, the primal problem of Example 13.2 is given by (13.9) - (13.12):

$$\text{Max } Z = 30x_1 + 20x_2 \tag{13.9}$$

subject to:

$$250x_1 + 500x_2 \leq 10{,}000 \tag{13.10}$$
$$640x_1 + 320x_2 \leq 16{,}000 \tag{13.11}$$
$$x_1, x_2 \geq 0 \tag{13.12}$$

As this is already in canonical form, the dual can readily be written as

$$\text{Min } W = 10{,}000y_1 + 16{,}000y_2 \tag{13.40}$$

subject to:

$$250y_1 + 640y_2 \leq 30 \tag{13.41}$$
$$500y_1 + 320y_2 \leq 20 \tag{13.42}$$
$$y_1, y_2 \geq 0 \tag{13.43}$$

Writing each problem in standard form we obtain for the primal problem:

$$\text{Max } Z = 30x_1 + 20x_2 \tag{13.44}$$

subject to:

$$250x_1 + 500x_2 + S_1 = 10{,}000 \tag{13.45}$$
$$640x_1 + 320x_2 + S_2 = 16{,}000 \tag{13.46}$$
$$x_1, x_2, S_1, S_2 \geq 0 \tag{13.47}$$

and, for the dual problem:

$$\text{Min } W = 10{,}000y_1 + 16{,}000y_2 \tag{13.48}$$

subject to:

$$250y_1 + 640y_2 - S_3 = 30 \tag{13.49}$$
$$500y_1 + 320y_2 - S_4 = 20 \tag{13.50}$$
$$y_1, y_2, S_3, S_4 \geq 0 \tag{13.51}$$

The solution to the primal problem was found previously (Sections 13.4 and 13.7) to be

$$Z = 800,\ x_1 = 20,\ x_2 = 10,\ S_1 = S_2 = 0$$

The dual problem defined by (13.40) to (13.43) can be solved graphically or by using an LP computer package to give:

$$W = 800, y_1 = 1/75, y_2 = 1/24, S_3 = S_4 = 0$$

thus verifying Equation (13.39). Furthermore it can be shown that the simplex solution to a primal problem contains the solution to the dual (and vice versa).

Most LP computer packages (including Excel Solver) will give the optimal values of the dual variables when they solve the primal problem.

Interpretation of the dual variables

Refer back to the primal problem defined by (13.13) - (13.15). In many problems, the right-hand side element of the *i*th primal constraint, b_i, may be interpreted as the availability of the *i*th resource. Combining Equations (13.36) and (13.39), we have at optimality:

$$Z^* = W^* = \sum_{i=1}^{m} b_i y_i^* \tag{13.52}$$

where y_i* is the value of y_i at optimality ($i = 1, \ldots, m$).

Therefore, the optimum value of the *i*th dual variable, y_i*, may be interpreted as the additional output Z which could be obtained by a unit increase in the *i*th resource. Thus, if b_i were increased by one unit, we would expect Z* to increase by y_i* units.

In the above problem, the dual variable y_1 represents the additional net benefits (in thousands of dollars) if the right-hand side of constraint (13.10) were increased by one unit. In this case this represents an additional person-day of labour. Similarly y_2 represents the additional net benefits (in thousands of dollars) if the right-hand side of constraint (13.10) were increased by one unit, i.e. by having one extra machine-hour available. Therefore

y_1 = 1/75 thousand dollars/person-day

y_2 = 1/24 thousand dollars/machine-hour

y_1 = \$13.33/person-day

y_2 = \$41.67/machine-hour

Expressed in this form the dual variables are *shadow prices* for the corresponding resources. The shadow price of a resource is the economic value of having an additional unit of the resource available.

If additional labour can be hired at less than the shadow price of \$13.33 per day the Council would make a net gain by doing so. Similarly if additional machine time can be obtained at less than its shadow price of \$41.67 per hour a net gain can be made. Of course, this gain by employing additional labour or machines will only apply for increases that are not so large as to cause a change in the optimum solution.

13.11 NON-LINEAR OPTIMISATION

In Sections 13.3 to 13.10 a special case of optimisation called linear programming (LP) is examined. This applies to problems with a single linear objective function and a set of linear constraints. Many general concepts of optimisation can be demonstrated using LP. However, the assumption of linearity is often not appropriate for real engineering systems. In general, the objective function and/or constraints may be non-linear, if indeed they can be expressed in mathematical form.

There are many techniques available for the solution of non-linear optimisation problems. In the following sections, the techniques of separable programming and dynamic programming are considered. Many practical engineering problems do not lend themselves to the use of traditional optimisation techniques. However, it is usually possible to develop a simulation model of the system under consideration. Computer simulation can then be used in a trial-and-error search for the optimum solution. In recent years, a new branch of optimisation techniques called heuristic techniques (or evolutionary algorithms) has developed. These are guided search techniques that work in conjunction with a simulation model to search the decision space in an efficient manner and identify a number of "near optimal" solutions. Heuristic optimisation methods are discussed in more detail in Section 13.15.

The general single-objective, non-linear optimisation problem may be written as:

$$\text{Max } Z = f(x_1, x_2, \ldots, x_n) \tag{13.53}$$

subject to:

$$h_i(x_1,\ldots,x_n)\begin{Bmatrix} \le \\ = \\ \text{or} \ge \end{Bmatrix} b_i \; (i = 1,\ldots,m) \tag{13.54}$$

This problem differs from the general LP problem (Conditions 13.6 - 13.8) in the following three ways:

1. The feasible solution space for a non-linear problem may be either convex or non-convex. (This follows from the fact that the boundaries of the feasible region may, in general, be generated by non-linear functions. Figure 13.6 illustrates some non-convex feasible regions.)
2. An optimum solution to a non-linear problem may occur at any point in the feasible region (not necessarily at an extreme point).
3. It is possible for more than one 'optimum' to exist. The solution corresponding to the absolute maximum (or minimum) of the objective function is called the *global optimum*; all other optima are called *local optima*.

Two examples are used to illustrate the second point.

Example 13.3 Consider the following problem:

$$\text{Min } Z = (x_1 - 2)^2 + (x_2 - 3)^2$$

subject to:

$$2x_1 + x_2 \le 14$$
$$x_1 + 3x_2 \le 15$$
$$x_1, x_2 \ge 0$$

The feasible solution space and contours of the objective function for this problem are shown in Figure 13.13. The minimum value of Z equals zero and occurs at $x_1 = 2$, $x_2 = 3$. This is in the interior of the feasible region.

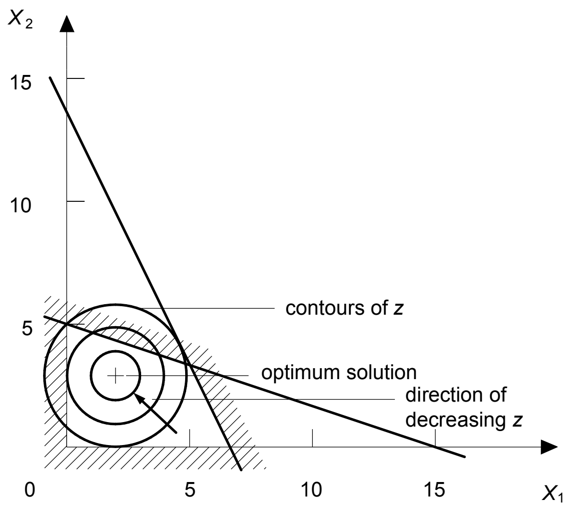


Figure 13.13 Optimum solution at interior of the feasible region.

Example 13.4 Consider the following problem:

$$\text{Max } Z = x_1 x_2$$

subject to:

$$x_2 + 0.1x_1^2 \le 10$$
$$x_1 \le 7$$
$$x_1, x_2 \ge 0$$

The feasible region and contours of Z are shown in Figure 13.14. In this case the optimum solution ($Z = 38.49$ at $x_1 = 5.77$, $x_2 = 6.67$) lies on the boundary of the feasible region but not at an extreme point.

In relation to the existence of local optima, it is useful to think of the objective function surface as a range of mountains. In a maximisation problem every peak is a local optimum, and the highest peak is the global optimum. Of course, in many problems, the objective function has only one peak and there are no local optima. Depending on the constraints this may or may not correspond to the optimum solution.

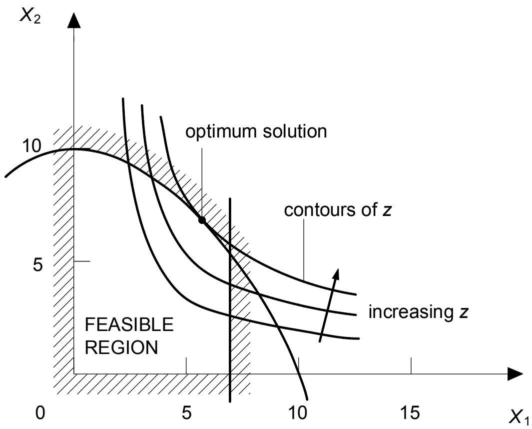


Figure 13.14 Optimum solution at boundary of the feasible region.

In linear programming, the feasible solution space is always convex, the optimum solution always occurs at an extreme point and local optima do not occur. The three differences between linear and non-linear optimisation listed above make non-linear optimisation considerably more difficult than linear optimisation. In fact, there is no single solution technique that works for all non-linear optimisation problems. Many specific techniques have been developed, each of which is applicable to a certain category of non-linear optimisation problems. The techniques may be grouped into the following categories:

- the graphical technique;
- classical optimisation techniques; and
- numerical techniques.

The graphical technique may be used for problems with only one or two decision variables. It involves plotting the constraints and the objective function in terms of the decision variables in order to identify the optimum solution. The technique is illustrated by Examples 13.3 and 13.4 considered earlier. It is generally only suitable for simple problems.

13.12 UNCONSTRAINED PROBLEMS USING CALCULUS

The general unconstrained optimisation problem is:

$$\text{Max } Z = f(x_1, x_2, \ldots, x_n) \tag{13.55}$$

where $f(x_1, x_2, \ldots, x_n)$ is any non-linear function of the n decision variables $x_1, \ldots, x_n$. This is perfectly general, as a minimisation problem can be converted into a maximisation problem by multiplying the objective function by -1. Firstly consider the single-variable case:

$$\text{Max } Z = f(x) \tag{13.56}$$

Classical differential calculus can be used to find the maxima of $f(x)$ provided that $f(x)$ and its first- and second-order derivatives exist and are continuous for all values of x.

A *necessary condition* for a particular solution x^o to be a maximum of $f(x)$ is:

$$\left.\frac{df(x)}{dx}\right|_{x=x^0} = 0 \tag{13.57}$$

The solutions to Equation (13.57) are called *stationary points* and may include maxima, minima, and points of inflection. A *point of inflection* is a point where $f(x)$ has zero slope but is neither a maximum nor a minimum. For example, the function $f(x) = x^3$ has a point of inflection at $x = 0$ as shown in Figure 13.15.

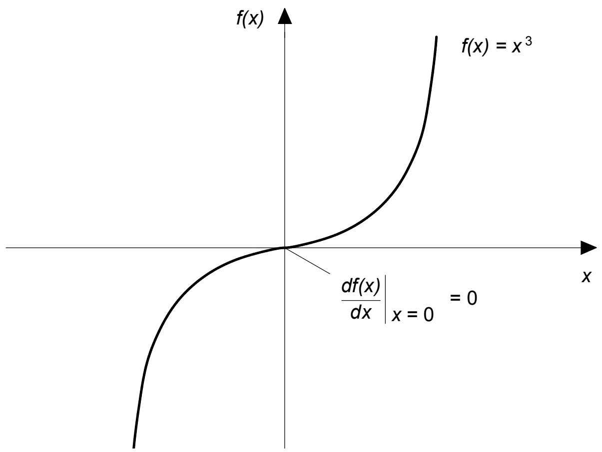


Figure 13.15 Point of inflection.

A *sufficient* condition for a stationary point x^0 to be a maximum of $f(x)$ is:

$$\left.\frac{d^2 f(x)}{dx^2}\right|_{x=x^0} < 0 \tag{13.58}$$

A sufficient condition for a minimum at x^o is:

$$\left.\frac{d^2 f(x)}{dx^2}\right|_{x=x^0} > 0 \qquad (13.59)$$

Example 13.1 (repeated): A tank of volume 200 m^3 is to be made from stainless steel. It will have a cylindrical body and hemispherical ends as shown in Figure 13.15. If the material for the ends costs twice as much per square metre as that for the body, find the dimensions of the tank that minimise its cost.

