Education in the Asia-Pacific Region:
Issues, Concerns and Prospects 18

Magdalena Mo Ching Mok *Editor*

Self-directed Learning Oriented Assessments in the Asia-Pacific

UNESCO UNEVOC
INTERNATIONAL CENTRE

ASIA-PACIFIC EDUCATIONAL
RESEARCH ASSOCIATION

Springer

Self-directed Learning Oriented Assessments in the Asia-Pacific

EDUCATION IN THE ASIA-PACIFIC REGION: ISSUES, CONCERNS AND PROSPECTS

Volume 18

For further volumes:
http://www.springer.com/series/5888

Magdalena Mo Ching Mok
Editor

Self-directed Learning Oriented Assessments in the Asia-Pacific

Springer

Editor
Magdalena Mo Ching Mok
The Hong Kong Institute of Education
Lo Ping Road 10
New Territories
Tai Po, Hong Kong

ISBN 978-94-007-4506-3 ISBN 978-94-007-4507-0 (eBook)
DOI 10.1007/978-94-007-4507-0
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012946747

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Introduction by the Series Editors

The ways in which learners are assessed by their teachers have a profound influence on student and teacher behavior, including such things as the teaching methods adopted, teaching and learning materials used, and organizational arrangements within the classroom and school. This is one of the key reasons why assessment reform has attracted, and continues to attract, considerable attention from education policy makers, researchers, and practitioners worldwide, particularly when it comes to their interest in improving learning outcomes and strengthening the accountability of teachers, schools, and education systems as a whole. This is to be expected since national education systems need to (and should) be accountable to taxpayers, parents, teachers, and to the learners themselves, all of whom want to ensure that the considerable financial (and other) resources devoted to the education enterprise are being put to the best possible use with regard to achieving high quality and reliable outcomes.

The Asia-Pacific region is vast and diverse, with countries at different stages of economic development. However, despite such diversity, all countries in the region have a shared concern that of seeking to improve access to, and the quality and relevance of, education and schooling.

This edited volume examines the various ways in which assessment methods have (and currently are) being reformulated and reformed in the Asia-Pacific region, with particular reference to self-directed learning orientated assessment (SLOA). As Professor Magdalena Mok points out, these reforms share a common emphasis on assessment *for* learning and on assessment *as* learning.

This book provides a comprehensive survey of assessment reform in countries in Asia-Pacific and an insightful analysis of how assessment reform has been used in many education systems throughout the region to drive educational change. In addition to examining the theory of self-directed learning orientated assessment, the volume surveys tools for implementing self-directed learning orientated assessment and provides case studies of SLOA in countries in the region.

This is an important book on an important topic and as such is certain to enjoy a large readership from educational researchers, policy makers, and practitioners who are interested in improving the quality and effectiveness of the learning which occurs in classrooms, schools, and education systems as a whole.

Hong Kong Institute of Education
National Institute for Educational
Policy Research (NIER) of Japan
February 27, 2012

Rupert Maclean
Ryo Watanabe

Acknowledgements

I would like to express my sincere gratitude to all those who have helped me with this book project. I am deeply indebted to Professors Rupert Maclean and Ryo Watanabe, the series editors, for their encouragement and for writing the foreword. I thank them for their generous support and advice throughout the book project. I am obliged to the thoughtful and constructive comments of the anonymous reviewers. My deepest appreciation goes to the preeminent scholars who contributed their chapters to this book. I would like to thank Sze Ming Lam, who has typeset this book, and Liz Stone and Hilary Tolley for their support on English editing. I thank Springer for trusting me with editorship and for publishing this book. My sincere appreciation is extended to The Hong Kong Institute of Education for enabling support on research in association with this book project.

Contents

Part I Theory of Self-Directed Learning Oriented Assessment

Part II Tools for Implementing Self-Directed Learning Oriented Assessment

Part III Case Studies of Self-Directed Learning Oriented Assessment in the Region

List of Figures

List of Tables

Contributors

Carl Martin Allwood Department of Psychology, University of Gothenburg, Gothenburg, Sweden

William John Boone Department of Educational Psychology, Miami University, Oxford, OH, USA

Paisley Tsz Mei Cheung Education Assessment Services, Hong Kong Examinations and Assessment Authority, Wanchai, Hong Kong

Hye-Jeong Choi Office of Research and Department of Psychology, University of Nebraska-Lincoln, Athens, GA, USA

Michelle Davidson Teaching and Learning (Assessment), Trinity Grammar School, Sydney, Australia

Jimmy de la Torre Graduate School of Education, Rutgers, The State University of New Jersey, Newark, USA

Lorna Earl Aporia Consulting Ltd., Toronto, Canada

International Centre for Educational Change at OISE, University of Toronto, Toronto, Canada

Pauline Swee Choo Goh Faculty of Education and Human Development, Universiti Pendidikan Sultan Idris (Sultan Idris Education University), Perak Darul Ridzuan, Malaysia

Chi Ming Ho Formerly Centre for Assessment Research and Development, The Hong Kong Institute of Education, Tai Po, Hong Kong

Connie Chia-Ling Hsu Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong

Slava Kalyuga School of Education, University of New South Wales, Sydney, NSW, Australia

Steven Katz Psychological Foundations of Learning and Development at OISE, University of Toronto, Toronto, Canada

Sabina Kleitman School of Psychology, University of Sydney, Sydney, NSW, Australia

Sze Ming Lam Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong

Doris Ching Heung Lau Formerly Centre for Assessment Research and Development, The Hong Kong Institute of Education, Tai Po, Hong Kong

Henry Kai On Lee School of Continuing and Professional Education, The Hong Kong Institute of Education, Tai Po, Hong Kong

Anthony Wai Chi Leung HHCKLA Buddhist Wisdom Primary School, Sheung Shui, Hong Kong

Karina Kar Lee Mak School of Psychology, University of Sydney, Sydney, NSW, Australia

Bobbie Matthews School of Nursing and Midwifery, Flinders University, Bedford Park, SA, Australia

Magdalena Mo Ching Mok Department of Psychological Studies, and Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong

Ming-Yan Ngan Department of Curriculum and Instruction, The Hong Kong Institute of Education, Tai Po, Hong Kong

Min Pan Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD, USA

Somwung Pitiyanuwat Chairman, National Institute of National Testing Service (NIETS) and Royal Associate,The Royal Institute, Bangkok, Thailand

Tan Pitiyanuwat Interior Design Department, Faculty of Architecture, Kasem Bundit University, Bangkok, Thailand

André A. Rupp Department of Human Development and Quantitative Methodology (HDQM), University of Maryland, College Park, MD, USA

Lazar Stankov National Institute of Education, Jurong West, Singapore

Centre for positive Psychology and Education, School of Education, University of Western Sydney, Sydney, NSW, Australia

John R. Staver Department of Curriculum and Instruction, College of Education, Purdue University, West Lafayette, IN, USA

Hak Ping Tam Graduate Institute of Science Education, National Taiwan Normal University, Taipei City, Taiwan

Stephen Yin Chuen Ting Formerly Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong

Jim Tognolini Senior Vice President, Research and Assessment, Pearson plc and Senior Research Fellow at Oxford University, Sydney

David Tzuriel School of Education, Bar Ilan University, Ramat Gan, Israel

Wen-Chung Wang Department of Psychological Studies, and Assessment Research Centre, The Hong Kong Institute of Education (HKIEd), Tai Po, Hong Kong

Michael Ying Wah Wong Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong

Margaret Wu Work-based Education Research Centre, Victoria University, Melbourne, Australia

Jacob Kun Xu Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong

Melissa Seward Yale Department of Educational Studies, Purdue University, West Lafayette, IN, USA

Zi Yan Department of Curriculum and Instruction, The Hong Kong Institute of Education, Tai Po, Hong Kong

Jing Jing Yao Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong

Department of Psychology, Zhejiang Normal University, Jinhua, China

Sarah Young School of Psychology, University of Sydney, Sydney, NSW, Australia

George C. Yu Director and Senior Consultant of the Hong Kong Language and Culture Institute, Centre for Assessment Research and Development, The Hong Kong Institute of Education, Tai Po, Hong Kong

Yue Zhao Assessment Research Centre, The Hong Kong Institute of Education (HKIEd), Tai Po, Hong Kong

Part I
Theory of Self-Directed Learning Oriented Assessment

Chapter 1
Assessment Reform in the Asia-Pacific Region: The Theory and Practice of Self-Directed Learning Oriented Assessment

Magdalena Mo Ching Mok

1.1 Background: The Broader Context for Change

1.1.1 Assessment Reforms in the Region

Assessment reform has been at the heart of education reforms in major systems in the Asia-Pacific region since the turn of the century (Hogan et al. 2009; Goh and Matthews 2012, this volume; Mok et al. 2003; Ng 2010; Pitiyanuwat and Pitiyanuwat 2012, this volume; Yu 2012a, this volume). Such reform movements are driven by two motives: *accountability* and *improvement*. National systems are accountable to taxpayers, teachers and parents on public money, and there is increasing public scrutiny of government expenses in developed countries as the internet makes it easier to access information compared with the last century. In parallel, in facing challenges brought about by globalization and the rapid speed at which knowledge is created, continuous improvement is seen by many governments as the only way to stay on a par with the rest of the world (Sahlberg 2006). Research (Hogan et al. 2009; Goh and Matthews 2012, this volume; Mok et al. 2003; Ng 2010; Pitiyanuwat and Pitiyanuwat 2012, this volume; Yu 2012a, this volume) has found that reforming the education system and repositioning assessment as a means to build up capacity for continuous self-improvement is common amongst systems in the Asia-Pacific region.

In the wave of education reforms, assessment reform has often been used by systems in the region as a means to drive the changes. For instance, under the vision of 'Thinking Schools, Learning Nation', in 1997, the Singapore government put forward new assessment policies at school level and changed assessment from

M. Mo Ching Mok (✉)
Department of Psychological Studies, and Assessment Research Centre,
The Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, N.T., Hong Kong
e-mail: mmcmok@ied.edu.hk

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_1, © Springer Science+Business Media Dordrecht 2013

summative to formative in the implementation of the School Excellence Model and the School Awards Model (Ng 2010). In conjunction with the 'Teach Less, Learn More' policy of Singapore, these models of school excellence emphasized the development in students of the capacity for self-directed learning, deep conceptual understanding, sharing of knowledge and knowledge construction.

Similar to Singapore, assessment for learning is high on the agenda of education reform in Hong Kong (Berry and Adamson 2011; Lee 2012, this volume). Since 2001, the Hong Kong government has launched a large-scale curriculum reform which emphasized the shift from knowledge transmission to building the students' capacity of learning how to learn (Cheung and Wong 2011; Curriculum Development Council 2001; Lee 2012, this volume). Governmental determination in reforming the culture and practice of assessment to promote independent learning and thinking in Hong Kong is unambiguously identified in the review of public examination system in Hong Kong report or the ROPES report (Hong Kong Baptist University and Hong Kong Examinations Authority 1998). There are clear directives from the government that 'Assessment is an integral part of the curriculum, pedagogy and assessment cycle. It involves collecting evidence about student learning, interpreting information and making judgements about students' performance with a view to providing feedback …' (Curriculum Development Council 2009, Booklet 4, p. 1).

1.1.2 Commonalities of Assessment Reforms in the Asia-Pacific Region

Mok et al. (2003) found in their review of education reforms in eight systems in the Asia-Pacific region (Australia, Hong Kong, Japan, Korea, New Zealand, Papua New Guinea, Singapore and Taiwan) that assessment reform was not only core to the education reforms in these systems, but also that these reforms shared many commonalities across the systems:

- Purpose of assessment: changing from selecting the best candidate for further education as the sole purpose of assessment to serving multiple purposes including the provision of feedback to support learning
- Philosophy of assessment: changing from summative assessment to both formative and summative assessment being used, with a strong emphasis on the relationship between teaching, learning and assessment
- Directive of assessment: changing from evaluative to learning-directed
- Methods of assessment: changing from paper and pencil to multiple formats and methods including electronic, performance, portfolio and project-based
- Analysis of assessment data: changing from traditional test theory to Rasch-based or item response theory
- Drivers of assessment: changing from teacher-initiated to self- and peer-initiated assessment
- Frame of reference: changing from norm-referenced to criteria- and standards-referenced

- Domains of coverage: changing from assessing cognitive domains only to assessing multiple domains including cognitive, affective and social domains in whole-person development of students

1.1.3 Assessment as Learning Reform: Self-Directed Learning

The use of assessment reform as a means to achieve the national education aim of building sustainable capacity for self-directed learning is the most prevalent feature in the assessment reforms in the region. There is an increasing emphasis on preparing students who are '*trainable* rather than *trained*' and who are capable of self-evaluation and of continuous learning throughout life (Maclean 2010, p. iv). The strong intention to build up capacity for self-directed learning through education reform in developed countries in the region is reflected in their respective aims of education. For instance, Hong Kong education aims to 'enable everyone to develop their full and individual potential … so that each individual is ready for continuous self-learning…' (Education Commission 2000); Japan education aims 'to raise the ability of self-education' and 'the ability to shape their own lives' (Abiko 2011, p. 359; Japan Ministry of Education 2000); Korean education aims 'to raise a self-reliant individual equipped with a distinct sense of independence, a creative individual with a sense of originality, and an ethical individual with some sound morality and democratic ideology' (Korean Ministry of Education 2000); Singapore education is to 'develop self-directed learners who take responsibility for their own learning, and who question, reflect and persevere in their pursuit of learning' (Singapore MOE 2010); for Taiwan, 'Education and culture shall aim at the development among the citizens of the national spirit, the spirit of self-government…' (Taiwan Ministry of Education 2001) and 'to encourage people's planning for self-directed learning based on theory of career development' (Taiwan Ministry of Education 2011); and Thailand's education system aims 'to develop student's learning capabilities in the areas of: self-learning, creative thinking and basic academic learning' (Thailand Ministry of Education 2000; Pitiyanuwat and Pitiyanuwat 2012, this volume).

The capacity for self-directed learning is labelled as 'Assessment *As* Learning' by Lorna Earl (2003) and 'Learning How To Learn' by Paul Black and associates (Black et al. 2006, 2011). Learning to Learn is one of four pillars of education for the twenty-first century identified by UNESCO (Delors et al. 1996), along with Learning to Be, Learning to Do, and Learning to Live Together. It is the foundation for lifelong learning.