Solution: Let the length of the body of the tank be b m and its radius be r m. The formulation given in Section 13.1 is repeated below.

$$\text{Min } Z = 2\pi rb + 8\pi r^2 \qquad (13.2)$$

subject to the volume constraint:

$$\pi r^2 b + 4\pi r^3/3 = 200 \qquad (13.3)$$

This constraint can be written as an equality because the minimum cost tank will not have excess volume in this case. Using Equation (13.3) we can substitute for b in the objective function

$$b = \frac{\left[200 - 4\pi r^3/3\right]}{\pi r^2} \qquad (13.60)$$

Thus (13.2) becomes:

$$\text{Min } Z = \frac{400}{r} + \frac{16}{3}\pi r^2 \qquad (13.61)$$

which is an unconstrained single-variable optimisation. The necessary condition (13.57), is found by differentiating Z with respect to r,

$$\frac{\mathrm{d}Z}{\mathrm{d}r} = \frac{-400}{r^2} + \frac{32\pi r}{3} = 0$$

then solving for r gives

$$r = 2.285 \text{ m}$$

substituting into Equation (13.3),

$$b = 9.146 \text{ m}$$

Now checking the sufficiency condition (13.58):

$$\frac{d^2Z}{dr^2} = \frac{+800}{r^3} + \frac{32\pi}{3}$$

which is positive for all positive values of r, therefore the above solution is a minimum. In practice the chosen dimensions would be rounded off, e.g. $b = 9$ m, $r = 2.3$ m.

The necessary and sufficient conditions for the unconstrained multivariate case are discussed in Appendix 13A. Classical optimisation techniques can be extended to solve constrained multivariate problems (particularly those with equality constraints). This is beyond the scope of this book and the interested reader is referred to Taha (2017) or Ossenbruggen (1984) for a treatment of this topic.

Numerical optimisation techniques are often preferred to calculus-based techniques because of the difficulty in applying the latter to problems of reasonable size. A large number of numerical techniques are available for solving non-linear optimisation problems.

The techniques considered in this chapter are separable programming and dynamic programming. Taha (2017) and Beightler et al. (1979) contain a more detailed treatment of various numerical techniques suitable for solving non-linear optimisation problems.

13.13 SEPARABLE PROGRAMMING

Separable programming is an extension of linear programming. It works by converting a non-linear optimisation problem into an LP problem that approximates the original non-linear one. Separable programming can only be applied to separable functions. A function $f(x_1, x_2, \ldots, x_n)$ is said to be separable if it can be expressed as the sum of n single-variable functions $f_l(x_l), \ldots, f_j(x_j), \ldots, f_n(x_n)$.

Although separable programming can be used to approximate any continuous non-linear function it is most effective when the following special conditions apply:

- the individual functions, $f_j(x_j)$, of the objective function are concave for a maximisation problem or convex for minimisation problems; and
- the constraints define a convex feasible region.

A *convex function* is one that has the property that a straight line joining any two points on the function always lies on or above the function. For a *concave function* a straight line joining any two points on the function always lies on or below the function. Examples of convex and concave functions are shown in Figure 13.16. A convex region has been defined in Section 13.4.

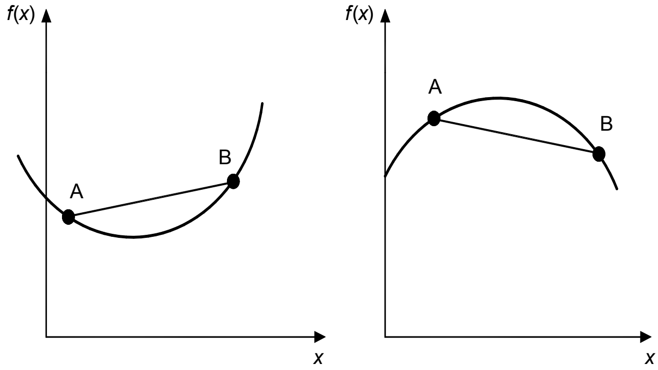


Figure 13.16 Convex and concave functions

If the properties defined above do not apply, separable programming involves the use of integer variables. This is beyond the scope of this book. The interested reader is referred to Taha (2017).

Separable programming always yields a global optimum to the approximate linear problem. To illustrate the separable programming technique, consider the minimisation of the convex objective function shown in Figure 13.17. The non-linear function is approximated by a series of straight line segments between selected values of x, namely λ_1, λ_2, ….., λ_K which must be selected to span the full range of likely values of x.

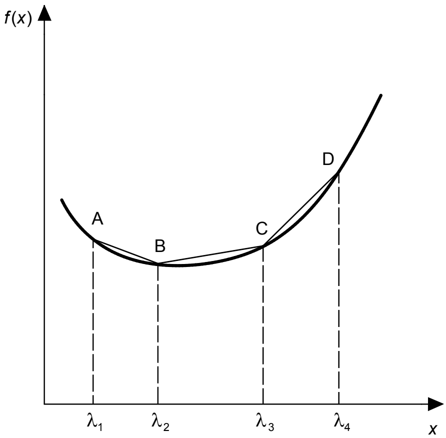


Figure 13.17 A convex objective function

We define a new set of variables $w_1, \ldots, w_K$ such that

$$x = \sum_{k=1}^{K} w_k \lambda_k \qquad (13.62)$$

$$\sum_{k=1}^{K} w_k = 1 \tag{13.63}$$

and $$w_k \geq 0 \qquad (k = 1, \ldots, K) \tag{13.64}$$

In the final solution, x is a weighted average of the values (λ_1, λ_2, ..., λ_K) where w_k (k = 1, . . ., K) are the relative weights. Provided at most two adjacent values of w_k are non-zero, $f(x)$ can be approximated by

$$f(x) = \sum_{k=1}^{K} w_k f(\lambda_k) \tag{13.65}$$

For example, if $w_2 = w_3 = 0.5$ and all other values of w_k are zero, Equations (13.62) and (13.65) give:

$$x = 0.5\lambda_2 + 0.5\lambda_3$$
$$f(x) = 0.5 f(\lambda_2) + 0.5 f(\lambda_3)$$

which is a point on the straight line midway between points A and B in Figure 13.18.

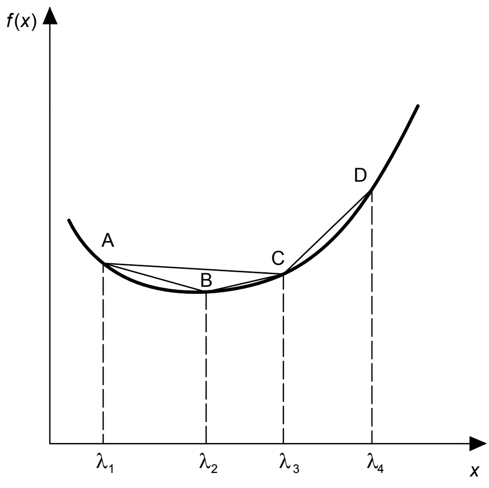


Figure 13.18 A convex objective function

It can be shown that, at most, two adjacent weights w_k will be non-zero when using LP and the special conditions above apply. This can be demonstrated intuitively for the case of minimising a convex function as shown in Figure 13.18. Suppose as a particular solution we obtain w_1 = 0.6 and w_3 = 0.4 with all other w_k = 0. Then

$$x = 0.6\lambda_1 + 0.4\lambda_3$$
$$f(x) = 0.6f(\lambda_1) + 0.4f(\lambda_3)$$

This is a point on the straight line segment between points A and C. Clearly the line segments AB and BC lie below this and an LP solution which is minimising $f(x)$ will achieve a solution on one of the lower line segments – i.e. with only two adjacent weights non-zero.

Example 13.5: A cylindrical tank is to be made from a 6 m square sheet of metal as shown in Figure 13.19. The tank will be h metres high and d metres in diameter. The height of the tank must not be more than twice its diameter. Use separable programming to find the dimensions of the tank that maximise its volume.

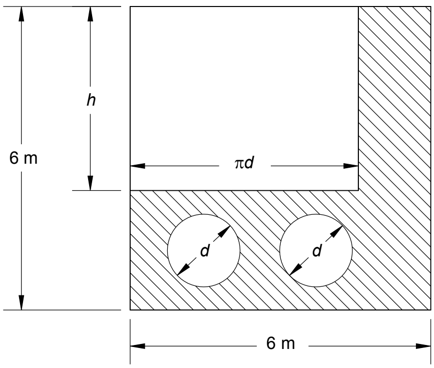


Figure 13.19 Plan for side and ends of a cylindrical tank

Solution: The optimisation problem is:

$$\text{Max V} = \pi \text{d}^2 h/4 \tag{13.66}$$

subject to

$$\pi d \leq 6$$

i.e. $$d \leq 1.910 \tag{13.67}$$

$$h + d \leq 6 \tag{13.68}$$

$$h/d \leq 2$$

i.e. $$h - 2d \leq 0 \tag{13.69}$$

$$h, d \geq 0 \tag{13.70}$$

The objective function can be made separable by taking its logarithm (to the base 10), i.e.

$$\log V = \log(\pi/4) + 2\log d + \log h \qquad (13.71)$$

Clearly, maximising log *V* is equivalent to maximising *V*. As log *d* and log *h* are concave functions in a maximisation objective, we can use separable programming without integer variables.

It is now necessary to approximate log *d* and log *h* over a suitable range. First consider log *h*. From Equation (13.68) it is obvious that *h* must be less than 6 m. An approximation to log *h* for *h* between 0.5 m and 6 m is shown in Figure 13.20. Values for λ_k and $\log_{10}\lambda_k$ are given in Table 13.4. Applying Equations (13.62) – (13.65) to approximate the function log (*h*) gives

$$h = 0.5w_1 + w_2 + 2w_3 + 4w_4 + 6w_5 \qquad (13.72)$$

$$w_1 + w_2 + w_3 + w_4 + w_5 = 1 \qquad (13.73)$$

$$w_i \geq 0 \qquad (i = 1,...,5) \qquad (13.74)$$

$$\log h = -0.301w_1 + 0.w_2 + 0.301w_3 + 0.602w_4 + 0.778w_5 \qquad (13.75)$$

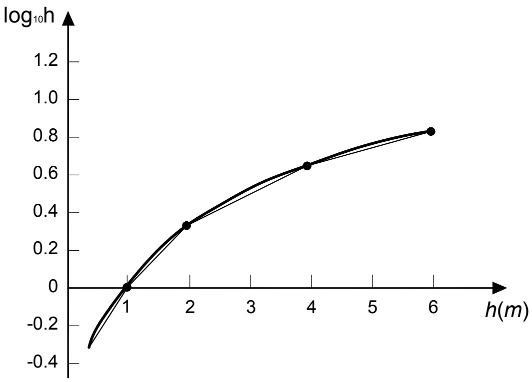


Figure 13.20 Approximation to log10 h

As *d* must be less than or equal to 1.91 m, log *d* will be approximated over the range 0.5 m – 2 m. Using Table 13.4, this gives

$$d = 0.5u_1 + u_2 + 2u_3 \qquad (13.76)$$

$$u_1 + u_2 + u_3 = 1 \qquad (13.77)$$

$$u_j \geq 0 \qquad (j = 1,...,3) \qquad (13.78)$$

$$\log d = -0.301u_1 + 0.u_2 + 0.301u_3 \qquad (13.79)$$

Thus, combining Equations (13.71), (13.75), and (13.79), the objective function becomes

$$\text{Max } \log V = \log(\pi/4) - 0.602u_1 + 0.602u_3$$
$$-0.301w_1 + 0.301w_3 + 0.602w_4 + 0.778w_5$$

subject to the constraints (13.67) to (13.70), (13.72) to (13.74) and (13.76) to (13.78).

Solving this problem using LP gives the solution: h = 3.82 m, d = 1.91 m, and V = 10.95 m^3, which can be shown to be the global optimum.

Table 13.4 Values of λ_k and log10 λ_k for the cylindrical tank problem

k	λ_k	$\log_{10} \lambda_k$
1	0.5	-0.301
2	1.0	0.000
3	2.0	0.301
4	4.0	0.602
5	6.0	0.778

13.14 DYNAMIC PROGRAMMING

Dynamic programming is a non-linear optimisation technique that is applicable to a wide class of multi-stage decision problems. The term *dynamic* indicates the application of the technique to problems that involve movement from one point in time or space to another point by means of a number of intermediate stages (intervals of time or distance). At each stage a decision has to be taken, i.e. a choice from several alternatives has to be made, and the problem is to find the optimum sequence of decisions. The method will be illustrated by the following example.

Example 13.6: Goods are to be transported by truck from city A to city K via the highway network shown schematically in Figure 13.21. The cost of transport on each length of highway in dollars is indicated on the figure. The problem is to find the minimum-cost route from city A to city K.

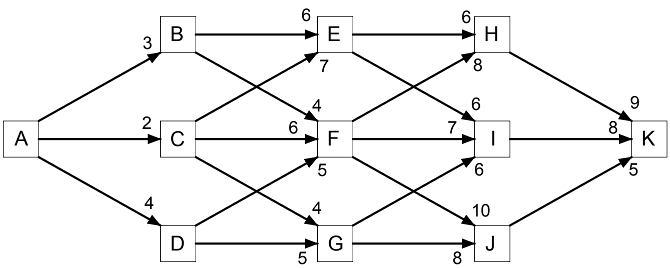


Figure 13.21 Highway routes and travel costs in dollars.

Solution: Inspection of Figure 13.21 indicates that, regardless of the route taken, the truck must pass through three other cities. Thus we may consider the problem to consist of four sequential decisions. We say that the problem consists of four stages. A *stage* is defined as the interval that separates consecutive decision points.

One method for determining the minimum-cost route would be to evaluate all alternatives. This is called *complete enumeration*, and in this case it would involve the costing of 17 feasible routes. For a larger problem containing, for example, 6 stages and 10 cities at each stage, there would be 10^5 alternative paths to evaluate. Clearly a more efficient approach is needed. Dynamic programming provides such an approach.

Dynamic programming is based on the *Principle of Optimality*, which may be stated as follows:

> *An optimal policy must have the property that regardless of the decisions leading to a particular state, the remaining decisions must constitute an optimal sequence for leaving that state.*

As discussed in Section 2.4, the 'state' of a system summarizes concisely its entire previous history and therefore allows the future behaviour to be determined for any future input. In this example the states are the cities passed through. When applied to the current problem, the principle of optimality means that if the optimum route passes through state F, it must conclude with the optimum route from F to K.

Thus the principle of optimality can be applied one stage at a time through the network in order to identify the minimum-cost path. Although it is possible to proceed either backwards or forwards in time, it is usually advantageous to work backwards. The present problem will be worked backwards.

With one stage to go, the truck could be in state H, I, or J. There is no choice of destination, which must be K. Nevertheless the calculations are set up in tabular form in order to be consistent with the rest of the solution. The information with one stage to go is summarized in Table 13.5(a). Regardless of the initial state, the decision must be to move to K. The optimum decisions and corresponding minimum costs for these trips are summarized in Table 13.5(a).

With two stages to go the truck could be in city E, F, or G and a decision must be made regarding the next state. The minimum cost for each decision is the sum of the cost for the first stage plus the minimum cost for the remaining stages. This information is summarized in Table 13.5(b).

If starting in city E there is no direct route to city J, so the choice is between cities H and I. The minimum cost of proceeding through H is the direct cost ($6) plus the minimum cost from H onwards which is obtained from Table 13.5(a) as $9. Similarly, the minimum cost of proceeding from E through I is the direct cost ($6) plus the minimum cost from I onwards ($8). The optimum decision from E corresponds to the smallest cost, namely $14 via city I. Similar calculations are made for the initial states F and G.

Table 13.5(b) indicates the optimum decisions from all cities with two stages to go. Thus if a truck driver should ever find himself in city G, the optimum decision is to go to city J, and so on.

At this point we have evaluated all possible paths over the last two stages. Now with three stages to go, dynamic programming enables us to evaluate only a subset of all possible paths.

Table 13.5 Decision tables for highway route problem.

(a) With one stage remaining

Initial state	Decision: *K*	Optimum decision	Minimum cost ($)
H	9	K	9
I	8	K	8
J	5	K	5

(b) With two stages remaining

Initial state	Decision: H	I	J	Optimum decision	Minimum cost ($)
E	6 + 9	6 + 8	—	I	14
F	8 + 9	7 + 8	10 + 5	I or J	15
G	—	6 + 8	8 + 5	J	13

(c) With three stages remaining

Initial state	Decision: E	F	G	Optimum decision	Minimum cost ($)
B	6 + 14	4 + 15	—	F	19
C	7 + 14	6 + 15	4 + 13	G	17
D	—	5 + 15	5 + 13	G	18

(d) With four stages remaining

Initial state	Decision: B	C	D	Optimum decision	Minimum cost ($)
A	3 + 19	2 + 17	4 + 18	C	19

Table 13.5(c) summarizes the calculations with three stages to go. If starting in city B, the choice is between proceeding through city E or F. The minimum cost of travelling through E is the sum of the direct cost ($6) plus the minimum cost from E onwards. The latter term is given in Table 13.5(b) as $14. The minimum cost of travelling through city F may be assessed in a similar way. Thus the optimum route from B is via city F at a cost of $19. The minimum cost route from states C and D are also indicated in Table 13.5(c).

Finally, with four stages to go the initial state is A and the choice is between cities B, C, and D. The minimum cost via B is the direct cost of $3 plus the minimum cost from B onwards ($19 from Table 13.5(c)). The minimum-cost route is via C at a cost of $19.

The optimum route through the network can be determined by working forwards from the starting state A using Table 13.5. Table 13.5(d) indicates that from A we should proceed to C. Table 13.5(c) shows that G is the optimum choice from C. From G we go to J (Table 13.5(b)) and from J to K (Table 13.5(a)). Therefore the minimum cost route is A-C-G-J-K at a cost of $19. Note also that the dynamic programming solution identifies the optimum solution extending from any non-optimal intermediate state. For example, should an incorrect initial decision lead to B instead of C, the optimal subsequent sequence is B-F-I-K or B-F-J-K.

The benefits of using dynamic programming to solve this problem should be apparent from Table 13.5. With three stages to go only seven comparisons had to be made, even though there are seventeen possible routes through the last three stages. For larger problems, the advantage of using dynamic programming is more marked. For example, a problem with six stages and ten states at each stage contains 10^5 paths, but dynamic programming involves only 410 comparisons.

It should be noted that dynamic programming involves the implicit evaluation of all possible paths. Therefore it always achieves the *global optimum***.** It is clearly a very powerful technique for problems that can be organized into an appropriate form.

The dynamic recursive equation

Although computer packages for solving general dynamic programming problems are available, it is often better for the analyst to write a computer program for the specific problem being solved. The tabular form of solution given above is suitable for hand solution of dynamic programming problems. Computer codes for solving dynamic programming problems rely on the development of a suitable recursive equation as follows:

- Let $c_n(i,j)$ be the cost of proceeding from state i with n stages to go to state j with $(n$-1) stages to go.
- Let $f_n(i)$ be the cost of the optimum policy to the end of the final stage when starting in state i with n stages to go.
- Let N be the total number of stages in the problem.

Then the dynamic recursive relationship for a minimisation problem may be written as:

$$f_n(i) = \min_{j} \left[c_n(i,j) + f_{n-1}(j) \right] \quad (n = 1, \ldots N) \qquad (13.80)$$

Thus, the cost of the optimum policy when starting in state i with n stages to go is the minimum over all feasible states j of going from i to j plus the cost of the optimum policy when in state j with $n-1$ stages to go.

The use of the dynamic recursive equation will be demonstrated for a problem of reservoir operation. This is typical of inventory type problems that involve carry-over storage of various commodities such as raw materials, finished product or, as in this case, water.

Example 13.7: A reservoir of capacity 6 units is situated on a river and is used to regulate flow releases to downstream irrigators. A maximum of 4 units can be released in one season. An optimum release policy is required over the next four seasons beginning with winter. The economic benefits to irrigators for various releases are shown in Table 13.6. The expected inflows to the reservoir over the next four seasons are 4, 3, 1, and 2 units, respectively. At the beginning of winter the reservoir starts with 3 units in storage and it must have at least 3 units in storage at the end of autumn. Use dynamic programming to find the optimum release policy.

Table 13.6 Seasonal benefits for release of water for irrigation.

Release (units)	Benefits ($m) for release during:			
	winter	spring	summer	autumn
1	0.2	1.5	1.3	0.5
2	0.4	2.5	2.3	1.0
3	0.5	3.3	3.0	1.2
4	0.6	3.8	3.3	1.3

Solution: The stages in this problem are the four seasons; the state variable is the storage in the reservoir at the end of each season as this summarizes the effects of all previous inputs and decisions. The objective function (to be maximised) is the total benefits of water released for irrigation.

Figure 13.22 shows the feasible states for this problem. Given the initial storage of 3 units and the winter inflow of 4 units the minimum storage at the end of winter is 3 units corresponding to a release of 4 units. The maximum storage is 6 units which corresponds to the release of 1 unit. (A release of zero units is not possible because a spill of 1 unit will occur in any case.) At the end of spring a minimum storage of 2 units can be achieved by releasing 4 units from the initial storage of 3 units.

At the end of summer, the minimum feasible storage is 1 unit which can be achieved via several paths. Although a storage of zero units can be achieved, it is not possible to come up to the desired storage of 3 units at the end of autumn from this state. The storage at the end of autumn will be 3 units in most cases as there is no advantage in having more water in storage. However, if the storage at the end of summer is 6 units, the minimum achievable storage at the end of autumn is 4 units.

The dynamic recursive equation (13.80) will be applied backwards in time to this problem. In this case, let i_n be the reservoir storage with n stages remaining, x_n be the reservoir release during the current stage with n stages remaining, and q_n the inflow during the current stage with n stages remaining (known).

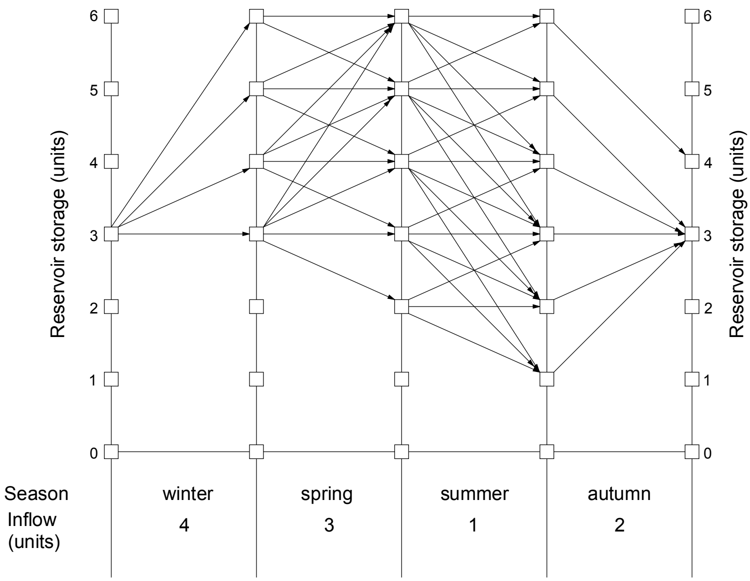


Figure 13.22 Feasible states for the reservoir release problem.

Clearly

$$i_{n-1} = i_n - x_n + q_n \qquad (13.81)$$

i.e. the reservoir storage with (n-1) stages remaining is the storage with n stages remaining minus the release plus the inflow during the current stage. Therefore, the dynamic recursive relation for this problem becomes:

$$f_n(i_n) = \max_{0 \le x_n \le 4} \left[c_n(x_n) + f_{n-1}(i_{n-1}) \right] \quad \begin{matrix} (n = 0,1,\dots,4) \\ \text{all feasible } i_n \end{matrix} \qquad (13.82)$$

where $f_n(i_n)$ is the maximum benefit when in state i with n stages remaining and $c_n(x_n)$ is the benefit of releasing x_n when n stages remain. Substituting Equation (13.81) into Equation (13.82) gives

$$f_n(i_n) = \max_{0 \le x_n \le 4} \left[c_n(x_n) + f_{n-1}(i_n - x_n + q_n) \right] \quad \begin{matrix} (n = 0,1,\dots,4) \\ \text{all feasible } i_n \end{matrix} \qquad (13.83)$$

Now to apply Equation (13.83) backwards in time, we start by defining

$$f_0(i_0) = 0 \quad (i_0 = 0,1,\dots,6) \qquad (13.84)$$

(a) *With one* stage *remaining:*

$$f_1(i_1) = \max_{0 \le x_1 \le 4} \left[c_1(x_1) + f_o(i_1 - x_1 + 2) \right] \quad (\text{for } i_1 = 1,...,6)$$

where values of $c_1(x_1)$ are given in the right hand column of Table 13.6. Clearly, there is no choice at this stage, so we simply determine:

$$f_1(1) = [c_1(0) + f_0(1 - 0 + 2)] = [c_1(0) + f_0(3)] = 0 + 0 = \$0 \left(\text{with } x_1^*(1) = 0\right)$$

where $x_n*(i_n)$ denotes the optimum release when in state i_n with n stages remaining.

$$f_1(2) = [c_1(1) + f_0(2 - 1 + 2)] = 0.5 + 0 = \$0.5\,\text{m (with } x_1^*(2)=1)$$
$$f_1(3) = [c_1(2) + f_0(3 - 2 + 2)] = 1.0 + 0 = \$1.0\,\text{m (with } x_1^*(3)=2)$$
$$f_1(4) = [c_1(3) + f_0(4 - 3 + 2)] = 1.2 + 0 = \$1.2\,\text{m (with } x_1^*(4)=3)$$
$$f_1(5) = [c_1(4) + f_0(5 - 4 + 2)] = 1.3 + 0 = \$1.3\,\text{m(with } x_1^*(5)=4)$$
$$f_1(6) = [c_1(4) + f_0(6 - 4 + 2)] = 1.3 + 0 = \$1.3\,\text{m (with } x_1^*(6)=4)$$

These calculations are summarized in Table 13.7(a).