1.1.4 Assessment for Learning Reform

The second prevalent phenomenon in assessment reforms in the region is the move from assessment *of* learning to assessment *for* learning or the generation of feedback to inform teaching and learning. In facing challenges of increasing global competition between nations in the new century, there is strong consensus amongst governments in the region that, for their nations to succeed, they must build a capacity for knowledge creation and knowledge transfer (Sahlberg 2006). Huge resources have been invested by these governments in making explicit policy shifts to assessment *for* learning. Assessment is to generate information to 'feedforward' for subsequent learning (Berry and Adamson 2011; Black et al. 2011; Carless 2007; Hogan and Gopinathan 2008; Lee 2012, this volume; Mok et al. 2003; Ng 2010).

Assessment *as* learning and assessment *for* learning are two new conceptions of assessment, and together, they form the foundations for assessment reforms in major education systems in the Asia-Pacific region. Assessment reform is particularly important to learners in the region because many of the education systems in the region have very strong cultures and traditions of assessment *of* learning. Assessment in China, Hong Kong, Japan, Korea, Macau, Singapore and Taiwan, for instance, is traditionally in the form of high-stake norm-referenced examinations that determine future prospects of education and employment of the examinees (Berry and Adamson 2011; Hogan and Gopinathan 2008).

1.1.5 Resolving Tensions in Assessment Reforms

In systems where assessment is mainly used for selection purposes, there are only a few winners and many losers. It is understandable that attention can easily focus on marks and grades instead of on learning in such systems. Research has shown that competitive assessment not only has pervasive debilitating effects on current learning including narrowing of learning but also induces stress and depression, deteriorates sleep quality and increases self-blaming, learned helplessness and other maladaptive beliefs as well as students' motivation for further learning (Berry and Adamson 2011; Putwain 2009). Nevertheless, it is not easy to uproot the deeply entrenched parental and societal beliefs on the functions served by examinations. In reality, many teachers still have to carry out assessment for summative purposes in their day-to-day practice in the middle of their local assessment *for* learning reform, resulting in teacher stress and resistance to change (Ballet and Kelchtermans 2009; Cheung and Wong 2011; Choi and Tang 2009; Day 2008; Ho et al. 2012, this volume; Goh and Matthews 2012, this volume).

The message is clear: there is an urgent need to revise and redesign pedagogy in order to reconcile the tension between assessment *as*, *for* and *of* learning and to glean the benefit of each for enhanced learning and teaching. The framework of self-directed learning oriented assessment (SLOA) discussed in this book offers a feasible solution to that new pedagogy. SLOA is a coherent framework of assessment, deliberately designed to capitalize on the integrative impact of assessment *of*, *for* and *as* learning in the construction of assessment activities for optimal learning and for the cultivation of self-directed learning capacities in students (Mok 2010).

The overall aim of this book is to present to readers – teachers, parents, educators and education policymakers – a set of theory-driven assessment strategies, guidelines and practical examples for the successful implementation of assessment reforms in schools and classrooms. The genesis of this book is the 3-year longitudinal assessment project in Hong Kong and China (2005–2008) and the SLOA projects in Macau (2008–2011) reported in Mok (2010) and Yu (2012a, b, this volume). This book provides further elaborations on the theoretical foundations of SLOA, examines actionable assessment strategies and tools that can facilitate teachers' work and presents practical examples where SLOA has been applied to teaching and learning in primary and secondary classes in the region. The book comprises 20 chapters and is divided into three parts: Theory of Self-Directed Learning Oriented Assessment, Tools for Implementing Self-Directed Learning Oriented Assessment and Case Studies of Self-Directed Learning Oriented Assessment in the Region. The rest of this chapter will be devoted to the theory and practice of SLOA and an introduction to the other chapters in the book.

1.2 Conceptions of Self-Directed Learning Oriented Assessment

Mok (2010) proposed an a priori conceptual framework to guide research on self-directed learning oriented assessment (SLOA). As the name implies, SLOA focuses attention on assessment that can support and advance learning and assessment that is self-directed by the learner. This section will expand on the meaning of these two aspects of SLOA. Furthermore, this section will explain how the SLOA framework draws from, and is being informed by, recent research in a number of domains in learning psychology (notably, self-directed learning, metacognition and feedback) and in psychometrics. Lastly, this section will explain how the three concepts of assessment *of* learning, assessment *for* learning and assessment *as* learning integrate and supplement each other in the SLOA framework.

1.2.1 Learning Oriented Assessment

The name SLOA is made up of two parts, namely, 'LOA' and 'S'. 'LOA' comes from Carless (2007), who coined the term Learning Oriented Assessment (LOA). 'Learning' was deliberately placed before 'assessment' in order to highlight the centrality of learning in all assessment activities. LOA means that (a) assessment activities should be designed as learning tasks, (b) assessment should engage students in the evaluation of the learning progress and (c) feedback from assessment should be used as feedforward to inform current and future learning (Carless 2007). Through these three principles, LOA gets to the spirit of assessment *for* learning.

1.2.2 Self-Directed Learning

The ‘S’ in SLOA means self-directed learning (Earl 2003; Knowles 1975; Lee 2012, this volume; Paris and Paris 2001; Pintrich 2004; Schunk 2008; Shute 2008). The capacity for self-directed learning is fundamental to sustainable development in the twenty-first century, given the rapid speed at which knowledge is created. Knowledge and skills that students will need when they join the workforce have not yet been created today when they are at school. Consequently, education in the new century has to go beyond the transmission of knowledge to students. Rather, the core mission of education is to engender in students the capacity for knowledge creation, knowledge management, knowledge transfer and knowledge acquisition. In other words, education in the new century means learning how to learn (Delors et al. 1996). In the process of knowledge creation, management, transfer and acquisition, the learner must be able to set learning goals, plan the course of action, manage resources, monitor his/her learning progress, assess the level of achievement so far, generate feedback and adjust and self-regulate accordingly. The learner holds the key to success in the learning process. Unless and until the learner is capable of directing his/her own action in this process, there will be no real learning. In this regard, the SLOA framework is very much inspired by the work of Earl and associates (Earl 2003; Earl and Katz 2008, reprinted in this volume), in which she argued that assessment is actually learning and labelled the concept as ‘assessment *as* learning’. Engaging the learner as his/her own assessor, or assessment *as* learning, is the ultimate goal of assessment *for* learning. Earl (2003) wrote:

> *The student is the link. Students, as active, engaged and critical assessors can make sense of information, relate it to prior knowledge, and master the skills involved. This is the regulatory process in metacognition. It occurs when students personally monitor what they are learning and use the feedback from this monitoring to make adjustments, adaptations and even major changes in what they understand. Assessment as learning is the ultimate goal, where students are their own best assessors.* (Earl 2003, p. 47)

1.2.3 Metacognition

Taking after Earl (2003; Earl and Katz 2008, reprinted in this volume), assessment *as* learning in the SLOA framework means an assessment process in which the learner actively considers and sets learning goals, deliberates upon and selects learning strategies, monitors learning, assesses learning progress, evaluates feedback information and, as a result, reaches new understanding, connects new information with existing knowledge or even revises learning goals or strategies. In other words, assessment *as* learning means the learner is exercising the self-regulatory process of metacognition (Brown 1987; Earl 2003; Flavell 1979; Loyens et al. 2008; Schunk 2008).

The SLOA framework incorporates a range of metacognitive tools and mechanisms, including the provision of timely feedback from assessment and explicit teaching of a range of strategies, so as to raise students' self-awareness (metacognitive knowledge) of their own learning process and to enrich their repertoire of self-regulation skills (self-regulation of cognition). These skills include identifying key issues in the learning task, posing questions, selecting learning strategies and monitoring progress by situating these strategies in learning tasks of curriculum subjects, as well as modelling and scaffolding the strategies (Black and Wiliam 1998a; Boone et al. 2012, this volume; de la Torre 2012, this volume; Hattie and Timperley 2007; Mok et al. 2012, this volume; Hsu et al. 2012, this volume; Kalyuga 2012, this volume; Lee 2012, this volume; Choi et al. 2012, this volume; Tzuriel 2012, this volume; Yu 2012a, b, this volume).

1.2.4 Feedback

Assessment is formative (assessment *for* learning) when feedback generated from the assessment is directed towards the quality of the task or learning process, identifies misconceptions and supports the development of more effective learning strategies (Black and Wiliam 1998a; Hattie and Timperley 2007; Lee 2012, this volume; Shute 2008).

Feedback also contributes to the metacognition of the learner (assessment *as* learning) through generating cues for the learner to internalize three key feedback questions (Hattie and Timperley 2007, p. 86):

1. 'Where am I going': What is the desired outcome (long-, intermediate-, short-term) of my learning endeavour? What is the anticipated outcome if I approach the problem this way? How is this new learning related to my previous learning?
2. 'How am I going': What does the assessment evidence tell me about the effectiveness of my learning strategies and is there a gap between my desired goal and my current progress? If there is a gap, what are the possible causes?
3. 'Where to next': What should be my next steps? Do I have to keep going this way or should I modify my learning strategies? Should I change my goal (set higher/lower goal, change direction)? Should I seek help and, if so, from where should I get help?

1.2.5 SLOA: Integrating Assessment Of, For and As Learning

The SLOA framework comprises three integrative components: assessment *of* learning, assessment *for* learning and assessment *as* learning. They refer to the purposes of assessment or how the assessment outcomes are to be used. The relationship between them is best described as a recurrent three-component learning process (Fig. 1.1). First, assessment *of* learning in the SLOA framework refers to assessment activities of the teacher and their students that aim to generate evidence

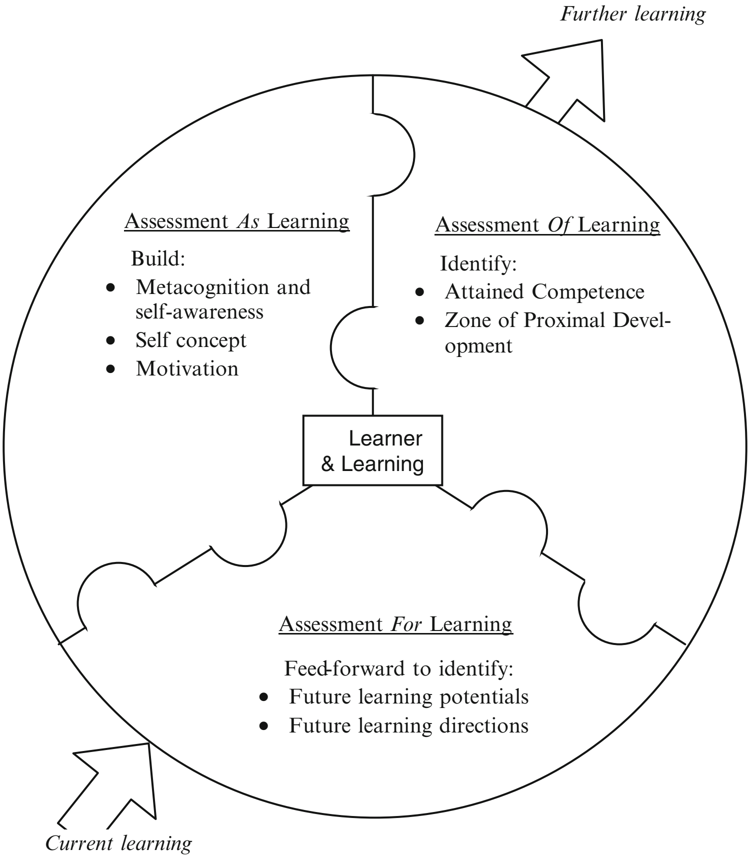

Further learning
Assessment *As* Learning
Build:
Metacognition and self-awareness
Self concept
Motivation
Assessment *Of* Learning
Identify:
Attained Competence
Zone of Proximal Development
Learner & Learning
Assessment *For* Learning
Feed-forward to identify:
Future learning potentials
Future learning directions
Current learning

Fig. 1.1 Assessment *of*, *for* and *as* learning

about current learning. In assessment *of* learning, the teacher and students ask, ‘Where are we in the learning?’

Next, after the evidence is generated, the teacher and students ask, ‘Is there a gap between the desired learning goal and the current level of learning?’ In order to address this question, the desired learning goal has to be established and made clear to both parties. Consequently, in the SLOA framework, assessment *for* learning often begins with goal setting and clarification of the desired learning goal. Even though the question concerning the gap can be addressed by assessment *of* learning, information generated from such assessment is often inadequate to address the next question, ‘If there is a gap, how can we close the gap?’ Assessment *for* learning in the SLOA framework refers to assessment activities by the teacher and the students to collect evidence with an aim to feedforward to inform further learning in terms of directions and potentials. That is, assessment *for* learning enables the teacher to ‘modify the teaching and learning activities’ and to ‘adapt the teaching work to meet the needs [of individual students]' (Black and Wiliam 1998b, p. 2). Third, assessment *as* learning in the SLOA framework means that the learner internalizes the questions of, 'Where am I going? How am I doing? How can I learn better? How can I keep up my motivation?' and acts upon them in a constant process of self-monitoring during learning.

1.2.6 Theoretical Underpinnings of SLOA

Four theories in assessment, psychometrics and learning underpin the SLOA framework. They are standards-referenced assessment, cognitive diagnostic assessment (CDA), Rasch measurement and metacognition. According to Tognolini and Davidson (2012, this volume), a standards-referenced system consists of a curriculum which clearly articulates learning outcome standards and performance standards, and assessment tasks which are set according to the expected learning outcomes for interpretation of student performance. Through checking the student's progression against expected outcomes, the teacher can get a clear idea about the student's growth in that area of learning, and from this, the teacher can determine subsequent actions to enhance further learning. As such, standards-referenced assessment provides a means to align curriculum, assessment and teaching and so gives meaning to assessment, enabling assessment *of* learning to be developmental. Instead of ranking students according to their performance, assessment is used to provide teachers with information about where their students are in their learning. In their chapter, Tognolini and Davidson (2012, this volume) explain how standards are defined and how they are used to improve classroom learning and assessment.

Assessment *for* learning is made possible through cognitive diagnostic assessment (CDA) which aims to generate diagnostic insights from assessment data in order to inform subsequent instructional decisions. Three chapters in this volume (Choi et al. 2012; de la Torre 2012; Kalyuga 2012) are devoted to the theory of CDA and how, under this theory, assessment can be designed to generate specific and fine-grained information about the learner's current knowledge and skills in order to facilitate assessment *for* learning.