(b) With two stages remaining:

Equation (13.82) becomes:

$$f_2(i_2) = \max_{0 \le x_2 \le 4} \left[c_2(x_2) + f_1(i_2 - x_2 + 1) \right] \quad (\text{for } i_2 = 2,3,...,6)$$

with $i_2 = 2$ only releases of 0, 1, or 2 units are feasible. Comparing these,

$$f_2(2) = \max \begin{bmatrix} c_2(0) + f_1(2 - 0 + 1) \\ c_2(1) + f_1(2 - 1 + 1) \\ c_2(2) + f_1(2 - 2 + 1) \end{bmatrix} = \max \begin{bmatrix} 0 + 1.0 \\ 1.3 + 0.5 \\ 2.3 + 0 \end{bmatrix}$$

= \$2.3m (with $x_2*(2) = 2$)

Note that the values of $f_1(1)$ - $f_1(3)$ are given in part (a) above and in Table 13.7(a). The calculations with two stages remaining are summarized in Table 13.7(b).

(c) With three stages remaining:

Equation (13.82) becomes

$$f_3(i_3) = \max_{0 \le x_3 \le 4} \left[c_3(x_3) + f_2(i_3 - x_3 + 3) \right] (\text{for } i_3 = 3,4,5,6)$$

The calculations are summarized in Table 13.7(c).

Table 13.7 Decision tables for reservoir release problem.

(a) With one stage remaining

Initial storage i_1	Release (x_1)					Optimum release $x_1*(i_1)$	Maximum benefit ($M) $f_1(i_1)$
	0	1	2	3	4		
1	0+0	—	—	—	—	0	0
2	—	0.5+0	—	—	—	1	0.5
3	—	—	1.0+0	—	—	2	1.0
4	—	—	—	1.2+0	—	3	1.2
5	—	—	—	—	1.3+0	1	1.3
6	—	—	—	—	1.3+0	1	1.3

(b) With two stages remaining

Initial storage i_2	Release (x_2)					Optimum release $x_2*(i_2)$	Maximum benefit ($M) $f_2(i_2)$
	0	1	2	3	4		
2	0+1.0	1.3+0.5	2.3+0	—	—	2	2.3
3	0+1.2	1.3+1.0	2.3+0.5	3.0+0	—	3	3.0
4	0+1.3	1.3+1.2	2.3+1.0	3.0+0.5	3.3+0	3	3.5
5	0+1.3	1.3+1.3	2.3+1.2	3.0+1.0	3.3+0.5	3	4.0
6	0+1.3	1.3+1.3	2.3+1.3	3.0+1.2	3.3+1.0	4	4.3

(c) With three stages remaining

Initial storage i_3	Release (x_3)					Optimum release $x_3*(i_3)$	Maximum benefit ($M) $f_3(i_3)$
	0	1	2	3	4		
3	0+4.3	1.5+4.0	2.5+3.5	3.3+3.0	3.8+2.3	3	6.3
4	0+4.3	1.5+4.3	2.5+4.0	3.3+3.5	3.8+3.0	3 or 4	6.8
5	0+4.3	1.5+4.3	2.5+4.3	3.3+4.0	3.8+3.5	3 or 4	7.3
6	0+4.3	1.5+4.3	2.5+4.3	3.3+4.3	3.8+4.0	4	7.8

(d) With four stages remaining

Initial storage i_4	Release (x_4)					Optimum release $x_4*(i_4)$	Maximum benefit ($M) $f_4(i_4)$
	0	1	2	3	4		
3	0+7.8	0.2+7.8	0.4+7.3	0.5+6.8	0.6+6.3	1	8.0

(d) *With four stages remaining:*

The initial storage is 3 units and Equation (13.82) becomes

$$f_4(3) = \max_{0 \le x_3 \le 4} \left[c_4(x_4) + f_3(3 - x_3 + 4) \right]$$

The calculations are summarized in Table 13.7(d). To find the optimum release policy over the four seasons we work backwards through Table 13.7 as follows: release 1 unit in winter (leaving a final storage of 6), release 4 units in spring (final storage = 5), release 3 units in summer (final storage =3), and release 2 units in autumn (final storage = 3). The economic benefits are \$8.0M for the year.

In actual reservoir operations the future inflows are usually not known with certainty, but only in a probabilistic sense. In this case an extension of dynamic programming called **stochastic dynamic programming** may be applied. This will not be discussed in this book.

Dimensionality in dynamic programming

The examples considered above both contain only one state variable. In the case of Example 13.6 the state variable is the city that the truck is located in at a particular time. In the case of Example 13.7, the state variable is the volume of water in storage at the start of any stage. Many real problems contain more than one state variable. An example involves the operation of a system of n interconnected reservoirs. Here the state of the system at any time is summarized by the storage in each of the reservoirs and there are n state variables.

An increase in the number of state variables significantly increases the amount of computational effort to solve dynamic programming problems. This has been referred to as the *curse of dimensionality*. It may be illustrated by the reservoir operating problem. Suppose it is desired to apply dynamic programming to a single reservoir. If we divide the volume of the reservoir into 10 discrete levels, there are 10 feasible states at each stage. Suppose we now consider a system of n reservoirs, with the storage in each one being able to take on any one of 10 discrete values. The total number of states at each stage of this problem is 10^n. For example, if n=4 there are 10 000 feasible states at each stage. In practical terms, dynamic programming problems are limited to around four state variables because of the curse of dimensionality.

13.15 HEURISTIC OPTIMISATION METHODS

In recent years quite a deal of research has been carried out into a new class of optimisation techniques that might loosely be called “heuristic optimisation algorithms”. These include techniques such as genetic algorithms, the evolutionary algorithm, differential evolution, ant colony optimisation, particle swarm optimisation and shuffled complex evolution (Michalewicz and Fogel, 2004). All of these methods have the following features in common:

1. They deal with a population of solutions rather than a single solution at any one time.
2. They are effectively "guided search methods" that utilise a simulation model of the system under study. The model can include any non-linear or logical constraints or variables. Thus any system that can be modelled can be optimised using one of these techniques.
3. They usually involve some random processes, so that they may be considered to be "stochastic" optimisation methods.
4. They usually do not reach a single optimum solution but progressively improve the population of solutions that they are working with over time.
5. One cannot prove that the true optimal solution has been obtained. Effectively the methods identify "near-optimal" solutions.

In this text only the genetic algorithm (GA) technique will be considered in detail as this is a commonly used method that has proven to be very efficient. This is discussed in the next section.

Genetic Algorithm Optimisation

Holland (1975) was one of the early advocates of genetic algorithms, although they could be viewed as descended from the work of Box (1957) who used evolutionary techniques for optimisation. The publication of the book by Goldberg (1989) inspired a large number of engineering applications of GA optimisation. The applications include structural optimisation (Goldberg and Samtani, 1986), optimising the operation of pumps in gas pipelines (Goldberg and Kuo, 1987), control system optimisation for aerospace applications (Krishnakumar and Goldberg, 1990) and the optimum design of water supply networks (Simpson et al., 1994).

In essence GAs are a set of guided search routines that work in conjunction with a mathematical or computer simulation model of a system in order to optimise aspects of the system's design or operation. Some distinguishing features of GAs compared to the traditional optimisation techniques such as linear programming and dynamic programming considered earlier in this chapter are as follows (Simpson et al., 1994):

1. GAs work directly with a population of solutions rather than a single solution. This population is spread throughout the solution space, so the chance of reaching the global optimum is increased significantly;
2. GAs deal with the actual discrete sizes available so that roundoff of continuous variables is not required;
3. Because GAs work with a population of solutions, they identify a number of near-optimal solutions. These solutions could correspond to quite different configurations that may have advantages in terms of non-quantified objectives such as environmental or social objectives; and
4. GAs use only information about the objective or fitness function and do not require the existence or continuity of derivatives of the objectives with respect to the decision variables.

GAs work by analogy to population genetics and involve operators such as selection, crossover and mutation. Unlike techniques such as linear programming, they do not necessarily converge to the global optimal solution. However, because they work in conjunction with a simulation model they can handle any non-linear, discontinuous or logical set of objective functions or constraints. In essence, any system that can be simulated on a computer can be optimised using GAs. Holland's Schema Theorem (Holland, 1975) suggests that increasingly fit solutions will be present in exponentially increasing numbers in succeeding generations in a GA run.

Application of Genetic Algorithms

Combinatorial optimisation was introduced in Section 13.2. It was noted that some engineering problems involve extremely large combinations of decision variables. The number of combinations can be so great that it is not feasible to simulate all alternatives in a reasonable amount of computer time. Genetic algorithms can be applied to these large combinatorial optimisation problems. An example of such problems is the New York Tunnels problem that was first studied by Schaake and Lai (1969). The basic layout of the system is shown in Figure 13.23.

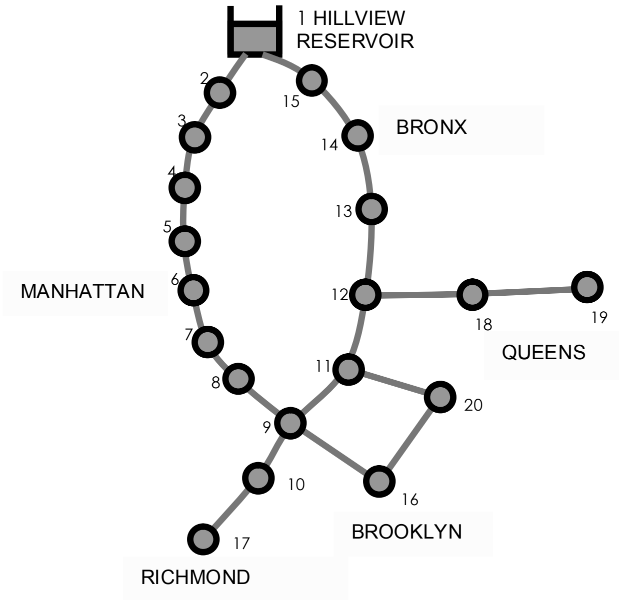


Figure 13.23 Layout of the New York Water Supply Tunnels

Water is supplied from Hillview Reservoir through 21 major tunnels ranging in size from 1.524 m to 5.182 m in diameter. In Figure 13.23 the lines represent the various tunnels and the nodes represent junctions in the tunnels where water is distributed to many smaller pipes in the network. These smaller pipes are not shown. At the time of the study, it was anticipated that the system would be unable to meet forecast increases in demand while supplying adequate pressures. Hence the problem was to identify which tunnels should be duplicated and with what diameter tunnels in order to meet the forecast increased demand.

The tunnels can only be drilled in discrete sizes of which 16 options (ranging from no duplication to duplication with a tunnel 5.182 m in diameter cover the practical sizes available for this problem. In practice this is a small network as the distribution system for a large city could contain more than 100,000 pipes. Nonetheless the number of options is significant and equals 16^{21} or 1.934×10^{25}. Such networks are usually designed using a trial-and-error procedure whereby the engineer chooses a trial design and then uses a hydraulic analysis computer model to estimate the pressures throughout the network under the design demand conditions. If the pressures are inadequate, some pipe (or tunnel) sizes are increased in the model and it is rerun. If the pressures are higher than the minimum requirements, some pipe sizes are reduced in the model and it is rerun. Eventually, the engineer decides that the design is reasonable i.e. provides the desired pressures at a reasonable cost.

Although this is a relatively small network, identification of the minimum cost solution is a non-trivial task. For example, if an evaluation of each of the 1.934×10^{25} solutions required 1 millisecond of computer time, the total run time would be 6.13×10^{14} years. This can be compared with the estimated life of the universe of 15×10^{9} years! In practice of course, engineers can eliminate many of the infeasible solutions using experience. Suppose that we can eliminate 99.99% of the potential solutions and then enumerate the remaining 0.01%. Unfortunately the computer run time will still be 6.13×10^{10} years or 4 times the life of the universe. Computer speeds are doubling every 18 months or so. If that trend continues, the computer run time to completely enumerate the New York tunnels problem will be down to 1 day in about 60 years time. Until that time, we need to rely on other techniques to identify optimal or near-optimal solutions to problems like the New York Tunnels problem.

It should be noted that the New York water supply system has changed considerably since the study by Schaake and Lai in 1969. However, it has been used by a number of researchers over the last 48 years as a benchmark problem to test their pipe network optimisation routines (Dandy et al., 1996). The best solution found so far is that by Maier et al. (2003) using the technique of Ant Colony Optimisation.

Representation of solutions

Many variants of GAs have been developed over the last 15 years or so. In this Section only a simple GA will be considered. A GA operates on a string of numbers that represents a set of values of the decision variables. Each string is solution to the optimisation problem. By analogy to population genetics, this string is sometimes called a chromosome while the individual numbers in the string are called genes.

For the New York tunnels problem, the string would consist of 21 integers, each one being between 0 and 15 as shown in Figure 13.24.

[3|12|0|5|0|1|0|0|9|3|0|7|5|11|0|2|2|0|7|10|0]

Figure 13.24 A GA String for the New York Tunnels problem

The first integer in the string represents the size of the duplicate tunnel for tunnel # 1, the second integer the duplicate for tunnel #2, and so on. The integers are decoded into actual pipe sizes using Table 13.8. Note that the problem has been converted into SI units from the original USA customary units of feet and inches.

Table 13.8 Coding of New Tunnels and their costs for the New York Tunnels problem (adapted from Dandy et al. 1996)

GA code	Tunnel Diameter (m)	Cost ($/m)
0	0	0
1	0.914	306.8
2	1.219	439.6
3	1.524	577.4
4	1.829	725.1
5	2.134	876.0
6	2.438	1036.7
7	2.743	1197.5
8	3.048	1368.1
9	3.353	1538.7
10	3.658	1712.6
11	3.962	1893.0
12	4.267	2073.5
13	4.572	2260.5
14	4.877	2447.5
15	5.182	2637.8

The solution given by the string in Figure 13.24 therefore represents duplicating Tunnel #1 with a tunnel of 1.524 m diameter, duplicating Tunnel #2 with a tunnel of diameter 4.267 m, not duplicating Tunnel #3, and so on.

Fitness function

The objective function in GA optimisation is called the fitness function. GA optimisation is usually aimed at maximising the fitness function, so if the problem involves minimisation, the fitness function is written as minus the value of the objective function or as the inverse of the objective function. The later transformation can only be used if the objective function is strictly positive.

Steps in the GA

The steps involved in the operation of a simple GA are given below:

1. generation of the initial population;
2. computation of the objective function for each solution in the population;

3. evaluation of the performance of each solution in the population relative to the constraints;
4. computation of the penalty cost for not meeting the constraints;
5. computation of the fitness function for each solution;
6. check for convergence of the population. If convergence has occurred, then stop; otherwise continue.
7. selection of a set of parent strings for the next generation;
8. crossover of pairs of parents; and
9. mutation of selected strings. Return to step 2.

Each of these steps will be explained in more detail in relation to the New York Tunnels problem. Firstly, some additional data are provided for the problem. Table 13.9 contains information on the design demands and elevations of each of the nodes.

The design demands are for a future target year, in this case 1994 (25 years on from the time of the study). Table 13.10 contains information about the lengths and diameters of the existing tunnels in the system (as they were in 1969).

The New York Tunnels system needs to be designed to meet the design demands at minimum cost while ensuring that the minimum total heads at all nodes (given in Table 13.9) are satisfied. The difference between the total head and the elevation at a node is the pressure head at that node. A minimum level of pressure head is required to ensure reasonable levels of service for consumers. The total heads at each node for any particular design can be computed using a combination of the continuity and head loss equations for the system. The details of computing the total heads are given in Appendix 13B. The Hazen-Williams coefficient (defined in Appendix 13B) for all tunnels is 100.

Table 13.9 Nodal Data for the New York City Water Supply Tunnels (adapted from Dandy et al., 1996)

Node	Demand (m^3/s)	Minimum total head (m)
1	0	91.44 (actual)
2	2.710	77.72
3	2.710	77.72
4	2.587	77.72
5	2.587	77.72
6	2.587	77.72
7	2.587	77.72
8	2.587	77.72
9	4.986	77.72
10	0.029	77.72
11	4.986	77.72
12	3.435	77.72
13	3.435	77.72
14	2.710	77.72
15	2.710	77.72
16	4.986	79.25
17	1.686	83.15
18	3.435	77.72
19	3.435	77.72
20	4.986	77.72

As this is a minimisation problem and the cost is strictly positive, the fitness function used will be the inverse of the total cost. The total cost will be the sum of the cost of constructing duplicate tunnels plus the penalty cost (if any) of not satisfying the minimum head constraints at nodes.

The steps involved in applying GA optimisation to the New York Tunnels problem are described below.

1. Generation of the initial population.

As stated previously, a GA works with a population of solutions at any one time. The population size is one of the parameters that must be chosen for a GA. It is typically in the range 20–1000. The initial population is usually generated randomly, so for the New York Tunnels each string would consist of a set of 21 randomly generated integers in the range 0–15. Each integer would have a specified probability of being selected. These probabilities are usually equal, but they could be selected to have a bias towards zero (i.e. no duplication) in the case of the New York Tunnels.

Table 13.10 Data for the existing New York Tunnels (adapted from Dandy et al., 1996)

Tunnel	Start Node	End Node	Length (m)	Existing Diameter (m)
[1]	1	2	3535.7	4.572
[2]	2	3	6035.0	4.572
[3]	3	4	2225.0	4.572
[4]	4	5	2529.8	4.572
[5]	5	6	2621.3	4.572
[6]	6	7	5821.7	4.572
[7]	7	8	2926.1	3.353
[8]	8	9	3810.0	3.353
[9]	9	10	2926.1	4.572
[10]	11	9	3413.8	5.182
[11]	12	11	4419.6	5.182
[12]	13	12	3718.6	5.182
[13]	14	13	7345.7	5.182
[14]	15	14	6431.3	5.182
[15]	1	15	4724.4	5.182
[16]	10	17	8046.7	1.829
[17]	12	18	9509.8	1.829
[18]	18	19	7315.2	1.524
[19]	11	20	4389.1	1.524
[20]	20	16	11704.3	1.524
[21]	9	16	8046.7	1.829

2. Computation of the objective function for each solution in the population.

The objective function can be now computed for each solution in the initial population. In the case of the New York Tunnels, this equals the cost of the solution and can be computed by knowing the diameters of the duplicate tunnels (from the values in the string), their lengths (given in Table 13.10) and the costs per unit length (given in Table 13.9).

3. Evaluation of the performance of each solution in the population relative to the constraints.

The hydraulic performance of each design can be evaluated by calculating the total head at each node in the network by solving the non-linear equations given in Appendix 13B using a hydraulic computer model or by other means.

4. Computation of penalty cost for not meeting the constraints.

If the total head at any node is below the minimum value (given in Table 13.9) a penalty cost is added, for example $16 million per metre of total head below the minimum required value. This is an arbitrary high figure chosen to penalise solutions that do not satisfy the minimum pressure constraints. If several nodes have total heads below the minimum values, the penalty will be based on the node with the largest deficit.

5. Computation of the fitness function for each solution.

The objective function value (i.e. the cost of the solution computed in step 2) is added to the penalty cost (step 4) to give the total cost for each design. The fitness function is the inverse of the total cost as this is to be minimised.

6. Check for convergence of the population. If convergence has occurred, then stop; otherwise continue.

Unlike a number of other optimisation techniques discussed earlier in this Chapter, genetic algorithms do not have an easy way of determining if the global optimum solution has been achieved. In practice the algorithm is run for a specified number of generations or until little improvement is achieved in the best solution over a number of generations. If one of these criteria is met, the algorithm is stopped, otherwise proceed to step (7).

7. Selection of a set of parent strings for the next generation.

By analogy to natural selection, not all solutions will be parents (i.e. produce offspring) in the next generation. Using the principle of "survival of the fittest", the GA will select solutions with the highest fitness values to be parents for the next generation. A common way to carry out this process is to use tournament selection. In tournament selection, two solutions are selected at random from the current population, they are compared in terms of their fitness and the string with the higher fitness becomes a parent for the next generation. This process is repeated (without replacing the first two strings) to give another parent. The process continues until all strings have been involved in tournaments. At this stage $N/2$ parents have been selected, where N is the population size. The full process is carried out a second time starting with the full population from the current generation, so that a total of N parents have been identified.

8. Crossover of pairs of parents.

Parents are "mated" to produce the next generation. This is carried out by randomly selecting a pair of parents from the N available. Crossover of the genetic material of the parents will occur with a probability p_c. A uniformly distributed random number between zero and one is selected. If this number is less than or equal to p_c, crossover occurs and some exchange occurs in the parental genetic material passed on to the offspring. This is simulated by determining a point to cut the two strings as shown in Figure 13.25(a).

The point at which the strings will be cut is determined randomly. With 21 numbers (genes) in the string, there are 20 possible places to cut the strings. An integer between 1 and 20 is chosen randomly and is used to identify where the strings will be cut. The right hand tails are then switched between the two strings. Suppose the random number chosen is 11. The offspring strings are shown in Figure 13.25(b).

Cut point for crossover

Parent string #1: [3|12|0|5|0|1|0|0|9| 3 |0|7|5|11|0|2|2|0| 7 |10|0]
Parent string #2: [8| 0 |0|3|2|0|0|0|0|15|0|0|4| 0 |0|5|1|9|11| 0 |0]

(a) parent strings before crossover

Cut point for crossover

Offspring string #1: [3|12|0|5|0|1|0|0|9| 3 |0|0|4| 0 |0|5|1|9|11| 0 |0]
Offspring string #2: [8| 0 |0|3|2|0|0|0|0|15|0|7|5|11|0|2|2|0| 7 |10|0]

(b) offspring strings after crossover

Figure 13.25 Crossover of two strings

If crossover does not occur, the two offspring strings are identical to their parents, so the two parent strings are passed into the next generation. This process is repeated for pairs of strings that are selected (without replacement) from the parent population, so that a population of offspring strings is formed.

9. Mutation of selected strings.

In nature, there is some chance that mutation will occur and the offspring will have some genetic differences from their parents. This serves to keep some diversity in the population and allows all species to adapt to a changing environment. The GA process allows some small percentage of strings to mutate every generation. This percentage is determined by the GA parameter called the probability of mutation, p_m. After the crossover process, each string is tested to see if mutation occurs. A uniformly distributed random number between zero and one is chosen. If it is less than or equal to $p_{m,}$ mutation occurs. One of the numbers in the string will change to a randomly chosen value. The number that mutates is chosen at random by selecting an integer random number between one and *l* (where *l* = the length of the string). The new value is randomly chosen over the range of possible values for this location in the string. In the case of the New York Tunnels, the new value will be randomly chosen between 0 and 15. Now return to step 2.

Convergence of the GA is illustrated in Figure 13.26 which shows the minimum cost solution in each generation. It can be seen from this figure that the GA converges rapidly for a start, flattens out and then only improves as a result of occasional mutations that identify areas of better solutions.

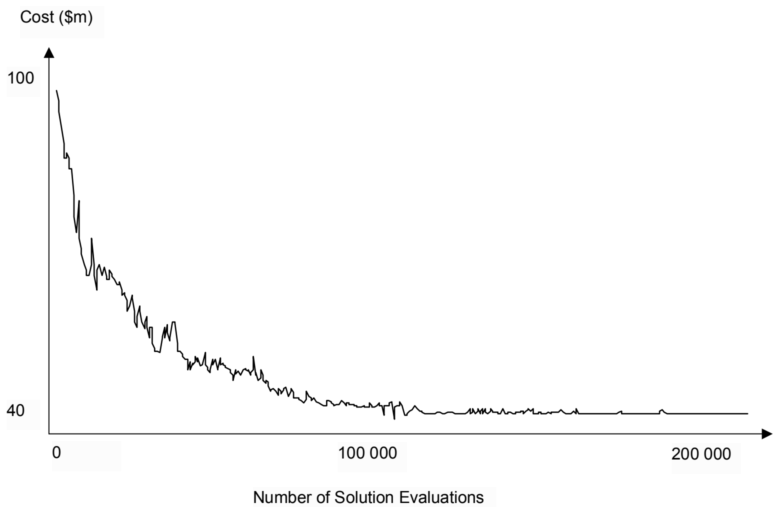


Figure 13.26 Convergence of the GA for the New York tunnels problem (adapted from Dandy et al. 1994)

The solution identified by Dandy et al. (1996) is given in Table 13.11 and has a total cost of $38.80 m compared to the original design of Schaake and Lai (1969) which had a cost of $78.09 m. Only the tunnels listed in the table are to be duplicated. Schaake and Lai (1969) used linear programming to solve the problem but this required assumptions to be made about the total head at each node. This method is unlikely to lead to the global optimal solution and the Schaake and Lai (1969) solution involved much more duplication of tunnels than is necessary. Over the years a number of other researchers have applied various optimisation techniques to solve this problem. The best solutions have been found using evolutionary algorithms such as genetic algorithms and ant colony optimisation.