Many models are available in the literature for CDA (see Choi et al. 2012, this volume) for an overview, or DiBello et al. (2007) for a review, but as a basic first step, the test designer needs to analyse the knowledge structure to identify and define the attributes underpinning the learning. Next, assessment items are constructed with contents designed to generate diagnostically relevant information on the knowledge and skills of interest. The matrix specifying the item and target attribute relationship is called a Q-matrix (Tatsuoka 2009). Construction of the Q-matrix involves many iterative rounds of theoretical mapping of attributes by content experts and empirical testing of the items for representation of attributes. Third, a psychometric model is selected to analyse the assessment data in order to identify the learner's strengths and weaknesses on the attributes. One family of psychometric models for CDA, namely the 'deterministic, input, noisy "and" gate'

(DINA) model, is highlighted with illustrative examples (using a mixed fraction subtraction problem) by de la Torre (2012) in this volume. Lastly, feedback on strengths and weaknesses of individual learners is given in order to facilitate instructional decisions by the teacher and each learner.

CDA has strong potential to give diagnostic insights into learning. Nevertheless, the construction of a Q-matrix is a very demanding task, and misspecification of the Q-matrix can lead to serious misinterpretations of performance data (Rupp and Templin 2008). Furthermore, there is no easily accessible computer software for analysis of assessment data involved in DINA or other models (Choi et al. 2012, this volume; de la Torre 2012, this volume). In particular, CDA usually requires a large number of examinees to be assessed on a considerable number of items to obtain reliable estimates. Kalyuga (2012, this volume) offers a rapid diagnostic assessment approach as an alternative. Carried out either as a first-step method, wherein a learner is invited to rapidly indicate their first step to solve a problem, or a rapid verification method, wherein the learner is asked to rapidly verify the accuracy of a series of steps towards a solution, the rapid diagnostic assessment method can be used to provide diagnostic information on the learner's current knowledge state (Kalyuga 2012, this volume).

Although CDA is gaining in popularity in education (Lee and Sawaki 2009; Leighton and Gierl 2007), its use remains limited because of its technical and psychometric complexities. Instead, the Rasch model (Boone et al. 2012, this volume) is perhaps more accessible to classroom teachers. The Rasch model is a statistical model that expresses the probability of a response (e.g. right/wrong answer) in terms of a logistic function of the difference between the ability of the person taking the test (represented by θ) and the difficulty level of an item (represented by δ). The probability of getting an item correct is given by $e^{(\theta-\delta)}/(1+e^{(\theta-\delta)})$,where e is the exponential function. It can be easily seen that if θ equals δ, the probability of getting an item correct is 0.5. However, if θ is greater than δ, i.e. if the person has more ability than what is demanded by the difficulty of the item, then the person has a greater than 0.5 probability of getting the item correct. And the converse is also true: if θ is smaller than δ, i.e. if the person has less ability than what is demanded by the difficulty of the item, then the person has less than 0.5 probability of getting the item correct. Graphically (see Fig. 1.2), the trait being tested can be represented by a vertical continuum, and the person ability and item difficulty on the left and right sides of the vertical continuum, respectively. It is then easy to illustrate the three situations: (a) the person has a high probability of passing the item, (b) the person has a 50/50 chance of passing the item and (c) the person has a low probability of passing the item (Fig. 1.2).

Most commercially available software packages of Rasch analysis, for instance Winsteps® (Linacre 2011), produce an item-person map for all persons taking the test and all items in the test. The example given in Fig. 1.3 shows that person A17 has ability well above items 1, 5, 30 and 7; thus she/he has a high chance of answering these items correctly. However, A17's ability is about the same as the difficult level of items 3, 9 and 18 and so she/he has only a 0.5 chance of getting these items correct. Items around this area of difficulty represent A17's zone of proximal

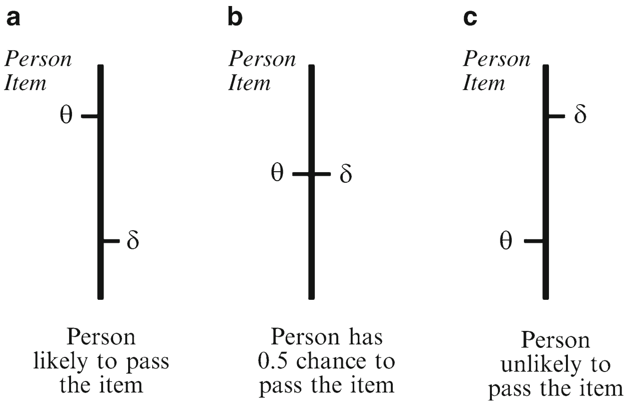
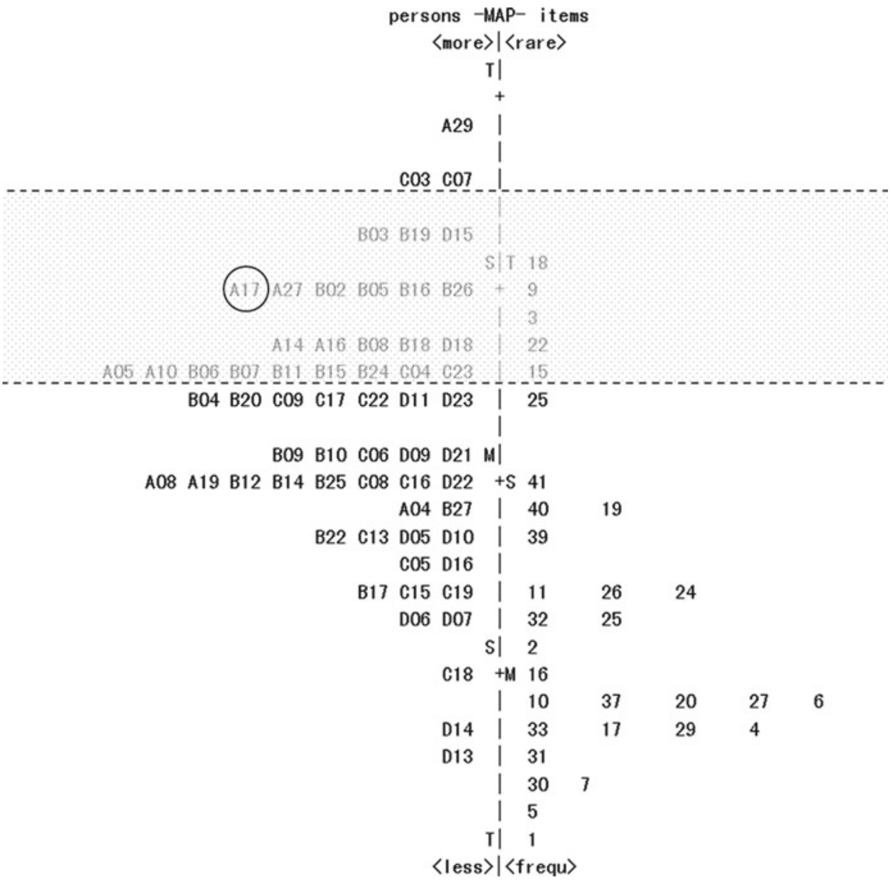

a
Person
Item
θ
δ
Person likely to pass the item
b
Person
Item
θ
δ
Person has 0.5 chance to pass the item
c
Person
Item
δ
θ
Person unlikely to pass the item

Fig. 1.2 Person ability verses item difficulty



```
                             persons -MAP- items
                                 <more>|<rare>
                                      T|
                                       +
                                A29    |
                                       |
                            C03 C07    |
-------------------------------------------------------------------------
                        B03 B19 D15    |
                                      S|T 18
              A17 A27 B02 B05 B16 B26  +  9
                                       |  3
                  A14 A16 B08 B18 D18  |  22
      A05 A10 B06 B07 B11 B15 B24 C04 C23  |  15
-------------------------------------------------------------------------
          B04 B20 C09 C17 C22 D11 D23  |  25
                                       |
                  B09 B10 C06 D09 D21 M|
      A08 A19 B12 B14 B25 C08 C16 D22 +S  41
                            A04 B27    |  40      19
                    B22 C13 D05 D10    |  39
                            C05 D16    |
                        B17 C15 C19    |  11      26      24
                            D06 D07    |  32      25
                                      S|  2
                                C18   +M  16
                                       |  10      37      20      27      6
                                D14    |  33      17      29      4
                                D13    |  31
                                       |  30   7
                                       |  5
                                      T|  1
                                 <less>|<frequ>
```

Fig. 1.3 Item-person map showing zone of proximal development for person A17

development (ZPD) (Vygotsky 1978). Using the item-person map, the teacher can find out the ZPD of every student and provide scaffolding accordingly.

Assessment *as* learning in the SLOA framework implies the engenderment in students of the habit of mind for self-monitoring and self-regulation in order to enhance further learning (Earl and Katz 2008, reprinted in this volume; Mok 2010; Pintrich 2004; Schunk 2008). It represents a shift in attention from assessment of subject matters to focusing on the learner's self-awareness. Critical to assessment *as* learning is the learner's metacognition. The literature distinguishes two components of metacognition: 'knowledge of cognition' (knowledge about oneself as learner and knowledge about strategies to learn) and 'regulation of cognition' (the conscientious control by the learner of various cognitive strategies for learning, including planning, regulation and evaluation) (Brown 1987).

Kleitman et al. (2012, this volume) identify self-confidence as an important aspect of metacognitive knowledge. Their programme of research in Australia and Sweden found that children as young as 9 years old can clearly articulate their own confidence judgments. They also found that the construct of self-confidence predicts school achievement even after controlling for students' cognitive ability, age and gender. Furthermore, self-confidence is affected by classroom goal orientation (Meece et al. 1988), teacher-student relations and after-school activities. These results have significant implications for the development of assessment *as* learning in students.

1.3 Implementation Strategies of SLOA in Schools

We have gained invaluable experiences from our 3-year longitudinal assessment project (2005–2008) in Hong Kong and the self-directed learning orientated assessment (SLOA) projects in Macau and China (2008–2009) (Mok 2010; Yu 2012a, b, this volume). These experiences show that successful implementation of SLOA has to be multilevel, instigating change at system, school, classroom, teacher and student levels, and that it also requires concerted effort by all the key actors, including parents, principals, teachers, students, government officials, educators and societal leaders.

Although there is no single set of strategies that suits all situations for successful reforms, previous experience with more than 100 schools suggests three strategies that tend to predict higher chance of success:

- Taking a whole-school approach to building a strong culture of SLOA in curriculum redesign
- Empowering teachers through development of knowledge and skills
- Activating students as learning partners

A whole-school approach means that all key stakeholders are involved and that clearly articulated management and implementation plans are established at all levels (Stringfield et al. 2008). Importantly, when faced with assessment data, new policies may need to be developed and curriculum may have to be redesigned. These actions are not achievable by individual teachers, learners or even the principal alone. Instead, a whole-school approach enables assessment data to be turned into actionable knowledge.

A whole-school approach creates a learning community in which teachers can experiment with new approaches in unison (Day 2008). School support strengthens teachers' identification with the school (Henkin and Holliman 2009); increases teacher work satisfaction (Day 2008), teacher commitment (Cheung and Wong 2011; Choi and Tang 2009; Day 2008) and teachers' willingness to implement innovations (Ballet and Kelchtermans 2009); and raises student achievement (Day 2008) (although an earlier research by Park (2005) has a different finding).

Research (Cheung and Wong 2011; Earl 2011; Fullan 2009; Pitiyanuwat and Pitiyanuwat 2012, this volume; Sahlberg 2006) found quality teachers to be a key factor to success of education reforms. Teacher professional development, especially at times of change, empowers teachers to initiate and sustain changes in their classrooms (Ballet and Kelchtermans 2009; Cheung and Wong 2011; Goh and Matthews 2012, this volume; Lieberman and Pointer Mace 2008; Pitiyanuwat and Pitiyanuwat 2012, this volume; Yu 2012a, b, this volume). The quality of professional development programmes as measured by their relevance, meaning, practical values and flexibility in choice on the format and time affects teachers' willingness to participate (Day 2008).

Although teachers can be drivers of reform, it is the students themselves who need to commit to change (Earl 2003; Earl and Katz 2008, reprinted in this volume). Self-regulated learning is facilitated by a learning environment with a community of learners and in the context of cooperative learning. An open and autonomous classroom encourages peer students to serve as resource persons and partners in the learning process (Mok 2010; Paris and Paris 2001; Slavin 1996).

1.4 Tools for the Implementation of SLOA

A number of new developments in assessment and psychometrics are now available to support the implementation of SLOA. Part II of this book focuses attention on how to harness new developments in psychometrics and information technology to facilitate assessment for learning. Six tools for the implementation of SLOA are introduced: item response theory, mathematics competency vertical scales, student-problem charts, dynamic assessment, two-tier items and computerized adaptive testing. These tools have in common an emphasis on the speedy generation of valid diagnostics feedback to inform instruction. They are presented as accessible alternatives to the traditional method of using the total score as an indicator of level of achievement.

The contrast between item response theory (IRT) and classical test theory (CTT) is presented by Wu (2012, this volume) with an example data set analysed using the ConQuest programme (Wu et al. 2007) that she developed. The IRT is a mathematical model representing the relationship between an examinee and a test item. Wu (2012, this volume) discusses the concepts of item difficulty, discrimination power and plausible values in IRT in this chapter.

A critical concept in IRT is the assumption of a single latent ability (construct) underpinning an examinee's performance on a test. Suppose the latent ability in question has meaning across several school years, then in theory, a vertical scale can be built to chart a student's progress across year levels on the construct. Yan et al. (2012, this volume) present a new method, entitled concurrent-separate approach, using the Rasch model (Boone et al. 2012, this volume) to develop vertical scale of measurement across several school levels. The authors demonstrated how a mathematics competency vertical scale (MCVS) with reasonable psychometric properties can be developed using the new method and made feasible to track Hong Kong students' development in mathematics from primary 2 (grade 2) to secondary 3 (grade 9) levels.

The Rasch model (Boone et al. 2012, this volume) was found by many teachers in Hong Kong and Macau to be helpful to their provision of quality feedback to students (Mok 2010). Nevertheless, to some teachers, the Rasch model can be mathematically demanding. The student-problem chart (SP chart) (Mok et al. 2012, this volume; Sato 1980, 1985) is an alternative for teachers to make sense of assessment data. The SP chart is a matrix of students' responses to individual items of a certain assessment in which the rows and columns are rearranged such that students are arranged from high to low ability (based on their total score on the assessment), and items are arranged in ascending order of difficulty from left to right (based on the number of students who answered the item correctly). After this rearrangement, the observed pattern of responses is matched against the expected pattern, which is computed based on the assumption that each student has a higher probability of answering correctly an easier item than a more difficult item and, likewise, each item has a higher probability of being answered correctly by a more able than a less able student. By inspecting the response pattern and interpreting it against the expected pattern, the teacher is able to identify aberrant response behaviours of students. Furthermore, a modified caution index (Sato 1980) can be computed based on the SP chart for each student and each item to enable the teacher to determine if the response pattern is too different from the expected pattern and, if so, how they are different. The teacher is able to give evidence-based feedback to the students on subsequent learning. A software package SP Xpress (Mok et al. 2011) is now available for producing the modified caution index, item reliability, student performance and other psychometric indices for use by teachers.

IRT, vertical scales and SP chart are helpful tools which can be used to analyse assessment data to support student learning. Nevertheless, it is impossible to undertake high quality analysis if the assessment data itself is of substandard quality. The chapter by Tam et al. (2012, this volume) presents the method of two-tier items to provide high quality diagnostic insights. A two-tier item is conceptualized by the authors as a mini-test (testlet) comprising two parts: the first part is usually designed to assess the examinee's ability to identify the targeted concept, and the second the extent to which the examinee can explain the rationale for the response on the first part. By design, the two parts of the testlet are related and thus violate the underlying assumption of local independence in the Rasch approach. In their chapter, Tam et al. (2012, this volume) propose a method to analyse two-tier items and illustrate with a real data set how the data can be analysed for diagnostic insights.

One important consideration in the implementation of SLOA is the speed at which assessment feedback is generated. This is especially so for classroom assessment. In Chap. 13, Tzuriel (2012, this volume) presented dynamic assessment as an attractive solution to speedy assessment feedback. Dynamic assessment is based on the author's three decades of work in this area and is underpinned by the theory of zone of proximal development (ZPD) developed by Vygotsky (1978). In this approach, assessment and learning are tightly integrated through an iterative process of 'assessment to ascertain the ZPD, teaching around the ZPD, learning and further assessment to ascertain new ZPD'. In a systematic presentation in six major sections, Tzuriel (2012, this volume) argues for the shift from standardized testing to dynamic assessment to support assessment for learning, and he also discusses the benefits, limitations and strategies in using this new approach.

Speedy assessment feedback can be achieved through computerized adaptive testing (CAT) as presented by Hsu et al. (2012, this volume). The CAT technology capitalizes on recent developments in psychometric theory, particularly IRT (Wu 2012, this volume) and information technology. A CAT system comprises an item bank that is constructed and calibrated according to a vertical scale about a trait (Yan et al. 2012, this volume); a set of item-selection strategies for the iterative process of 'selection of an initial batch of test items, response by examinee, analysis on response and generation of the next batch of test items' in order to elicit the optimal amount of information about the examinee's level of competence on the trait; and a stopping rule which specifies criteria for the iterative process to stop. With the availability of fast speed computers, CAT can be used for large-scale assessment as well as classroom applications.

1.5 Examples of Implementation in the Asia-Pacific Region

Part III of the book presents six case studies of SLOA being implemented in the Asia-Pacific region in Thailand, Malaysia, China and Hong Kong. The first case is contributed by Pitiyanuwat and Pitiyanuwat (2012, this volume), who write on the history of assessment reform in Thailand and how the reform has evolved from a 'non-formal' form of education in the period from year 1283–1883 to the contemporary period wherein the alignment between assessment and learning is emphasized. The analysis by the authors not only shows the pathway to SLOA, the hurdles, pitfalls, rewards and achievements involved but also gives hope and direction for other Asia-Pacific education systems who are tempted to try SLOA in their own system.

One of the major areas of education reform in the Asia-Pacific region is the building up of a strong teaching force to drive the reform (Mok et al. 2003). The second and third case studies are both concerned with teacher capacity in driving assessment reforms. In the second case study, in response to the desire of the Malaysian Ministry of Education to evaluate teacher education, Goh and Matthews (2012, this volume) examined pre-service teachers' ability for self-assessment. Through the voices of 16 pre-service teachers in Malaysia, the authors raise questions regarding the development of teacher self-metacognition – questions that are of critical importance for the successful implementation of SLOA.

The third case study also focuses on teacher capacity to implement assessment reform. This case study, undertaken by Ho et al. (2012, this volume), explores attitudes of teachers in Hong Kong towards Rasch measurement, particularly regarding the desirability and feasibility of the Rasch model as a tool for assessment for learning. Their findings suggest that although teachers recognize the Rasch model as a powerful alternative to traditional methods in generating assessment feedback, their adoption of the model for classroom applications is impeded by realistic workplace concerns including heavy workload and lack of technical support. Given that teachers' attitudes affect their instructional decisions and willingness to adopt new approaches in their teaching (Choi and Tang 2009; Day 2008), it is important that teachers are supported in their implementation of assessment reform. Targeted professional development workshops, partnership with universities and provision of assessment item banks are proposed by Ho et al. (2012, this volume) as possible solutions to overcome the difficulties perceived by teachers.

The fourth case study, reported by Yu (2012a, this volume), involves trial implementations of SLOA with 209 primary students in three schools in China. By using metacognitive reading strategies, a specially designed reading log and a Rasch-calibrated English reading assessment system, Yu demonstrated that the SLOA approach significantly affected several aspects of teaching and learning of English reading in these schools, including changes in the physical learning environment, teacher motivation and teacher knowledge, as well as improvement in students' English reading proficiency. Story book reading and metacognitive methods to promote reading are not entirely new strategies in the teaching of English reading, but in a country that has a long history of teacher-centred instruction, these approaches are innovative and have deep and far-reaching implications.

Encouraging results are also reported in the fifth case study (Lee 2012, this volume). Lee's study involves an intervention designed to support pre-service sports coaches in the implementation of assessment for learning in the teaching of sports in Hong Kong. Through a series of carefully designed experimental procedures, Lee successfully instilled in his pre-service sports coaches in the experimental group the skills and strategies for using feedback to promote sports learning.

The sixth case study presented by Yu (2012b, this volume) reports on the implementation of SLOA in the subject of English at Saint Margaret's Girls' College in Hong Kong. Yu has provided for the readers a very detailed account on the rationale behind the SLOA project and illustrated vividly, using quotations taken from students' and teachers' journals, the impacts of SLOA on English instruction across several different year levels at the school. Although the results found by the study do not have statistical significance, the case report carries with it great substantive significance because through its rich description and the testimonies given by the actors, there is strong evidence of how the study has changed the school's approach to assessment for learning.

1.6 Conclusion

Assessment is a concept with a long history. It has special meanings to people in the Asia-Pacific region where assessment and high-stake examination used to be synonymous. Since the turn of the century, however, systems in the region have initiated many reforms to catch up with worldwide paradigm shifts in the conception of assessment from assessment *of* learning to assessment *for* and *as* learning and to face the new challenges of the twenty-first century. Globalization and knowledge economies demand that we revise our vision on pedagogy to one that centres on learning how to learn. This volume presents a framework entitled self-directed learning oriented assessment (SLOA) which is strongly grounded in cognitive learning theory, powered by psychometric tools, and has been validated in several systems in the region as an enabling device for the betterment of learning and teaching in the new century.

References

Abiko, T. (2011). A response from Japan to TLRP's ten principles for effective pedagogy. *Research Papers in Education, 26*(3), 357–365.

Ballet, K., & Kelchtermans, G. (2009). Struggling with workload: Primary teachers' experience of intensification. *Teaching and Teacher Education, 25*(8), 1150–1157.

Berry, R., & Adamson, B. (Eds.). (2011). *Assessment reform in education: Policy and practice*. New York: Springer.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: School of Education King's College.

Black, P., McCormick, R., James, M., & Pedder, D. (2006). Learning how to learn and assessment for learning: A theoretical inquiry. *Research Papers in Education, 21*(2), 119–132.

Black, P., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to the navigation of learning progression. *Measurement, 9*, 71–123.

Boone, W. J., Staver, J. R., & Yale, M. S. (2012). Theory of self-directed learning oriented assessment: A non-technical introduction to the theoretical foundations and methodologies of cognitive diagnostic assessment. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. Weinert & R. Kluwe (Eds.), *Metacognition, motivation and understanding* (pp. 65–116). Hillsdale: Erlbaum Associates.

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International, 44*(1), 57–66.

Cheung, A. C. K., & Wong, P. M. (2011). Effects of school heads' and teachers' agreement with the curriculum reform on curriculum development progress and student learning in Hong Kong. *International Journal of Educational Management, 25*(5), 453–473.

Choi, P. L., & Tang, S. Y. F. (2009). Teacher commitment trends: Cases of Hong Kong teachers from 1997 to 2007. *Teaching and Teacher Education, 25*(5), 767–777.

Choi, H. J., Rupp, A. A., & Pan, M. (2012). Standardized diagnostic assessment design and analysis: Key ideas from modern measurement theory. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Commission, E. (2000). *Review of education system reform proposals: Consultation document: Education blueprint for the 21st century*. Hong Kong: HKSAR Government Printing Department.

Curriculum Development Council. (2001). *Learning to learn: Life-long learning and whole-person development*. Hong Kong: The Education Department, Hong Kong SAR.

Day, C. (2008). Committed for life? Variations in teachers' work, lives and effectiveness. *Journal of Educational Change, 9*, 243–260.

de la Torre, J. (2012). Application of the DINA model framework to enhance assessment and learning. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Delors, J., Al Mufti, I., Amagi, I., Carneiro, R., Chung, F., Geremek, B., Gorham, W., Kornhauser, A., Manley, M., Quero, M. P., Savane, M. P., Singh, K., Stavenhagen, R., Suhr, M. W., & Nanzhao, Z. (1996). *Learning: The treasure within*. Paris: UNESCO.

DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 979–1030). Amsterdam: Elsevier.

Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks: Corwin Press.

Earl, L., & Katz, S. (2008). Getting to the core of learning: Using assessment for self-monitoring and self-regulation. In S. Swaffield (Ed.), *Unlocking assessment: Understanding for reflection and application* (pp. 90–104). London: Routledge/Taylor & Francis. (Reprinted in this volume, with permission from Taylor & Francis.)

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911.

Fullan, M. (2009). Large-scale reform comes of age. *Journal of Educational Change, 10*, 101–113.

Goh, P. S. C., & Matthews, B. (2012). Concerns of student teachers: Identifying emerging themes through self-assessment. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Henkin, A. B., & Holliman, S. L. (2009). Urban teacher commitment: Exploring associations with organizational conflict, support for innovation, and participation. *Urban Education, 44*(2), 160–180.

Ho, C. M., Leung, A. W. C., Mok, M. M. C., & Cheung, P. T. M. (2012). Informing learning and teaching using feedback from assessment data: Hong Kong teachers' attitudes towards Rasch measurement. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Hogan, D., & Gopinathan, S. (2008). Knowledge management, sustainable innovation, and pre-service teacher education in Singapore. *Teachers and Teaching: Theory and Practice, 14*(4), 369–384.

Hogan, D., Towndrow, P., & Koh, K. (2009). The logic of confidence and the social economy of assessment reform in Singapore: A new institutionalist perspective. In E. Grigorenko (Ed.), *Assessment of abilities and competencies in the era of globalization*. New York: Springer.

Hong Kong Baptist University and Hong Kong Examinations Authority. (1998). *Review of public examinations system in Hong Kong: Final report (The ROPES report)*. Hong Kong: Author.

Hsu, C. C. L., Zhao, Y., & Wang, W. C. (2012). Exploiting computerized adaptive testing for self-directed learning. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Japan Ministry of Education. (2000). *Aims and objectives of education*. Retrieved August 28, 2010, from http://www.monbu.go.jp/aramashi/1999eng/e03/e03-1.htm

Kalyuga, S. (2012). Rapid dynamic assessment for learning. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Kleitman, S., Stankov, L., Allwood, C. M., Young, S., & Mak, K. (2012). Metacognitive self-confidence in school-aged children. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Knowles, M. (1975). *Self-directed learning: A guide for learners and teachers*. Chicago: Association Press and Follett Publishing Company.

Korean Ministry of Education. (2000). *Aims and objectives of education*. Retrieved August 28, 2010, from http://www.moe.go.kr/english/edukorea/edukorea1/htm

Lee, H. K. O. (2012). Physical education in higher education in Hong Kong: The effects of an intervention on pre-service sports coaches' attitudes towards assessment for learning used in sports. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Lee, Y.-W., & Sawaki, Y. (2009). Cognitive diagnosis and Q-matrices in language assessment. *Language Assessment Quarterly, 6*(3), 169–171.

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and application*. Cambridge: Cambridge University Press.

Lieberman, A., & Pointer Mace, D. H. (2008). Teacher learning: The key to educational reform. *Journal of Teacher Education, 59*(3), 226–234.

Linacre, J. M. (2011). *Winsteps* (version 3.72.3) [computer software]. Chicago: Winsteps.com.

Loyens, S. M. M., Magda, J., & Rikers, M. J. P. (2008). Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational Psychology Review, 20*(4), 463–467.

Maclean, R. (2010). Introduction by professor Rupert Maclean. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment: Assessment that informs learning and empowers the learner* (p. vi). Hong Kong: Pace Publications.

Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology, 80*, 514–523.

Mok, M. M. C. (2010). *Self-directed learning oriented assessment: Assessment that informs learning & empowers the learner*. Hong Kong: Pace Publications.

Mok, M. M. C., Gurr, D., Izawa, E., Knipprath, H., Lee, I. H., Mel, M. A., Palmer, T., Shan, W. J., & Zhang, Y. (2003). Quality assurance and school monitoring. In J. P. Keeves & R. Watanabe (Eds.), *International handbook of educational research in the Asia-Pacific region* (Kluwer international handbooks of education, Vol. 11, pp. 945–958). Dordrecht: Kluwer Academic.

Mok, M. M. C., Ting, Y. C., Ho, H. S., Wong, Y. W., Tse, C. N., Xu, J., & Yao, J.-J. (2011). Optimising learning oriented assessment: SP Xpress 2.2. Hong Kong: Pace Publications Ltd. (Published in Chinese: 莫慕貞、丁彥銓、何昊璇、黃英華、謝棹南、徐坤、姚靜靜 (2011). 優化學習導向評估之SP Xpress 2.2. Hong Kong: Pace Publications Ltd.).