Table 13.11 Solution to the New York Tunnels problem by Dandy et al. (1996)

Duplicate Tunnel Number	Duplicate Diameter (m)
[15]	3.048
[16]	2.134
[17]	2.438
[18]	2.134
[19]	1.829
[21]	1.829

Choice of GA Parameters

The following parameters need to be specified for a GA:

- population size;
- probability of crossover;
- probability of mutation;
- tournament size (which can be more than two); and
- maximum number of generations

These values need to be chosen based on experience in the particular problem regime. Indicative ranges for each parameter are given in Table 13.12.

In addition, penalty costs need to be chosen for each set of constraints. It is difficult to give general advice for these values because they depend on the units of the objective function and constraints. In general the penalties need to be large enough to ensure that the constraints are satisfied, but small enough to ensure that the optimum solution can be approached from both the infeasible as well as the feasible region. In practice, the GA is usually run a number of times using different combinations of parameters and a range of starting seeds for the sequence of random numbers and the best overall solution is identified.

Table 13.12 Indicative ranges for values of the GA parameters

GA parameter	Symbol	Indicative range
Population size	N	50–1000
Probability of crossover	p_c	0.7–1.0
Probability of mutation	p_m	0.001–0.02
Tournament size	T	2–4
Maximum number of generations	G	500–2000

13.16 SUMMARY

Optimisation is a process of attempting to find the “best” solution for an engineering planning or design problem. The “best” solution is the one that best achieves the defined objectives for the particular system while satisfying all of the constraints. There are a number of approaches to optimisation including subjective, combinatorial and analytical optimisation. Subjective optimisation is based on engineering judgement and experience and is used in practice when the problem is too complicated to allow combinatorial or analytical methods to be used.

Combinatorial optimisation involves evaluating and comparing all (or a large number) of the possible combinations of solutions for a particular problem. Complete enumeration of all alternatives can be used when there are a small number of possible combinations, otherwise heuristic methods (such as genetic algorithms) can be applied to identify near-optimal designs.

Analytical optimisation techniques can be used where the objective(s) and constraints for the system can be written in mathematical form. There are a large number of techniques that can be used to solve analytical optimisation problems. In this text the following techniques have been described in detail: linear programming, differential calculus, separable programming and dynamic programming.

The choice of the most appropriate optimisation technique to use for a particular problem is a skill based on experience and a good knowledge of the problem and the available techniques.

PROBLEMS

13.1 Solve the following LP problems graphically:

(a) Max $Z = 3x_1 + 5x_2$
subject to $2x_1 + x_2 \leq 12$
$4x_1 + 5x_2 \leq 40$
$x_1, x_2 \geq 0$

(b) Min $Z = 10x_1 + 8x_2$
subject to $x_1 \geq 4$
$x_1 + 3x_2 \geq 12$
$x_1 + x_2 \geq 8$
$x_1, x_2 \geq 0$

(c) Max $Z = 4x_1 - 3x_2$
subject to $x_1 \leq 10$
$2x_1 + 3x_2 \geq 12$
x_1, x_2 unrestricted in sign.

13.2 Formulate the following LP problems into canonical and standard form:

(a) Min $Z = 4x_1 + 2x_2$
subject to $2x_1 + 3x_2 \geq 9$
$3x_1 + x_2 \geq 6$
$x_1, x_2 \geq 0$

(b) Max $Z = x_1 + 2x_2 + 4x_3$
subject to $x_1 + x_3 \leq 5$
$2x_1 - x_2 \geq 4$
$x_1, x_3 \geq 0$ x_2 unrestricted in sign.

13.3 Portion of a road network is shown in Figure 13.27. The travel times and capacities in both directions for all roads are given in Table 13.13. During peak conditions, 5000 vehicles per hour are travelling from node 1 to node 6 and 4000 vehicles per hour are travelling from node 4 to node 2.

(a) Formulate an LP problem which could be used to find the volume of traffic in each direction on each road during peak conditions.
(b) Solve the LP problem using a standard computer package.

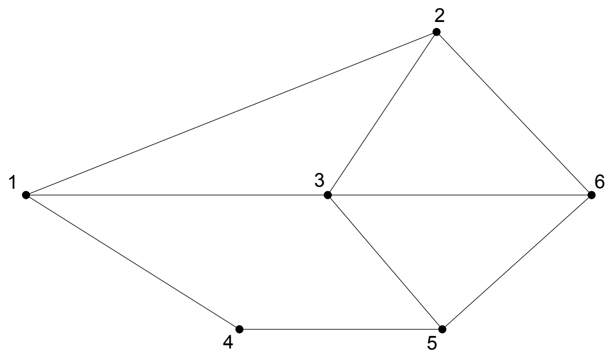


Figure 13.27 Portion of a road network.

Table 13.13 Travel times and capacities for a road network.

Road	Travel time (minutes)	Capacity (vehicles/h)
1–2	15	3000
1–3	10	4000
1–4	8	1000
2–6	6	3000
3–2	7	2000
3–5	6	3000
3–6	5	2000
4–5	9	3000
5–6	10	4000

13.4 A reservoir and pipe network is shown in Figure 13.28. The hydraulic grade line at locations C and D must be at or above RL 20 m and 0 m respectively. The elevation of the hydraulic grade line at B is to be determined. Three sizes of pipe are available according to Table 13.14.

Table 13.14 Pipe sizes and costs.

Pipe diameter (mm)	Cost ($/m length)
600	180
750	320
900	500

Each of the sections AB, BC, and BD may be made up of lengths of any or all pipe sizes. However, only 1000 m of 750 mm diameter pipe is available in time for the job.

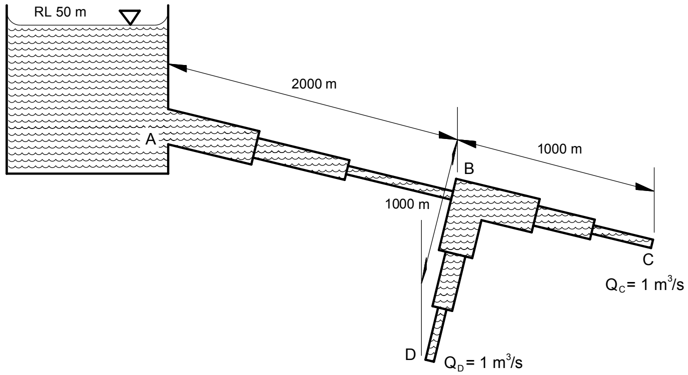


Figure 13.28 A reservoir and pipe network.

The head loss in any length of pipe is given by:

$$h_l = 0.001\frac{V^2 L}{D}$$

where h_l *is* the head loss (m), V is the average velocity in pipe (m/s), L is the pipe length (m), and D is the pipe diameter (m). Ignore the velocity head and head losses at entrance, exit, contraction, expansion, and bends.

(a) Set up an LP model which could be used to determine the lengths of pipe of each diameter to be used in each of the sections AB, BC, and BD so as to minimise the total cost.

(b) Solve the model using a suitable computer package.

13.5 A river receives organic waste from the two factories shown in Figure 13.29 The organic waste load is 1000 kg/day from the first factory and 1400 kg/day from the second. The streamflows (including waste input) at locations 1 and 2 are 3 m^3/s and 4 m^3/s (respectively).

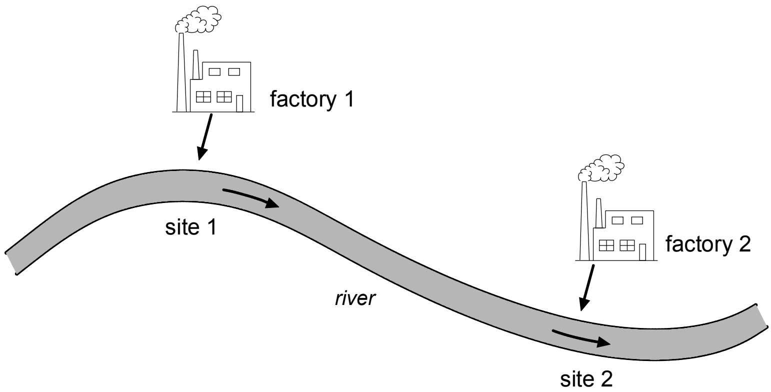


Figure 13.29 A river receiving organic waste from two factories.

It is proposed to install waste treatment plants at each factory so as to reduce the maximum waste concentration in the river to 0.002 kg/m^3 at all locations. The cost of waste removal is \$0.30/kg at factory 1 and \$0.20/kg at factory 2. A maximum of 90% of the waste from either factory can be removed. Waste decomposition in the river may be ignored.

Formulate an LP model which could be used to find the minimum cost combination of waste treatment plants at factories 1 and 2 such that the maximum waste concentration is not exceeded. Solve the model graphically.

13.6 Write the duals to Problems 13.1(a)-(c) and 13.2(a)-(b).

13.7 Carry out the following:

(a) Write the dual to the LP model formulated in Problem 13.5.
(b) From the graphical solution of the primal, determine the optimum values of the dual variables. (Hint: the dual variables will be zero for those primal constraints which have some slack or surplus at the optimum solution. Why?)
(c) Verify the optimum values of the dual variables found in part (b) by solving the dual problem using the simplex method.
(d) Use the value of the dual variables to estimate the reduction in treatment cost if the maximum allowable waste concentration at site 2 is increased to 0. 0022 kg/m^3.

13.8 An optimal policy is required to maximise the net daily benefits due to the operation of a reservoir from which water may be extracted for both power generation and irrigation purposes.

Each 1000 m^3 of water released through the power station results in the generation of 150 kWh of electrical energy, with all of the water passing from the turbines into the river downstream of the dam. The net benefit to the community of power generation at this site is 10 c/kWh for the first 3000 kWh, 6 c/kWh for the next 3000 kWh and 4 c/kWh thereafter (up to a maximum of 15 000 kWh).

Each 1000 m^3 of water released for irrigation purposes produces a net benefit to the community of $10. In addition, 30% of the water released for irrigation is returned to the river downstream of the dam.

The total volume of water available for power generation and irrigation purposes is 100 000 m^3/day. In order to maintain water quality standards in the river downstream of the dam, a minimum flow of 10,000 m^3/day must be maintained.

Formulate the problem of finding the daily volumes of water to be used for power generation and irrigation as an LP problem using the separable programming technique.

13.9 Rework Problem 13.5 assuming that the cost of waste removal is not constant at each site but is given in Table 13.15. Linear interpolation may be used for intermediate values of plant capacity. Why is it possible to use separable programming for this problem?

Table 13.15 Costs of waste removal

Capacity of plant (pollutant removal kg/day)	Cost (x10^3$) of installing a plant at	
	Industry 1	Industry 2
0	0	0
500	100	120
1000	250	300
1400	—	500

13.10 A trucking company wants to transport goods from city 1 to city 11 at minimum cost. The cost of travel on each link (in dollars) is shown in Figure 13.30. Find the minimum cost route from A to B. (Note that it may be necessary to introduce dummy states in order to apply dynamic programming to this problem.)

13.11 A quarry produces crushed rock for road-making. It can produce up to 5000 m^3 per month. The quarry's orders and cost of production per 1000 m^3, for the next four months are given in Table 13.16

The quarry may store up to 4000 m^3 of crushed rock from one month to the next at a cost of $50/1000 m^3.

Use dynamic programming to find the quarry's minimum cost production schedule over the four months assuming it starts and finishes with no crushed rock in storage.

Table 13.16 Orders and production costs.

Month	1	2	3	4
Orders (x 1000 m^3)	2	4	4	3
Production cost for the first 1000 m^3 ($)	500	550	600	550
Production cost for each additional 1000 m^3 ($)	300	250	350	300

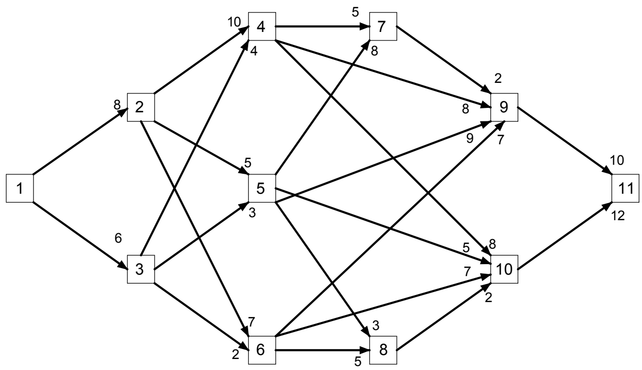


Figure 13.30 A highway network.

13.12 Water is pumped from a river into a reservoir and is used to supply water to a nearby city. The maximum volume of water which can be pumped into the reservoir in any month is 30 000 ML.