Mok, M. M. C., Lam, S. M., Ngan, M. Y., Yao, J. J., Wong, M. Y. W., Xu, J. K., & Ting, S. Y. C. (2012). Student-problem chart: An essential tool for SLOA. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Ng, P. T. (2010). The evolution and nature of school accountability in the Singapore education system. *Educational Assessment, Evaluation and Accountability, 22*(4), 275–292.

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist, 36*(2), 89–101.

Park, I. (2005). Teacher commitment and its effects on student achievement in American high schools. *Educational Research and Evaluation, 11*(5), 461–485.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385–407.

Pitiyanuwat, S., & Pitiyanuwat, T. (2012). Learning assessment reform in Thailand. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Putwain, D. W. (2009). Assessment and examination stress in key stage 4. *Educational Research, 51*(3), 391–411.

Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96.

Sahlberg, P. (2006). Education reform for raising economic competitiveness. *Journal of Educational Change, 7*(4), 259–287.

Sato, T. (1980). The S-P chart and caution index. In *NEC educational information bulletin* (pp. 80–81). Tokyo: C&C Systems Research Laboratories, Nippon Electric Co., Ltd.

Sato, T. (1985). *Introduction to student-problem curve theory analysis and evaluation*. Tokyo: Meiji Tosho.

Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational Psychology Review, 20*, 463–467.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Slavin, R. E. (1996). Research for the future. Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology, 21*, 43–69.

Stringfield, S., Reynolds, D., & Schaffer, E. C. (2008). Improving secondary students' academic achievement through a focus on reform reliability: 4- and 9-year findings from the high reliability schools project. *School Effectiveness and School Improvement, 19*(4), 409–428.

Taiwan Ministry of Education. (2001). *Aims and objectives of education*. Retrieved December 12, 2001, from http://www.edu.tw:81/english/

Taiwan Ministry of Education. (2011, October 26). *Towards a learning society – Appendix*. Retrieved August 28, 2010, from http://english.moe.gov.tw/content.asp?CuItem=751

Tam, H. P., Wu, M., Lau, D. C. H., & Mok, M. M. C. (2012). Using user-defined fit statistic to analyse two-tier items in mathematics. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge Academic.

Thailand Ministry of Education. (2000). *Aims and objectives of education*. Retrieved August 28, 2010, from http://www.moe/go.th/English/Mla/default.htm

Tognolini, J., & Davidson, M. (2012). Assessment, standards-referencing and standard setting. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Tzuriel, D. (2012). Dynamic assessment of learning potential. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Vygotsky, L. S. (1978). Interaction between learning and development (M. Lopez-Morillas, Trans.). In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes* (pp. 79–91). Cambridge, MA: Harvard University Press.

Wu, M. (2012). Using item response theory as a tool in educational measurement. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2: Generalised item response modelling software*. Camberwell: Australian Council for Educational Research.

Yan, Z., Lau, D. C. H., & Mok, M. M. C. (2012). A concurrent-separate approach to vertical scaling. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Yu, G. (2012a). Accelerated approach to primary school English education in China: Three case studies. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Yu, G. (2012b). The case of St Margaret's Girls' college: How SLOA promotes self-assessment and peer assessment to enhance secondary school student English learning. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer.

Chapter 2
Assessment, Standards-Referencing and Standard Setting

Jim Tognolini and Michelle Davidson

2.1 Assessment

2.1.1 The Meaning of Assessment

Assessment can be defined as the collection of information about student performance for a purpose. In education, students are generally assessed for the purpose of improving their learning and monitoring and certificating their performance or achievement.

Teachers collect information about student performance (assess them) in numerous ways. These have been summarised along a continuum of assessment methods (Fig. 2.1) that range from 'less formal or unstructured methods' to 'more formal or highly structured methods' of collecting information.

At the highly structured end, there are examinations, published tests and tests such as national and state-based testing programmes. These are highly structured in that the conditions of administration are tightly controlled and standardised and the tests have been through rigorous test construction processes.

Classroom tests, checklists, practical work, project work, etc., are also methods for collecting information about students. They are not as formal in their structure, but they provide information that is just as pertinent and relevant about a student as the more highly structured means of collecting information and they happen much more often.

J. Tognolini (✉)
Senior Vice President, Research and Assessment, Pearson plc and Senior Research Fellow at Oxford University, Sydney
e-mail: Jim.Tognolini@Pearson.com

M. Davidson
Teaching and Learning (Assessment), Trinity Grammar School, Sydney, Australia
e-mail: michelle.davidson@live.com.au

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_2, © Springer Science+Business Media Dordrecht 2013

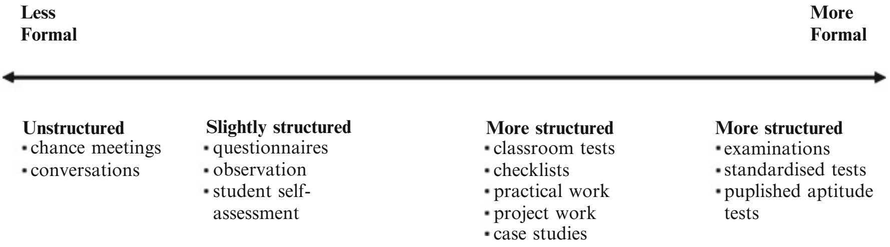


Fig. 2.1 Methods for collecting information on student performance

2.1.2 Reporting of Assessments

Once teachers have collected information and formed the image, they then have the task of reporting that image to students, parents, employers and the general community. Reporting happens fairly regularly in schools. In most cases, the report is prepared by the school and presented to the students as a testimony of their performance at this stage in their development.

Traditionally, in most cases of school reporting, the student marks are compared with the other student marks in the same subject in the same school. As such, comparability is not a major issue as long as the teacher is internally consistent.

However at other key stages in the educational process, the information is used to formally certify that students have reached a ‘milestone’. In this case, the marks have to be comparable across all students across the country in that subject for that year group. The most common way to achieve such comparability is to ensure that all students have the same tasks given under standardised conditions. The assessments that are used in this latter case are generally based upon formal examinations. The results that evolve from such examinations have to be directly compared because they are generally centrally certified.

It is possible, for example, to directly compare a score of 65 from one school to a score of 65 from another school when everyone has taken the same test or examination under standardised conditions. However, if this is not the case, then such comparisons are dubious, to say the least, without some effort to ensure comparability.

The next section considers how meaning has been given to marks over the history of testing.

2.2 Standards-Referencing

2.2.1 Giving Meaning to Student Achievement: Norm-Referencing

While marks have traditionally been used to summarise student achievement, it must be remembered that marks by themselves have no clear meaning. For example, what does a mark of 65 mean? One piece of information that is required to fully understand the meaning of a score of 65 is the maximum score for the examination; not all examinations are necessarily scored out of 100. If the examination was out of 150 or 500, then the meaning of 65 changes quite significantly.

A second piece of information that is required is the mean or average mark of the group of students on the examination. Such a summary statistic gives an idea of 'what marks everybody else got'. This average gives an indication of the relative difficulty of the test for the group of students taking the examination. In some situations, the spread of marks on the test (standard deviation) is also provided to give an indication of how the scores of the students on the test were spread out around the average or mean. If the 65 is obtained on an examination where the average is 50, then the 65 would suggest a very good performance as the examination is relatively hard for the students; if the average for the group is 90, then the examination would appear to be relatively easy, and the mark of 65 would suggest a poor performance on the examination relative to the other students. In other words, the mark is given meaning by comparing it to the performance of a group of students who sat the examination (commonly referred to as the 'norming group').

Giving marks meaning by referencing the marks to the marks of the norming group is referred to as *norm-referencing*. It has served educators (and the educational community as a whole) well since the introduction of formal examination procedures.

One of the main advantages of norm-referencing is that the marks, grades or awards are interpreted in the same way from situation to situation (year to year; subject to subject). For example, a distinction can be awarded each year to those students who are in the top 20% of the group taking the examination. Similarly, a mark of 50 can be set to be the pass mark by assigning it to the highest mark of the bottom 30% of students. This means for example, that each year 20% of students in a subject will receive a distinction and 30% will fail.

This approach to reporting marks is the method used by teachers and examining authorities to give meaning to student marks around the world. Generally, most people do not realise that the marks that they get have been adjusted (scaled) by referencing them to the achievement of the norming group.

There are numerous limitations associated with referencing marks to the performance of the group. One obvious weakness is that the result has no reference to the standard of the performance. While the examination (or any other assessment) potentially provides a wealth of content information, the resulting report indicates the location of the student relative to the norming group and that is all. It provides little meaningful information about what students know and can do other than the implicit understanding made about the standard of the group performance remaining constant from year to year and from subject to subject.

Peddie (1992) makes the point that it has been concerns like the one expressed above that has led to a search for more appropriate ways of reporting student achievement:

> … a number of related concerns have led to many teachers and some members of the public wanting a different system. They wanted each learner to be tested simply to see what that learner knew and could do. (Peddie 1992, p. 23)

Over the years, there have been attempts to move the process of reporting achievement from referencing to norms, towards methods that better capture the image of what it is the students know and can do.

2.2.2 Giving Meaning to Student Achievement: Criterion-Referencing

Criterion-referencing did just that; all the behaviours that students might be expected to demonstrate during the course are listed, and the teachers then record and eventually report when (and how often) the students actually demonstrate the behaviour.

The process of giving meaning to student performance by referencing it to specified criteria is called criterion-referencing (Popham 1978).

The main problem with criterion-referencing is that it is very labour intensive. It is also characterised by an atomisation of the curriculum into multiple behaviours, and it has become inextricably linked with the behavioural objectives and mastery learning movements. However, it does go some way towards reporting the image that the teacher and the examiner develop during the teaching/learning and examining processes.

More recently, a number of education systems around the world have introduced a different way to reference achievement. It builds upon criterion-referencing, but instead of referencing achievement to the myriad of behaviours that comprise an examination, course or subject, the achievement is referenced to pre-determined standards of performance. It is referred to as standards-referencing.

2.2.3 Giving Meaning to Student Achievement: Standards-Referencing

One of the main differences between norm- and standards-referencing is that with the latter there is no inherent limit to the percentage of students achieving a particular standard. In theory, it is possible for all students to achieve a performance standard although in practice this is unlikely because the standards have generally been constructed drawing on the experience of normative data. In other words, norms generally underpin performance standards. This point is made by Elley (2004) when he states:

> Of course, the standard we set for an SBA (Standards-based Assessment) decision is strongly influenced by what the norms are, or have been in the past. If the standard set is too high, then hardly anyone will pass. And if it is too low, then everyone will pass. Setting the appropriate standard is a problem in its own right. (Elley 2004, p. 1)

With a standards-referenced system, performance relative to standards can be measured and monitored over time. This is not possible in a norm-referenced system in which the distributions are pre-determined and the reported performance relative to standards appears fixed.

The trade-off for meaningful and relevant performance data is that the items, questions and tasks that comprise the examinations have to be more carefully designed and more closely linked to the learning outcomes specified in the curriculum being assessed.

Elley (2004) captures some of the dangers associated with adopting a standards-referenced system when he comments on the New Zealand experience:

> …, in my view, SBA (Standards-based Assessment) works when the standards are clearly defined, or when the knowledge is well organised, with clear limits, a limited number of decisions, a progression of difficulty, or when the tasks are pre-tested for difficulty and uniform for all students, or when the assessment is low-stakes. BUT, what do we have?
>
> In NCEA (National Certificate of Educational Achievement), the knowledge base for academic subjects is complex and multi-dimensional, the so-called standards are fuzzy, the questions are not pre-tested for difficulty, and they are not standard from year to year, and the assessments are definitely high stakes. (Elley 2004, p. 5)

The problems outlined by Elley (2004) are not unique (see Donnelly 2000; Manno 1994; Shanker 1994). They are prevalent in the introductory stages of all standards-referenced systems and as such, need to be addressed in any attempt to introduce a standards-referenced system.

2.2.4 Characteristics of Standards-Referenced Systems

A standards-referenced system comprises a curriculum (syllabus or framework) that describes through its statement of aims, objectives, learning outcomes (content standards) and content, what it means to grow in an area of learning.

Teaching and learning is based on the curriculum. The most important sources of information for the curriculum are the learning outcomes or content standards.

Assessment tasks in a standards-referenced system must be directly linked to the learning outcomes or content standards and the performance standards. They provide students with the opportunity to demonstrate what it is they know and can do in relation to the curriculum for the learning area.

While the process of referencing performance to standards is quite straightforward, there are numerous points in the process that require judgement and interpretation and present significant challenges to examiners and teachers. For example, the learning outcomes are intended to describe what it means to grow or progress through an area of learning. This path is not deterministic, and hence there is scope for this developmental sequence to be challenged by data; setting examination questions that accurately assess the learning outcomes, that are consistent with the requirements of the performance standards and that are technically correct is a challenging task; setting marking keys and rubrics that are both fair and accurate is a challenging task; ensuring that marking keys and rubrics are consistently applied is a challenge; accurately establishing the performance standards (levels) and presenting them to teachers, examiners and students in a manner in which they will all interpret them consistently is a challenge; and operationally defining the boundaries of the performance standards in the context of external and internal assessments is a challenge.

It is important to reinforce the point here that there is an equally imposing set of challenges (basically the same ones) confronting the establishment of a norm-referenced system. The difference is that the manifestation of the uncertainty that emerges with these challenges is made explicitly obvious in a standards-referenced system through the observed variance in the distribution of results, whereas in a norm-referenced system the errors of measurement are masked by the representation of fixed distributions of results. In a norm-referenced system, for example, if an examination paper is poorly targeted and is too hard, it does not matter. The final result can still have a determined distribution: 10% of the students will be given an A, 20% will be given a B, etc. No one will know that the paper did not provide an adequate opportunity for students to demonstrate what they knew. So long as there is some differentiation of performance, the results reported will have the same final distribution.

Central to a standards-referenced system is the whole notion of standards: how they are defined, how they are set, how they can be used to summarise student performance, how they can be used to report student performance and how they can be used to improve classroom test setting and examination construction of teachers and examiners. The remainder of this chapter considers each of these aspects of standards.

2.2.5 Defining Standards

Donnelly (2000) makes the point that different systems have different meanings for the term ‘standards’:

> The US New Standards Project defines ‘standards’ as: ‘what students should know and be able to do’. This definition is also used by the American Federation of Teachers (see AFT, 1999). Victoria’s CSF, on the other hand, uses ‘standards’ as a synonym for ‘learning outcomes’ which are described as: ‘benchmarks or standards against which student achievement can be measured’. (Donnelly 2000, p. 30)

Victoria’s Curriculum Standards Framework uses ‘standards’ as a synonym for ‘learning outcomes’ which are described as: ‘benchmarks or standards against which student achievement can be measured’.

In New Zealand, standards provide assessment targets and describe the level of work that each learner’s performance can be evaluated against in order to earn credit for a standard on the National Qualifications Framework.

The New South Wales Board of Studies (1999) makes a distinction between *syllabus standards* and *performance standards*.

Syllabus standards describe *what* students are expected to know and understand as a result of studying a course. Other names for standards which have the characteristics of syllabus standards include content standards, grade level standards, core standards and learning outcomes.

Performance standards describe *how well* the students are expected to be able to know and perform the skills included in the syllabus standards. Other names in the literature which generally mean the same as performance standards include achievement standards, benchmark standards, proficiency standards, reporting standards and accountability/target standards.

Performance standards for a subject are generally partitioned into levels of achievement or proficiency levels of performance for that subject. Each level has performance descriptors associated with it. These descriptors may be generic or subject specific. The number of levels specified varies across systems with five or six levels commonly specified. New South Wales, for example, has six-performance levels (bands) for each subject in the New South Wales Higher School Certificate (HSC). Table 2.1 shows the performance levels for HSC Biology.

Table 2.1 Performance bands for New South Wales HSC biology

	The typical performance in this band
Band 6	Demonstrates an extensive and detailed knowledge and superior understanding of biological concepts, including complex and abstract ideas
	Demonstrates an extensive understanding of the historical development of biological concepts, their applications and implications for society and the environment, and the future directions of biological research
	Communicates succinctly, logically and sequentially using a variety of scientific formats, including diagrams, graphs, tables, flow charts and equations relating to biology
	Analyses and evaluates data effectively, identifying biological relationships, quantifying explanations and descriptions and synthesising information to draw conclusions
	Uses precise biological terms extensively and correctly in a wide range of contexts
	Designs valid experimental processes using appropriate technologies and incorporating the thorough knowledge of the use of a control, variables and repetition to solve biological problems
	Applies knowledge and information to unfamiliar situations and designs an original solution to a biological problem
Band 5	Demonstrates thorough knowledge and understanding of most biological concepts
	Demonstrates a thorough understanding of the historical development of biological concepts and their applications and implications for society and the environment
	Communicates effectively in a variety of scientific formats including diagrams, graphs, tables, flow charts and equations relating to biology
	Explains qualitative and quantitative biological relationships and ideas coherently and identifies patterns in data to draw conclusions
	Uses precise biological terms frequently and correctly in a range of contexts
	Identifies the correct application of scientific experimental methodology to solve biological problems
Band 4	Demonstrates sound knowledge and clear understanding of some biological concepts
	Demonstrates a sound understanding of the historical development of biological concepts and their applications for society and the environment
	Communicates using clear written expression and incorporating diagrams of biological structures
	Provides qualitative and quantitative descriptions of biological phenomena and explains straightforward biological relationships
	Uses general biological terms frequently and correctly in a range of contexts
	Identifies the correct components of the experimental scientific method in biology

(continued)

Table 2.1 (continued)

	The typical performance in this band
Band 3	Recalls basic knowledge and understanding of some biological concepts Demonstrates a basic understanding of the historical development of biological concepts and their applications for society and the environment Uses fundamental written communication with some use of simple scientific diagrams relating to biology Provides qualitative descriptions of fundamental biological phenomena and explains some straightforward biological relationships Uses some general biological terms correctly in a limited range of contexts Recalls some aspects of the experimental scientific method in biology
Band 2	Recalls limited knowledge and has elementary understanding of some straightforward biological concepts Demonstrates a limited understanding of the historical development of biological concepts Uses fundamental written communication relating to biology Provides simple qualitative descriptions of biological phenomena Uses general biological terms occasionally
Band 1	

The GCSE (General Certificate of Secondary Education) in the United Kingdom has an eight-point scale A* to G and U. Performance descriptors have been written for the A/B boundary and the C/D boundary for each subject. Ultimately, a single grade is produced for each student in each subject, and the grade is referenced to a description of performance.

In Queensland, results for each subject are reported on a five-point scale: Very High Achievement (VHA), High Achievement (HA), Sound Achievement (SA), Limited Achievement (LA) and Very Limited Achievement (VLA).

In New Zealand, each performance standard is composed of four categories (equivalent to bands or performance levels): students do not achieve the standard (NA), students achieve the standard (A), students achieve the standard with merit (M) and students achieve the standard with excellence (E).

In some US states, each standard (there may be 5–10 for a subject across grades) is partitioned by benchmarks at key locations at the end of a segment work (e.g. Colorado has benchmark descriptors for grades 1–4, 5–8 and 9–12)

The Australian Curriculum Assessment and Reporting Authority (ACARA) uses one performance standard for each grade level, and the standard states what students should be able to know and do by the end of the nominated grade.

2.3 Standard Setting

2.3.1 Setting Standards

It would be wrong to think that teachers have not always worked with standards. Ever since teachers have been marking work, they have always marked according to standards; it is just that the standards have been internalised by the teachers.

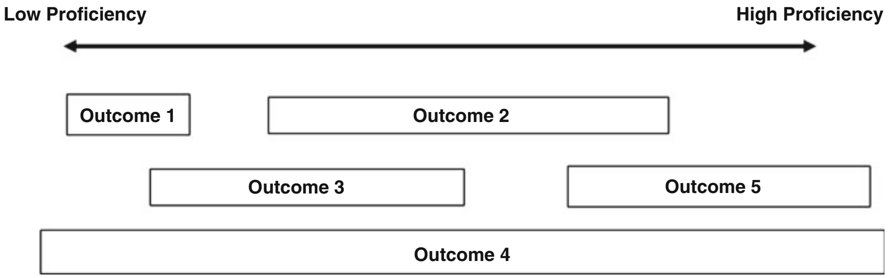


Fig. 2.2 Schematic representation of a developmental continuum

This means that comparability from one teacher to the next is practically non-existent, and the marks of students vary from school to school and teacher to teacher according to the severity of the teacher as a marker.

This is not fair. The same work should be given equivalent marks irrespective of which teacher does the marking. In order to do this, teachers have to make explicit what it is that they think students know and can do at different levels along a developmental continuum.

Before looking at how standards are set (and validated), it is worthwhile exploring the notion of a developmental continuum with regard to assessment, a term which has been used a number of times in this chapter.

One of the main ideas that have emerged in relatively recent times is the notion of developmental assessment. This is the process of monitoring a student's progress in a subject so that decisions can be made about how to improve learning for the student. Developmental assessment shifts the focus of attention in assessment from comparing one individual to another, towards one of monitoring student progress. The key feature of developmental assessment is that the students' progress or growth in the subject is monitored along a linear continuum that is referred to as a *developmental continuum* (see Fig. 2.2).

The monitoring of student growth along a developmental continuum requires that the continuum be defined. Many countries have now defined continua for the various subjects in terms of learning outcomes. These outcomes typically describe what students know and can do at different stages along the continuum. These outcomes are usually contained in syllabus documents or frameworks and provide the basis for the development of the teaching and learning sequence and activity (including assessment) within the subject.

It can be seen from Fig. 2.2 that some of the learning outcomes extend across the whole continuum (e.g. reading for meaning), whereas others are relatively less extensive. The further the outcome extends across the continuum, the more demanding it is for the students and the more of knowledge, skill and understanding of the subject is required to demonstrate achievement of the outcome.

To progress along the continuum, students have to become more proficient in the subject. Similarly, learning outcomes that are further along the continuum are more demanding for the student. They require more of the 'property', 'trait' or 'thing'

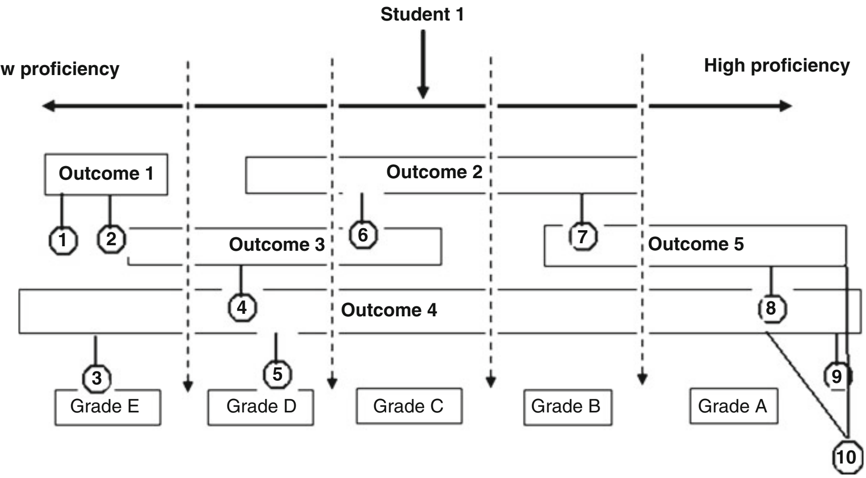


Fig. 2.3 Developmental continuum with items, students and grades

that defines the subject to be able to demonstrate proficiency. The whole idea is based upon growth.

Generally, the developmental continua are partitioned into levels, stages, bands or grades (see Fig. 2.3).

The grades have descriptors (grade-related descriptors) that try to capture the skills, understanding and knowledge that students have at different stages along the developmental continuum for the subject. These represent broad descriptions of standards, and teachers in schools and examiners are able to locate students along these continua by comparing their 'images' of students to these broad standards and using their professional judgement to say, on balance, that the student is located at 'grade D' or 'grade A' at this stage of their learning. Just as importantly, students can also locate themselves along this continuum by judging their own performance and work out what they have to do to go from a lower grade to a higher grade along the continuum. The continuum is cumulative in that what is required for a grade C is everything that is required for a grade D and grade E plus the extra for a grade C. Similarly, a grade A includes everything that is required for all grades up to A, plus the extra segment unique to grade A.

In order for students to demonstrate where they are along the continuum, they must be given the opportunity to demonstrate what they know and can do in relation to the outcomes of the subject. Tasks or items that examiners and teachers set provide this opportunity for the students to demonstrate what they are capable of doing.

In the case of formal assessments like public examinations and standardised tests, the examiners or test constructors must write items to match the student learning outcomes that are in the syllabus documents so that the results can be interpreted in terms of the same developmental continuum that is being used by the teacher in the classroom. In this way, the results should be providing one more piece of information about the location of the student and should supplement the teaching/learning process that is going on in the classroom.

Figure 2.3 shows items and a student along the continuum.

Item numbers in Fig. 2.3 are denoted by the numbers in the circles. It can be seen therefore that item 1 in the test is assessing outcome 1 and is relatively easy (because it is located towards the left on the continuum); item 2 is further to the right of item 1 on the continuum so it is demanding more of the student, and hence it is harder than item 1 and still measures outcome 1. Item 3 is a bit harder than item 1, not as hard as item 2 and is measuring outcome 4. Item 10 is a bit harder (hence it is more demanding of the students) than all the other items and actually measures two outcomes: outcomes 4 and 5.

Similarly, it can be seen from Fig. 2.3 that student 1 is located within grade C on this particular continuum. Because this student is located at that point along the continuum, it could be expected that the student would get the items that are less demanding (easier – i.e. they are located to the left of the student) correct and the more demanding items (harder – i.e. they are located to the right of the student) incorrect. Of course, students do not always behave in such an orderly fashion. They will probably get some easier items incorrect and some of the harder items correct. This is useful diagnostic information for both the student and the teacher.

In the classroom situation, teachers can make an 'on-balanced judgement' about the location of the student on the continuum.

In the case of the formal assessments (assessment of learning), the number of marks that the student gets on the examination locates the student along the continuum: the more marks students get on the assessment, the further they are located along the continuum.

A common approach used to establish performance standards (grades in the case just discussed) related to different levels of performance in a subject is to ask a group of experienced teachers and other subject specialists to describe, in summary from the knowledge, skills and understanding typically demonstrated by students, who will achieve each standard. In some cases, as a starting point to assist the group, these different levels might simply be given labels such as 'outstanding achievement', 'high achievement', 'satisfactory achievement' and the like. In other cases, an initial step may include giving the writers some guidance by providing brief general descriptions.

Whatever approach is used, the group preparing the descriptions has the task of summarising the level of knowledge, skills and understanding typically demonstrated by students who will achieve each performance standard. The developmental continuum can be also used to help develop these descriptions.

Before the introduction of performance standards for the NSW Higher School Certificate in 2001, teams of subject specialists prepared statements that summarised the nature and extent of the knowledge, skills and understanding typically demonstrated by students at six different levels in each course. These statements are associated with 'performance bands'. Band 6 represents the highest performance standard. No description was written for band 1, the level referred to as 'below the minimum standard expected'.

In some subjects, the results from past examinations can be used to assist the subject specialists in describing these different levels of achievement for a course. As an example, Table 2.1 shows the band descriptions produced for the HSC biology in New South Wales.

In the UK, New South Wales and Queensland standards-referenced assessment systems described above, results are generally reported at the level of a subject.

Student performance on the HSC, for example, is referenced directly to the performance band. Thus, when a student has achieved a score that locates him or her within band 2 on the performance band (see Table 2.1 for biology example), he or she can be reported as a student who typically demonstrates 'that they can recall limited knowledge and have an elementary understanding of some straightforward biological concepts; demonstrate a limited understanding of the historical development of biological concepts; use fundamental written communication relating to biology; provide simple qualitative descriptions of biological phenomena; and, use general biological terms occasionally'.

There are numerous models and procedures for moving from performance on an examination or assessment task to a standard. The next section shows how one of these methods is used to reference student performance to grades or, in this particular case, bands.

2.3.2 Using Performance Standard to Summarise Student Performance

In formal assessment situations like examinations, there are some common standard-setting methods that are typically used. These procedures focus on establishing the examination mark (cut-off) of the students who are on the borderline between grades, bands and levels. Once the cut-off marks have been established, anyone who scores at or above the cut-off mark is awarded the appropriate grade.