It is desired to determine the minimum-cost pumping schedule over the next four months so as to meet the known demands of the city. The demands and pumping costs are given in Table 13.17.

Table 13.17 Demands and pumping costs.

Month	Demand (ML)	Cost ($M) of pumping a volume of:			
		0	10 000 ML	20 000 ML	30 000 ML
1	20 000	0.0	1.0	2.5	3.0
2	40 000	0.0	1.5	2.7	3.3
3	30 000	0.0	1.2	2.0	3.0
4	10 000	0.0	0.9	2.6	3.5

The reservoir contains a volume of 30 000 ML, at the start of the first month and must contain at least 20 000 ML, at the end of the fourth month. Evaporation and seepage losses are negligible.

Using dynamic programming and discrete volumetric units of 10 000 ML, find the minimum-cost set of monthly volumes to be pumped to the reservoir over the period.

13.13 The expansion of a wastewater collection system for a major city is to be designed. It consists of 20 major pipelines (each with 8 possible sizes), 50 smaller pipelines (6 possible sizes each), 12 pumps (12 possible sizes each) and 6 detention storages (5 possible sizes each).

(a) What is the number of possible combinations for this design?
(b) How does this compare with the estimated number of atoms in the Universe (10^{75})?
(c) Based on the comparison in part (b), comment on the difficulty of finding the true optimum solution to this problem.
(d) If it takes 0.001 seconds to simulate the performance of each solution to this problem using an appropriate simulation package, how long would it take to simulate all of the possible combinations?

13.14 In order to demonstrate the operation of a simple genetic algorithm, consider the problem of minimising the following function:

$$Z = 0.5x_1 + x_2 + 2x_3 + 3x_4 + 5x_5$$

where all x_i *(i = 1., ...5)* equal either 0 or 1.

A binary string of length five can represent each solution to this problem. The optimum solution to the problem should be obvious.

(a) Using a random number generator or table of random numbers, generate a starting population of 20 solutions and determine the value of the objective function of each one.
(b) Using tournament selection, a crossover probability of 0.7 and mutation probability of 0.02 per bit, develop a further population of 20 solutions from the initial population.
(c) Evaluate the objective function for each new solution and compare the average value and the lowest value of the objective function for each generation. Comment on the results.
(d) Repeat this process for another four generations and verify that the best solution is improving over the generations.
(e) How would you modify the genetic algorithm solution process if the following constraint needed to be included: $x_1 + x_2 + x_3 + x_4 + x_5 \geq 3$?

13.15 The westbound portion of a proposed new freeway in a city in the USA is shown in Figure 13.31.

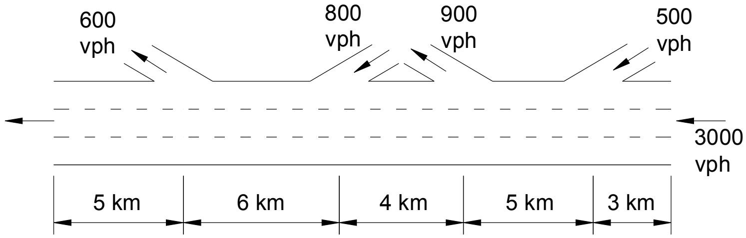


Figure 13.31 Portion of a proposed new freeway (not to scale)

The average travel time on each section of the freeway is given by the following equation:

$$t_i = t_{i0}[1 + (V_i / C_i)^4]$$

where t_i = the travel time on section i (minutes); t_{0i} = the "free speed" travel time on section i (minutes) based on a speed of 100 km /hour; V_i = volume of traffic on section i (vehicles per hour); and C_i = the nominal "capacity" of section i (vehicles per hour).

The nominal capacity per lane is 1200 vehicles per hour and the cost to build 1 km of freeway is \$1m per lane. There is a minimum of 2 lanes and a maximum of 6 lanes per section of freeway.

(a) Write the objective function and constraints for an optimisation model that can be used to find the optimal number of lanes for each section of the freeway so as to minimise the total travel time by all vehicles on the freeway subject to a budget of \$70m.
(b) Outline how this problem could be solved using the genetic algorithm (GA) technique. Describe the structure of a typical string and outline the steps involved in running the GA.

REFERENCES

Beightler, C., Phillips, D. & Wilde, D., 1979, *Foundations of Optimisation,* 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall).

Box, G.E.P., 1957, Evolutionary operation: a method for increasing industrial productivity, *Applied Statistics*, 6(2), 81–101.

Dandy, G.C., Simpson, A.R. and Murphy, L.J., 1993, A review of pipe network optimisation techniques. *Proceedings of Watercomp 93*, (Melbourne, Institution of Engineers, Australia) National Conference Publication No 93/2, 373–383.

Dandy, G.C., Simpson, A.R. and Murphy, L.J., 1996, An improved genetic algorithm for pipe network optimisation, *Water Resources Research*, 32(2), 449–458.

deNeufville, R. & Stafford, J.H., 1971, *Systems Analysis for Engineers and Managers* (New York: McGraw-Hill).

Goicoechea, A., Hansen, D.R. & Duckstein L., 1982, *Multiobjective Decision Analysis with Engineering and Business Applications* (New York: Wiley).

Goldberg, D. E., 1989, *Genetic Algorithms in Search, Optimisation, and Machine Learning* (Addison-Wesley Publishing Company, Inc.)

Goldberg, D.E. and Samtani, M.P., 1986, Engineering optimization via genetic algorithms. *Proceedings 9th Conference on Electronic Computation*, ASCE, New York, NY, 471–482.

Goldberg, D.E. and Kuo, C.H., 1987, Genetic algorithms in pipeline optimization, *Journal of Computing in Civil Engineering*, 1(2), 128–141.

Holland, J.H., 1975, *Adaptation in natural and artificial systems,* University of Michigan Press, Ann Arbor

Krishnakumar, K. and Goldberg, D.E., 1990, Control system optimization using genetic algorithms. *Proceedings AIAA Guidance, Navigation and Control Conference*, American Institute of Aeronautics and Astronautics (AIAA).

Maier H.R., Simpson, A.R., Zecchin, A.R., Foong, W.K., Phang, K.Y., Seah, H.Y. and Tan, C.L., 2003, Ant colony optimisation for design of water distribution systems, *Journal of Water Resources Planning and Management, ASCE,* 129(3), 200–209.

Meredith, D.D., Wong, K.W., Woodhead, R.W. & Wortman R.H., 1985, *Design and Planning of Engineering Systems,* 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall).

Michalewicz, Z. and Fogel, D.B., 2004, *How to Solve It: Modern Heuristics*, 2nd ed. (Berlin: Springer).

Ormsbee, L.E. and Wood, D.J., 1986, Hydraulic design algorithms for pipe networks, *Journal of Hydraulic Engineering*, ASCE, 112(12), 1195–1207.

Ossenbruggen, P.J. 1984. *Systems Analysis for Civil Engineers* (New York: Wiley).

Schaake, J., and Lai, D. 1969. Linear Programming and Dynamic Programming Applications to Water Distribution Network Design. *Report 116*, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Shamir, U. and Howard, C.D.D., 1968, Water distribution systems analysis. *Journal of Hydraulics Division*, ASCE, 94(HY1), 219–234.

Simpson, A.R., Dandy, G.C. and Murphy, L.J., 1994 Genetic algorithms compared to other techniques for pipe optimisation. *Journal of Water Resources Planning and Management, ASCE*, 120 (4), 423–443.

Smith, A.A., Hinton, E. and Lewis R.W., 1983, *Civil Engineering Systems. Analysis and Design.* (New York: Wiley).

Stark, R.M. & Nicholls, R.L., 1972, *Mathematical Foundations for Design. Civil Engineering Systems.* (New York: McGraw–Hill).

Streeter, V.L. and Wylie, E.B., 1979, Fluid mechanics. (McGraw–Hill, New York).

Taha, H.A., 2017, *Operations Research. An Introduction,* 10th ed. (Upper Saddle River , NJ: Pearson Education).

APPENDIX 13A: NECESSARY AND SUFFICIENT CONDITIONS FOR THE SOLUTION OF UNCONSTRAINED MULTIVARIATE OPTIMISATION PROBLEMS

The necessary and sufficient conditions for an unconstrained multivariate optimisation problem are given below. For convenience define $\underline{\mathbf{X}} = (x_1, x_2, \ldots, x_n)$, then condition 13.55 becomes:

$$\text{Max } Z = f(\underline{\mathbf{X}}) \tag{13.85}$$

As in the single-variable problem, differential calculus can be used to solve this problem provided $f(\underline{\mathbf{X}})$ and its first- and second-order partial derivatives exist and are continuous for all values of $\underline{\mathbf{X}}$.

A **necessary** condition for a particular solution $\underline{\mathbf{X}}^{\mathbf{o}}$ to be a maximum of $f(\underline{\mathbf{X}})$ is:

$$\left.\frac{\partial f(\underline{X})}{\partial x_j}\right|_{\underline{X} = \underline{X}^{o}} = 0 \quad (j = 1, \ldots, n) \tag{13.86}$$

The solutions to Equation (13.86) are called stationary points and include maxima, minima, points of inflection, and saddlepoints. In order to determine which of the stationary points are maxima (or minima) it is necessary to examine the second-order partial derivatives of $f(\underline{\mathbf{X}})$. These are contained in the Hessian matrix, $\underline{\mathbf{H}}$, which is defined as follows:

$$\underline{\mathbf{H}} = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

A **sufficient** condition for a stationary point, $\underline{\mathbf{X}}^{\mathbf{o}}$ to be a minimum is that the Hessian matrix evaluated at $\underline{\mathbf{X}}^{\mathbf{o}}$ is **positive definite.** For $\underline{\mathbf{X}}^{\mathbf{o}}$ to be a maximum, the Hessian matrix evaluated at $\underline{\mathbf{X}}^{\mathbf{o}}$ should be **negative definite.**

A matrix is positive definite if all of its **principal minor determinants** λ_k $(k = 1, \ldots, n)$ are positive.

The kth principal minor determinant of $\underline{\mathbf{H}}$ is defined as:

$$\lambda_k = \begin{vmatrix} h_{11} & h_{12} & \ldots & h_{1k} \\ h_{21} & h_{22} & \ldots & h_{2k} \\ . & . & & . \\ . & . & & . \\ h_{k1} & h_{k2} & \ldots & h_{kk} \end{vmatrix}$$

A matrix is negative definite if its kth principal minor determinant has the sign $(-1)^k$ for $k = 1, 2, \ldots, n$.

APPENDIX 13B: HYDRAULIC ANALYSIS OF PIPE NETWORKS

This appendix explains how flows and pressure heads are determined in a pipe network. It is based on Dandy et al (1993).

Consider a general pipe network with *NP* pipes, *NJ* junction nodes (excluding fixed grade nodes or reservoirs) and *NF* fixed grade reservoirs. The following are assumed to be known:

1. the demands at all nodes ($q_1, q_2, \ldots, q_{NJ}$);
2. the diameters of all pipes ($D_1, D_2, \ldots, D_{NP}$); and
3. the total head at one or more nodes (e.g. at fixed grade reservoirs).

The total head at each node is the sum of the elevation head and the pressure head at the node. It is assumed that velocity heads, minor losses and losses at junctions are negligible. The analysis described in this appendix will be a steady state analysis assuming that all demands are constant.

The basic unknowns are the discharge in each pipe, ($Q_1, Q_2, \ldots, Q_{NP}$) and the total head at each node ($H_1, H_2, \ldots, H_{NJ}$). The pressure head above ground level at each node can be determined by subtracting the elevation of ground level at the node from the total head at that node.

The basic equations are:

(i) the continuity equations:

$$\sum_{j=1}^{NPJ} Q_j + q_i = 0 \quad \text{for all nodes } i = 1, \ldots, NJ \qquad (13.87)$$

where the summation is made across the *NPJ* pipes connected to node *i* (flow away from the node is taken as positive).

(ii) the head loss equations:

for pipe j between nodes i and k, the head loss is related to the discharge in the pipe as well as its diameter. One head loss equation commonly used in water supply engineering is the Hazen-Williams equation given here in SI units (Streeter and Wylie, 1979).

$$H_i - H_k = h_j = \frac{10.675 L_j Q_j \left|Q_j\right|^{0.852}}{C_j^{1.852} D_j^{4.8704}} \qquad \text{for all pipes } j = 1, \ldots, NP \qquad (13.88)$$

where, h_j = head loss in pipe j
L_j = length of pipe j
C_j = Hazen-Williams coefficient for pipe j

Equation sets (13.87) and (13.88) provide $(NJ + NP)$ equations in the same number of unknowns (pipe discharges and nodal heads).