Two common methods that are used in systems around the world include the Angoff and the Bookmark Standard Setting methods. Both are well documented in the literature. The Angoff method is briefly described here to give an indication of how these methods work to ascribe a grade corresponding to a level of performance to a student in a formal (assessment of learning) assessment situation.

Both methods use the 'image' that has been referred to earlier as the key piece of information required for the exercise. Both rely on the professional judgement of judges who are usually teachers in an educational context.

In the case of the Angoff method, a number of (usually 6–10) teachers (who are referred to as judges) are selected for each subject. This means that within the team there is a good understanding of the range of achievement of students in the subject across different types of schools and geographic locations. At the same time, the team is small enough to enable each member of the team to contribute fully to the discussions throughout the process.

Prior to applying the procedure, the judges use the performance standards and work samples that are indicative of performance of students at the borderline between grades or bands to develop a personal ‘image’ of students at that borderline. Each judge then refines these images.

If, for example, judges are working in the New South Wales system which has the 6 performance bands, the judges would first focus on one borderline (e.g. band 5/band 6, band 4/band 5) at a time. Each judge, working independently from the rest of the team, then makes a decision as to what score *on each item* (*or test question*) in the examination students at that borderline would achieve. In the case of multiple-choice items, the judges would record the probability that a student at that borderline would answer the item correctly. In the case of extended response type items that are polytomously scored, the judges would indicate mark that they think the particular borderline student would achieve on the question.

Once a judge has recorded a decision for each item (question) and each borderline, the scores for each borderline are added. This gives that judge’s first estimate of the cut-off mark for each borderline. By averaging the cut-off scores for a borderline proposed by each judge, across judges, a first estimate of the team’s cut-off score for that borderline is obtained. Effectively, the judges are calculating the mark that the borderline student would get on the examination.

At the end of the first stage, each of the judges has their own cut-off marks. In the second stage of the procedure, the judges are brought together to discuss as a team the decisions they have made individually. The average mark across the team for each item and for the total examination is provided. To further assist in the discussion, the team is given statistical data showing the scores achieved by the students on each item and relating it to their total score on the examination.

In the third stage of the procedure, the judges are generally given a report showing the decisions each member of the team has made during the second stage and the resulting cut-off marks being proposed by the team. The judges are then given the examination scripts of some students who have achieved marks equal to the team’s proposed cut-off marks. The purpose of this is to enable the judges to satisfy themselves that the standard of knowledge and skills exhibited by these students in the examination is consistent with their expectations of students whose performances would place them at the borderlines between the performance standards that have been established for the subject. The judges review these scripts individually and then discuss their views with their team members. The judges take a holistic view of the scripts during this part of the process.

If the judges have any doubt as to whether the students’ performances as reflected in these scripts are not truly ‘borderline’, they are given other scripts to review. Judges are also able to have access to further scripts (if required) that are awarded the proposed cut-off marks, or ones that receive a slightly higher or lower mark. During this stage, judges have the opportunity to vary their cut-off marks for one or more items. This step is consistent with the practice advocated by Mills et al. (1991) and Berk (1996) that has been shown to produce reliable standards.

After any final adjustments are made, and the final values recorded by the judges are totalled and averaged as before, the procedure has produced the cut-off mark for each performance band.

It can be seen that the Angoff method used for formal assessment activities like examinations is premised on the professional judgement of judges, and these judges are generally teachers.

The procedure for establishing the cut-off scores has a number of advantages. Firstly, it has a strong theoretical base. The use of expert judgement to set standards in a systematic and professional manner is well tried and documented. The procedure outlined above builds upon a well-regarded standard-setting methodology.

Secondly, the judges (who in this case are the teachers) are very involved in the process. The professional development involved in a process that requires them to post hoc assess the skill that is being tested by each of the questions in the paper, internalise the image of a student on the border of each of the performance bands and assign a score for that student on the question, discuss and defend the score that was given on each question with other teachers and colleagues involved in the process, internalise statistical information regarding the performance of the group on each of the items and then use all of this information to arrive at a cut score will undoubtedly ensure that teachers have a significant opportunity to improve their teaching and assessment and at the same time improve their knowledge of the curriculum and the performance standards.

The same procedure could be used in the school situation.

Another procedure that can be used by teachers in schools is to aggregate the marks from all the different tests, term examinations, assignments, etc., during the year and order the students in the class or school for a subject on the basis of their total marks (added across all these assessments.), and then the teacher (or teachers) should come down the list until the image that they have for the borderline band 5/6 student corresponds with the student in the school list. This student is thus the band 5/6 borderline student, and the mark that the student obtained is the cut-off mark. All students with scores above this student are in band 6. The process can be repeated to obtain all the other cut-off marks for all the other bands.

The next section examines how marks that have been referenced to performance bands can be reported.

2.3.3 Reporting Student Performance

As can be seen from the previous section, student performance is referenced directly to the performance band. Thus, when a student has achieved a score that locates him or her within band 2 on the performance band (see Table 2.1 for biology example), he or she can be reported as a student who typically demonstrates ‘that they can recall limited knowledge and have an elementary understanding of some straightforward biological concepts; demonstrate a limited understanding of the historical development of biological concepts; use fundamental written communication relating to biology; provide simple qualitative descriptions of biological phenomena; and, use general biological terms occasionally'.

In addition, in a situation like the example described in New South Wales, it is also possible to retain the marks (scaled to accommodate the marking schema being used by the system) and even show the distribution of marks on the reporting scale (developmental continuum) as well (see Fig. 2.4).

2.3.4 Some Suggestions for Teachers and Examiners in Setting Examinations and Tests in a Standards-Referenced System

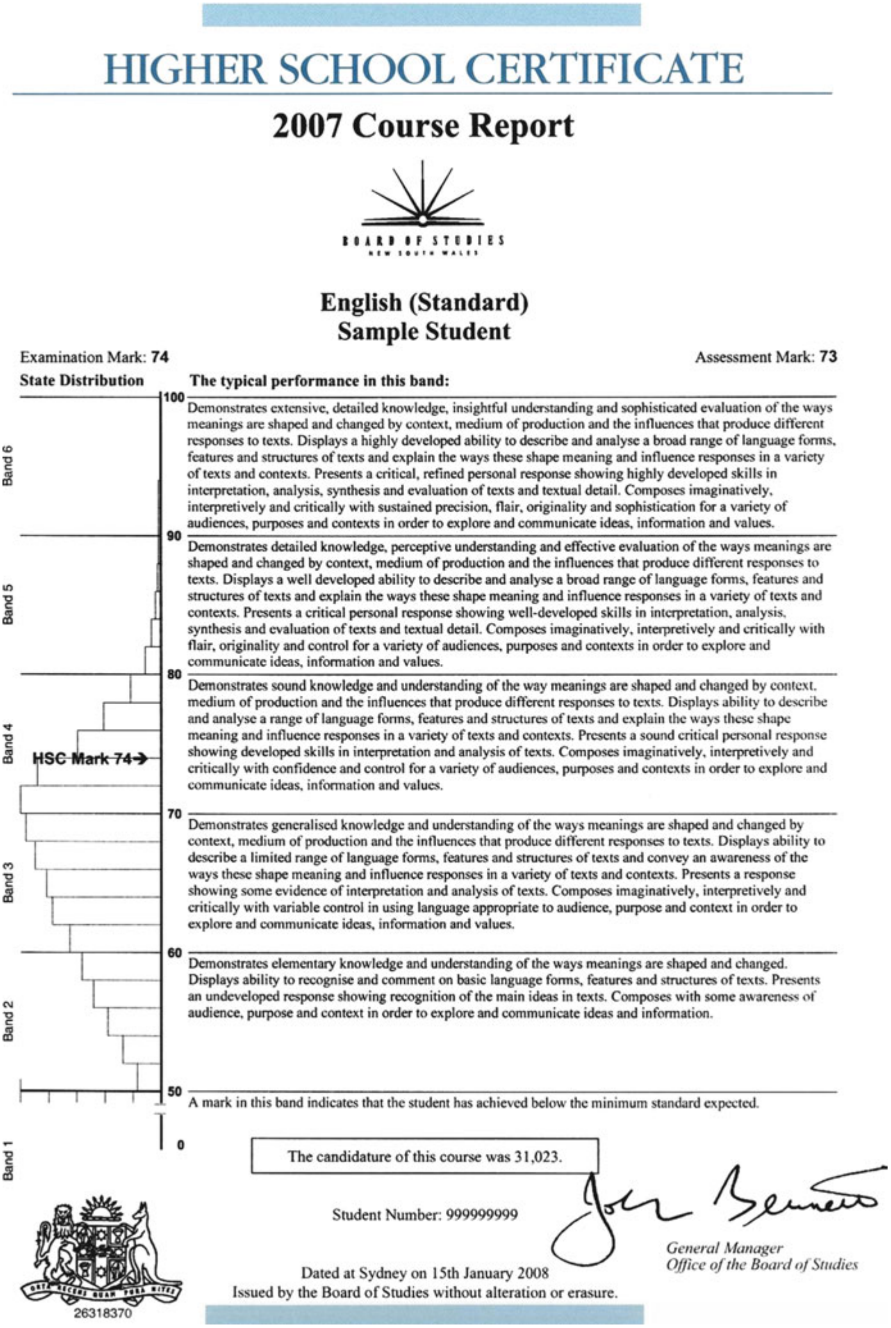
The main feature of standards-referenced assessment is that it focuses on the student and the progress of the student along the developmental continuum. Teachers and examiners must develop a quality image and be able to reliably compare the image of the child to the performance standard. As such, it has a number of requirements that are critical to enhancing the teaching and learning that is taking place in schools.

For example, if teachers have to monitor students' progress against outcomes and be fair and consistent in making decisions about where students are located along the continuum, then it means that they should ensure that the items and questions that they write and the assignments that they set actually match the content standards (outcomes) articulated in the syllabus.

While this may appear to be an obvious thing to do, it is not widely done, and as a result, the image that teachers form of student performance and is so critical to the process of standards-referencing is flawed. This is not fair to the students.

When constructing classroom tests (and examinations), it is important to ensure that the items and questions that are developed actually enable students at different stages in their learning (locations along the developmental continuum) to demonstrate what they actually know and can do. There should be items and questions that enable the students who are just beginning the outcome to demonstrate that they are at this stage; those students who are a fair way along the outcome should be able to provide evidence of their location along the outcome, while those who are well advanced will also be able to demonstrate that they are capable of achieving the more demanding learning outcomes associated with the subject. In other words, teachers and examiners should ensure that as the items are being written, the ones that are intended to be located further towards the top of the scale are, in fact, harder than those that are located towards the bottom of the scale and ensure that the reason that the items are harder is a function of the property/variable that is being measured and not a function of some other extraneous feature (validity).

Given the focus on the individual and what the individual knows and can do, it is important to construct tests and examinations which minimise the influence of factors not directly associated with the learning outcomes being assessed. A simple hint would be to construct tests so that wherever possible the items in the tests go



HIGHER SCHOOL CERTIFICATE

2007 Course Report

BOARD OF STUDIES
NEW SOUTH WALES

English (Standard)
Sample Student

Examination Mark: **74**

Assessment Mark: **73**

State Distribution

The typical performance in this band:

100
Band 6
90
Band 5
80
Band 4
HSC Mark 74→
70
Band 3
60
Band 2
50
0
Band 1

Demonstrates extensive, detailed knowledge, insightful understanding and sophisticated evaluation of the ways meanings are shaped and changed by context, medium of production and the influences that produce different responses to texts. Displays a highly developed ability to describe and analyse a broad range of language forms, features and structures of texts and explain the ways these shape meaning and influence responses in a variety of texts and contexts. Presents a critical, refined personal response showing highly developed skills in interpretation, analysis, synthesis and evaluation of texts and textual detail. Composes imaginatively, interpretively and critically with sustained precision, flair, originality and sophistication for a variety of audiences, purposes and contexts in order to explore and communicate ideas, information and values.

Demonstrates detailed knowledge, perceptive understanding and effective evaluation of the ways meanings are shaped and changed by context, medium of production and the influences that produce different responses to texts. Displays a well developed ability to describe and analyse a broad range of language forms, features and structures of texts and explain the ways these shape meaning and influence responses in a variety of texts and contexts. Presents a critical personal response showing well-developed skills in interpretation, analysis, synthesis and evaluation of texts and textual detail. Composes imaginatively, interpretively and critically with flair, originality and control for a variety of audiences, purposes and contexts in order to explore and communicate ideas, information and values.

Demonstrates sound knowledge and understanding of the way meanings are shaped and changed by context, medium of production and the influences that produce different responses to texts. Displays ability to describe and analyse a range of language forms, features and structures of texts and explain the ways these shape meaning and influence responses in a variety of texts and contexts. Presents a sound critical personal response showing developed skills in interpretation and analysis of texts. Composes imaginatively, interpretively and critically with confidence and control for a variety of audiences, purposes and contexts in order to explore and communicate ideas, information and values.

Demonstrates generalised knowledge and understanding of the ways meanings are shaped and changed by context, medium of production and the influences that produce different responses to texts. Displays ability to describe a limited range of language forms, features and structures of texts and convey an awareness of the ways these shape meaning and influence responses in a variety of texts and contexts. Presents a response showing some evidence of interpretation and analysis of texts. Composes imaginatively, interpretively and critically with variable control in using language appropriate to audience, purpose and context in order to explore and communicate ideas, information and values.

Demonstrates elementary knowledge and understanding of the ways meanings are shaped and changed. Displays ability to recognise and comment on basic language forms, features and structures of texts. Presents an undeveloped response showing recognition of the main ideas in texts. Composes with some awareness of audience, purpose and context in order to explore and communicate ideas and information.

A mark in this band indicates that the student has achieved below the minimum standard expected.

The candidature of this course was 31,023.

Student Number: 999999999

John Bennett
General Manager
Office of the Board of Studies

Dated at Sydney on 15th January 2008
Issued by the Board of Studies without alteration or erasure.

26318370

Fig. 2.4 Example of a standards-referenced report

from easy to hard. In this way, the students will not give up on the whole test because they have encountered an item that is much too difficult, and their performance on the test will not be adversely affected by a heightened anxiety level. The result will more closely resemble what is the standard of performance of the students.