The number of equations can be reduced by summing the head losses around loops and equating the sum to zero or by transposing Equation sets (13.88) to get expressions for Q_j $(j = 1,2, \ldots NP)$ in terms of H_i and H_k and then substituting into Equation sets (13.87) to eliminate the pipe discharges (Ormsbee and Wood, 1986).

All solution techniques involve the solution of simultaneous non-linear equations and may require the use of techniques such as the Newton-Raphson method (Shamir and Howard, 1968).

CHAPTER FOURTEEN

Epilogue

Planning and design are the key activities that are used, together with management, to bring engineering projects from the initial concept stage through to successful implementation. Our prime goal in preparing this book has been to explain how engineering planning and design work is carried out. In the first ten chapters we have introduced basic concepts and have described procedures that are used in engineering planning and design. In Chapters 11 to 13 we have presented some analytic techniques that find frequent use in the detailed phases of planning and design.

We are mindful that some of our readers will be students who are just commencing their engineering studies. In Chapter 1 we therefore describe the kind of work that engineers typically engage in. In brief, we suggest that engineering work is concerned with the creation, maintenance and extension of the physical infrastructure that modern society needs, in order to function. In Chapter 1 we also emphasise that planning and design are similar in nature, and are both problem-solving activities.

Engineering problems tend to be both complex and open-ended. Open-ended problems do not have a single "correct" solution; rather, there is a range of possible solutions, so that the aim is not to find the "correct" solution, but rather the "best" solution. If the problem is very complex, it may not be feasible to find the "best" solution, and the aim then is to find a relatively "good" solution.

The systems concepts introduced in Chapter 2 provide a means for dealing with complexity. A methodology for solving open-ended problems is explained in Chapter 3, and this provides the basis for planning and design work. Later chapters describe specific procedures that are also useful in planning and design. For example, the scheduling techniques described in Chapter 5 help ensure that a project is completed on time and within budget.

In engineering work, it is necessary to be able to think creatively and laterally, as well as rationally and logically. These contrasting ways of thinking are sometimes referred to as right-brain thinking and left-brain thinking. Because of their early education, training and background, engineering students today still tend to be best suited to tasks that involve analytic thinking and rational analysis. It can therefore come as a surprise that they need to develop skills in right-brain thinking, in order to be creative. Various aspects of creative thinking are discussed in Chapter 4, where some techniques for developing right-brain activity are introduced.

Engineering management is far too broad a topic to be covered in any detail in this book on planning and design. Nevertheless, in Chapter 6 we deal briefly with several aspects of management that are needed when engineering planning and design work is undertaken. The skills of self-management, time-management and people-management become especially important in the planning and design of large, complex projects, when engineers must work in teams, both large and small, and with other engineers and specialists from other disciplines.

Other relevant skills are discussed in later chapters. One of the most important is the ability to communicate effectively and unambiguously, not only with colleagues, but also with clients and with the community at large. In current engineering work, clients and other interested parties (the "stakeholders") can take a very active part in the relevant decision-making stages of a project. This places added demands on the communication skills of engineers, and in particular on their ability to explain technical problems to non-experts. The topic of communication is introduced in Chapter 7.

Engineers also need to be reasonably familiar with broader areas of knowledge such as economics and ethics, law and sociology, and environmental matters including sustainability. We have provided brief introductions to these areas in Chapters 8, 9 and 10 of this book. Sooner or later, of course, engineers encounter unusual problems for which their background knowledge and expertise is inadequate. They then have to be able to acquire and absorb new information and knowledge rapidly, so that they can communicate with, and work with, experts in the appropriate fields.

In all practical engineering work there are inherent risks that have to be evaluated and managed. In planning and design, the goal is to create engineering systems and processes that are safe and reliable. Chapter 11 introduces the concepts of engineering risk and reliability.

Decision-making is an important aspect of planning and design. In each phase of an engineering project, many decisions (or choices) have to be made. An introduction to mathematical decision making is presented in Chapter 12, not in the expectation that the idealised decision models presented will always be directly applicable to real situations, but rather with the intention of providing a conceptual framework for dealing with real decision problems. This should be of value to engineers when they are asked to explain and justify their technical decisions.

Chapter 13 introduces some of the optimisation techniques that are available to engineers, including traditional mathematical techniques and modern search approaches that are used in association with non-linear and discontinuous modelling and computer simulation. Optimisation techniques often find application in the detailed phases of planning and design.

There are many specialised design and planning techniques that we have not mentioned in this book. These tend to be field-specific and are dealt with in undergraduate and post-graduate coursework in many different engineering disciplines. However, we have tried to provide a general overview, and a framework of ideas, for the discipline-specific methods of planning and design.

New solutions to engineering problems sometimes come from scientific and mathematical knowledge via technological innovation. Modern engineering developments and innovations also come from ongoing engineering research. While there is a strong continuing need for empirical investigation in engineering work, the emphasis over many years has tended to be away from the use of empirical *ad hoc* approaches, towards the use of fundamental principles and general concepts. The ever-widening scope of engineering analysis is a result of ongoing research and an exponential increase in the capacity of computing facilities. There have been unparalleled developments in the computational methods of analysis that engineers use to create ever-more accurate models of physical systems and to obtain more efficient solutions to complex problems of analysis and design.

The history of distance-communication is an interesting example of the benefits of research. The earliest communications over long distances were visual: bonfires were lit successively on adjacent hills and high vantage points to send important (but simple and short) messages across a country, perhaps to announce important matters of state. Complex, detailed information had to be transported physically, and this was much slower. In the Roman Empire, important government documents were transported via the road systems and it could take months for bulky documents from Rome to be delivered to the most distant outposts; nevertheless, the business of empire was remarkably efficient. But it was not until the discoveries and inventions of people such as Hertz, Marconi, Morse and Edison, and more recently Claude Shannon, that the rapid and secure transmission of detailed and complex information over large distances commenced. Today, after generations of further research, and with the development of new areas of knowledge including information theory and cryptography, a person today can speak directly and privately with someone else located almost anywhere on earth and even in space. And it is also possible to transmit, confidentially, enormous amounts of information, including complex visual images, sounds, written documents and data files, almost anywhere and almost instantaneously, using modern channels of tele-communication.

Planning and design are not unique to engineering. They form the backbone of many diverse professional disciplines, such as architecture, financial planning and agricultural economics. For example, architects are involved in planning and design when, together with engineers and building contractors, they create new suburbs and living areas, shopping centres, retirement villages and multi-function tall buildings. Looking more broadly at society, we see a wide range of designed products, such as jewellery, fine art, furniture, botanical gardens and zoos, fun parks, entertainment centres and innumerable consumer products. What makes engineering planning and design unique is the nature of the problems that are dealt with.

We have suggested that engineering work is concerned with the physical infrastructure of society, and we have pointed to a close two-way relationship between society and its engineering infrastructure. Although the nature of engineering work is dictated by the needs of society, the nature of society is itself influenced by the engineering solutions to infrastructure problems. This inter-play between engineering work and human society is as strong today as it was in previous centuries, and in the earliest human civilisations.

Engineers in early civilisations had the task of creating basic infrastructure: buildings and fortifications to house and protect the population; clean water for drinking, cleaning and for irrigating crops; roads for transport and communication; and tools and machines to make manual work more efficient. And the most basic needs of our society today have not changed substantially; they still include shelter, protection and defence, clean water, systems for handling waste and sewerage, and efficient systems for transport and for communications. However, in today's affluent society the demands for infrastructure are much broader and are continually broadening. Manual labour has been largely replaced by machinery in the heavy and light industries. Labour-saving devices are to be found everywhere. There are washing machines, dishwashers, ceramic stoves, microwave ovens and air conditioners in our homes, and computers, printers, shredders, data storage devices and other labour-saving equipment in our offices. Year after year these devices become more intelligent and smarter, as do the robots in our factories.

However, we now use enormous amounts of energy to drive the innumerable devices that replace routine manual labour and to keep our infrastructure operating. When the ancient Egyptians constructed their buildings, temples and pyramids, the energy they used was derived from the sun and the soil. That energy was supplied to the work crews via the bread, beer and meat that they were fed, and it was stored for a short time in human muscle. We use vastly more energy today, in our offices, factories, homes and automated construction sites, but it still comes largely from the sun and the soil, but not via beer, bread and meat. Rather, it comes mainly from petroleum, natural gas, and coal, in which solar energy has been stored as potential energy over many millions of years.

Our forebears learnt how to run mills by extracting energy from wind and flowing water. They also used wind to sail ships around the world. They exploited animal muscle, as well as human muscle, and they were innovative in obtaining energy from wood-burning fires. Since the Industrial Revolution, however, the demand for energy has increased almost exponentially, and every available energy source is now being exploited.

Of course, there are many parts of the world that are not "advanced" in the sense that they use enormous amounts of energy to drive sophisticated engineering infrastructure. In much of the world today, poverty is far more common than affluence; starvation is a far, far more common problem than obesity. Inequity currently takes many forms. Indeed, the world-wide problems of inequity are "intractable", in the sense that this word was used in Chapter 3. Engineering work will certainly be needed to solve the important and difficult problems of inequity, but engineering input alone cannot be sufficient. Political and social solutions are needed. Engineering input can only become effective after difficult and astute political decisions have been taken.

As we come towards the end of this book in which we have discussed the nature of engineering, the way engineers undertake their work, and (very briefly) the history of engineering, it is natural to think about the future. Although we inevitably face unpredictable changes in an uncertain future that will be influenced by world events, we can be confident that some things in engineering are unlikely to change radically, and there are other things that will change in a more-or-less predictable fashion.

We can be confident that in fifty and a hundred years' time, engineers will still be providing for the basic human needs of most of the world's population, such as fresh water and shelter in a sustainable environment, and effective systems for communication and transport. Engineers may well be required to provide this basic infrastructure in new and remote locations, such as in Antarctica, and perhaps on the moon and on mars. Extrapolating from the present, we can also be confident that improvements to the physical infrastructure will occur progressively in the coming decades, as mature components are made obsolete by new, disruptive technologies and industries.

Catastrophic events such as earthquakes, floods and tsunamis will continue to occur regularly around the world, as at present, and engineers will still be dealing with the aftermath of these natural catastrophes for many years to come. Armed conflict seems to be inherent in human nature, and brings devastation to the physical infrastructure and incalculable loss of life. Sadly, we can expect that engineers, into the distant future, will be called on to rebuild, repair and replace infrastructure that has been destroyed through conflict.

Unfortunately, armed conflict in the future could be more devastating than at present. The effects of nuclear and biological warfare, or perhaps of nuclear and biological terrorism, would be horrendous. If all-out nuclear war were to occur, it would bring human society back, in the best-case scenario, to a primitive starting point, and engineers would then have to re-create the basic infrastructure that humans need for survival and comfort.

In thinking about the future, we need to remember that the Earth we inhabit is not in stasis: it is not a system that is, or ever was, or ever will be, in a state of static and stable equilibrium. It is continually evolving and changing. Even so, significant changes seem to occur relatively slowly if judged in terms of the human life span. This has led many people to the false impression that we live in an unchanging world that can be moulded to our wishes. In fact, our planet is currently undergoing accelerating changes that are caused, in part, by our own actions, and these changes seem not to be within our control.

The problems we face in regard to changing climate and increasingly scarce resources will only intensify in coming decades. Lack of fresh water has long been a severe problem in many parts of the world. As nations and regions of the earth that are now in poverty become more affluent, competition for scarce resources will only increase. Engineers can make a difference by improving the efficiencies of procedures for harvesting, using and reusing scarce resources such as water and energy.

Scenario planning, which was discussed briefly in Chapter 3, can be used to examine and, to a limited extent, plan for possible future events, both favourable and unfavourable. When we make extrapolations from the present and the past into the future, we find any number of topics in need of scenario planning studies.

Extrapolating from the present, but on a more optimistic note, we can expect great benefits to flow on from continuing developments in areas such as medical and biomedical engineering. In the longer term, we can expect to see improvements that will allow the replacement of most human parts, with longer and happier lives for all humans. Artificial intelligence and nanotechnology are at present are in their infancy, but applications of these technologies will be positive and spectacular.

Optimistically, we can hope and expect that humans will survive the severe adverse events they face in the coming decades and century, and that an improving engineered infrastructure will make life not only possible, but enjoyable for most humans. In the longer term, the aim must be to achieve a levelling out of the world's population, to provide adequate food, clean water and basic sanitation for all people, with better husbandry of the earth's resources, and better treatment of the environment and other living creatures.

In the immediate future, however, action is urgently needed to curb the current over-use of water, energy and other scarce resources. The alternative, of trying to meet ever-increasing demands, is simply not feasible. Demand will have to be attenuated, and improved efficiencies in the use of scarce resources will have to be achieved. This will require a great deal of innovative engineering planning and design in the immediate future.

Index