A second hint that should improve the validity of the results from tests and examinations is to ensure that the items and questions in the tests and examinations are written within a context that engages the students. In this way, the students are more likely to demonstrate their actual achievement, and the image that is used to compare against standards is more likely to faithfully reflect the student performance.

Where possible, teachers should use a range of different tasks to generate a reliable and valid estimate of the student's location along the developmental continuum. This does not mean that there should be a large amount of formal assessment. Rather, teachers should be collecting information constantly and then confirming what they know about the students with a few formal, well-constructed tasks. As such, they should be well constructed and should meet the requirements of well-constructed standards. Any marks that are awarded using the marking guidelines should be consistent with the requirements of a developmental assessment model. That is, more marks implies evidence that the performance is further along the developmental continuum, and hence, in this model, marks have a meaning which is different from what they traditionally have in a norm-referenced model.

Wherever possible, whether the results are derived from an examination or some less formal assessment, students should be provided with feedback that is designed to help them improve and move along the developmental continuum. The very fact that performance standards articulate what it means to improve in a subject empowers the students in the teaching learning process. Marking rubrics, when published, also enable information from tests and examinations to be used for improving learning.

In order for systems to use standards-referencing effectively and ensure comparability of results across schools, districts and systems, there is a need for teachers to have a shared understanding of the meaning of the learning outcomes (through the use of exemplars, work samples, teacher development meetings, etc.) and that there is consistency of teacher judgement in making the on-balanced decision about where the student is located along the continuum.

2.3.5 Standards-Referencing for School Executives

One of the key challenges confronting school executives is to make sure that the educational community is supportive of systems predicated on standards. To ensure that support there must be a conscious effort to let the students, parents and community know the value of using standards to reference performance. It must be remembered that they have come through a different system. They want marks. They want formal tests. They want to know where their child is relative to the other students. Of course, it is possible to keep supplying such information. But it should be continually subservient to the message of what their child knows and can do at this time; this is where he or she has to get to next; and, this is how 'we' can help the child improve. It is a difficult challenge and raises the whole issue of reporting. There is no doubt that reporting has to be improved in a standards-referenced system, maybe with digitised reporting coupled with the developmental continuum. This is a separate topic for another occasion.

The emphasis on teacher judgement and teacher skill in developing assessment tasks (items, questions, etc.) means that teachers need to be well trained in this aspect of their work. At the moment, little formal assessment training occurs in pre-service or in-service training courses. There has to be an acknowledgement that assessment is an important part of a teacher's repertoire of skills and an assurance that they are well versed in not only the assessment techniques but also the philosophy that provides the backdrop for assessment in schools.

2.4 Conclusion

This chapter outlines a model for giving meaning to achievement by referencing it to student learning or standards. This effectively shifts the focus in assessment from notions of rank ordering students (comparing their performance purely to each other) to those of monitoring growth or progress and measurement. More specifically, it introduces standards-referenced assessment: the concept and theory, and it provides some indication of how standards can be implemented at a system, school and individual student level.

There is no doubt that the introduction of standards-referenced systems will force systems to work hard on integrating assessment, teaching and learning. There is also no doubt that such a system will ensure that the children, who are the responsibility of educators and education systems, will move along their life journey at the speed that best suits them. Paradoxically, the transparency of student progress which emerges in a standards-referenced reporting system appears to reverse 'dumbing down' which can occur when developmental progression is not emphasised (Stanley and MacCann 2005).

References

American Federation of Teachers (AFT) (1999). Making standards Matto, American Federation of Teachers, Washington. http:// www.aft.org/edissues/standards99/judging.htm

Berk, R. (1996). Standard setting: The next generation. *Applied Measurement in Education, 9*, 215–235.

Donnelly, K. (2000). *New Zealand's National Certificate of Educational Achievement (NCEA): An international perspective*. Wellington: Daphne Brasell Associates.

Elley, W. (2004, October). *Facts and fallacies about standards-based assessment*. A paper presented at the Cambridge International Conference.

Manno, B. (1994, June). *Outcomes-based education: Miracle cure or plague*? Hudson Institute briefing paper number 165.

Mills, C., Melican, G., & Ahluwalia, N. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice, 10*, 7–10.

New South Wales Board of Studies. (1999, May). Assessment and reporting in the new higher school certificate. *Newsletter*, *14*. http://www.boardofstudies.nsw.edu.au/archives/stfreview/stf_14.html#Heading3

Peddie, R. (1992). *Beyond the norm*. Wellington: New Zealand Qualifications Authority.

Popham, W. J. (1978). *Criterion-referenced assessment*. Upper Saddle River: Prentice Hall.

Shanker, A. (1993). *Outrageous outcomes*. American Federation of Teachers. http://www.aft.org/stand/previous/1993/091293.html

Shanker, A. (1994). *A do-it-yourself kit*. American Federation of Teachers. http://www.aft.org/stand/previous/1994/100994.html

Stanley, G., & MacCann, R. G. (2005). Removing incentives for "dumbing down" through curriculum re-structure and additional study time. *Educational Policy Analysis Archives, 13*(2). Retrieved February 28, 2005, from http://epaa.asu.edu/epaa/v13n2/

Chapter 3
Rapid Dynamic Assessment for Learning

Slava Kalyuga

3.1 Introduction

Enhancing formative diagnostic assessment is a clear current trend in educational testing. Such assessment allows determining specific levels of acquisition of knowledge and skills and provides fine-grained diagnostic information about strengths and weaknesses of a particular learner. Teachers are encouraged to use more formative assessments throughout their courses to inform their classroom instruction.

In a major report on educational assessment, Pellegrino et al. (2001) emphasized that cognitive theories should be the cornerstone of the assessment design process directed toward evaluating students' schematic knowledge structures. Cognitive models of specific domains are usually based on task analyses, expert interviews, and verbal protocols of thinking processes and identify cognitive attributes required for successful learning and performance in these domains. The need for using cognitive theories of learning and models of expertise as foundations for the design of assessment has been recognized by many educational testing theorists (Embretson 1993; Mislevy 1996; Pellegrino et al. 1999; Pirolli and Wilson 1998; Snow and Lohman 1989; Tatsuoka 1990).

Cognitive diagnostic assessment is aimed at providing ongoing information about students' mastery of specific cognitive processes and operations required for learning and performing particular types of tasks. It combines cognitive models of corresponding domains and statistical models of students' response patterns. Empirical evidence shows that cognitive diagnostic assessment is capable of maximizing students' learning outcomes (e.g., Russell et al. 2009).

However, testing learners continuously without interfering with their learning is a challenging task. Testing time could not be increased considerably as it would

S. Kalyuga (✉)
School of Education, University of New South Wales, Sydney, NSW, Australia
e-mail: s.kalyuga@unsw.edu.au

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_3, © Springer Science+Business Media Dordrecht 2013

inevitably reduce instruction time. Traditional standardized multiple-choice tests are rather time consuming and not always represent the best way of diagnosing learner actual levels of knowledge in a domain. With most currently used diagnostic assessment techniques, developing and administering the tests, obtaining data, and interpreting results, as well as incorporating appropriate instructional interventions based on these results, may require considerable amount of time. As a consequence, many teachers may not be inclined to use cognitive diagnostic assessment to guide their instructional decisions.

A possible solution for this problem is to make diagnostic assessment rapid in order to accelerate the process (rapid diagnostic assessment). Another possibility is to use diagnostic assessment itself as an instructional means by integrating seamlessly testing and learning. With this approach, students learn while being tested or are assessed while learning (dynamic assessment). This chapter starts with the description of the general idea of a rapid diagnostic assessment approach and its theoretical framework based on cognitive nature of expertise, schema-based assessment, and cognitive load theory. Then, it describes a general design approach and its specific implementations as rapid diagnostic methods, as well as their possible integration with dynamic assessment methods (rapid dynamic assessment). A summary and directions for further research and development in this area conclude the chapter.

3.2 Theoretical Framework

3.2.1 Knowledge Base and the Nature of Expertise

Whether expertise is considered in a real professional sense (e.g., Ericsson and Charness 1994) or at a narrow task-specific level (e.g., secondary school students as experts in solving linear algebra equations), it includes a well-organized domain-specific knowledge base as its most important component (Bransford et al. 2000). This knowledge resides in long-term memory which represents one of the major components of human cognitive architecture that underlies cognition and learning. Another essential component of this architecture is working memory.

According to a contemporary model of human cognitive architecture (Sweller 2004; Sweller et al. 1998; Van Merriënboer and Sweller 2005), working memory represents our immediate conscious processor of information. It is limited in both duration and capacity when dealing with novel elements of information (Baddeley 1997; Miller 1956; Peterson and Peterson 1959). No more than a few elements of information could be processed and maintained consciously at the same time in working memory, and they would most likely be lost after a few seconds (unless intentionally rehearsed). For a simple example, consider dialing an unfamiliar mobile phone number after just having heard it from another person.

In familiar domains, the available knowledge base in long-term memory allows us to chunk many elements of information in larger units that could be treated as single elements in working memory. For example, it would be easier to dial the above phone number if you notice a well-known combination of digits as a part of it (e.g., “2010” that could be treated as a single unit of information instead of four). Therefore, the long-term memory knowledge base effectively influences the actual content and capacity of working memory and determines the efficiency of performance.

Studies of expert-novice differences in cognitive science have convincingly demonstrated that learner knowledge base is a single most important cognitive characteristic that influences learning and performance (e.g., Chi et al. 1981; Larkin et al. 1980; see Pellegrino et al. 2001, for a review). When experts face a problem in a familiar area, their available knowledge structures are rapidly activated and brought into working memory as problem-relevant chunks of information. Ericsson and Kintsch (1995) called such knowledge structures associated with currently active working memory elements as long-term working memory (LTWM) structure. They are capable of holding virtually unlimited amount of information due to the chunking effect. In the absence of appropriate domain-specific knowledge structures, novices have to resort to cognitively inefficient and time-consuming random search or weak problem-solving methods such as means-ends analysis or trial-and-error approach.

For example, in classical studies of chess expertise by De Groot (1965) and Chase and Simon (1973), professional grand masters performed considerably better than amateur players in reproducing briefly presented chess positions taken from real games, although there were no significant differences when random configurations of chess figures were used. Knowledge of effective moves for a large number of different real game patterns held in grand masters’ long-term memory allowed them to reproduce chess positions by large chunks of familiar patterns rather than by individual chess figures. During a short exposure to a real-game board configuration, they were able to form long-term working memory structures associated with presented configurations of chess figures using their available domain-specific knowledge base.

The organized generic knowledge structures that we use for categorizing information according to familiar patterns are called schemas. Since the levels of learner expertise in a specific domain are determined by the levels of acquisition of schematic knowledge structures in long-term memory, schemas should be the major target for diagnostic assessment of expertise. In cognitive science, laboratory studies using interviews, observations, and “think aloud” protocols are conducted for uncovering schemas held by individuals (Chi et al. 1989; Ericsson and Simon 1993; Magliano and Millis 2003). Although highly powerful and precise, these methods are very time consuming, slow, and not suitable for realistic educational settings. Combining high levels of diagnostic power with acceptable speed of assessment and simplicity of its implementation is a very challenging task. The next section describes an idea of a potentially suitable approach.

3.2.2 Rapid Schema-Based Assessment

Since long-term memory that contains schematic knowledge base cannot be accessed directly, we usually make inferences about learners' available knowledge structures based on the results of their problem-solving performance (e.g., answers to multiple-choice items or recorded problem solutions). However, such inferences may not be reliable because they are based on remote and indirect results of actual cognitive processes and structures. They could in fact be misleading for cognitive diagnosis.

For example, based on students' correct answers to multiple-choice items in solving algebra equations (e.g., $5x=-4$), it is not possible to say exactly what cognitive processes were involved. Some students could apply knowledge-based schematic solution procedures, but others could achieve the same outcomes by using a random search method. Even those students who relied on knowledge structures could use different levels of knowledge. Some students could apply fine-grained step-by-step procedures (dividing both sides of the equation by 5, $5x/5=–4/5$, then canceling the same numbers in the numerator and denominator on the left side of the equation), while others could use higher-level automated procedures by skipping intermediate steps with the final answer ($x=–4/5$) obtained immediately. Traditional multiple-choice tests would place these two groups of students who are at correspondingly intermediate and top levels of expertise in this task area, together with novices using weak problem-solving methods, in the same category of successful learners (Kalyuga 2006b, d).

A similar situation could be with traditional methods used for assessing reading skills that do not measure students' actual cognitive representations constructed during reading (Magliano and Millis 2003). Students are usually required to read segments of text and then answer multiple-choice questions related to the concurrently displayed texts. Correct answers to such multiple-choice questions would not indicate what actual cognitive processes were used before selecting those answers. Students who achieved correct answers by repeatedly searching the text for key question words (novice readers) and those who answered correctly by relying on their constructed coherent mental representations of the text (advanced readers) would not be distinguished.

Thus, obtaining evidence that is directly related to the assessed schemas is essential for ensuring the diagnostic validity of assessment tools. A possible approach could be based on directly observing what schemas (if any) learners use immediately when approaching a problem or trying to make sense out of the presented situation. Even though schema-based approaches to the assessment of students and to the design of test items have been suggested before (Marshall 1993, 1995b; Singley and Bennett 2002), the idea of registering rapidly if and how learners use their schemas while they approach a specific problem or situation has a potential value for enhancing cognitive diagnostic assessment (Kalyuga 2006d; Kalyuga and Sweller 2004). A general design methodology and specific implementations of this approach will be described in the following sections.

3.2.3 General Design Framework

The rapid schema-based diagnostic approach is based on observing task-relevant schemas from long-term memory (if any) that are rapidly activated and brought into working memory as learners approach a briefly presented specific task situation. Individuals who are more experienced in the task domain would be better able to recognize presented problem states and retrieve appropriate solution schema steps than less knowledgeable learners. Experts could immediately see a task situation within their higher-level knowledge structures and activate appropriate solution schemas, while novices could only locate some random lower-level components.

The design of a schema-based assessment may follow a general conceptual framework for the design of cognitive assessment containing three basic components: the student model, the task model, and the evidence model (Mislevy et al. 2002). The student model (or model of expertise) describes the cognitive constructs to be assessed, i.e., schemas that guide cognitive processing in a specific task area. The task model defines characteristics and patterns of tasks that would allow obtaining evidence about assessed cognitive knowledge structures. The evidence model defines observable variables, their scoring procedures, and a specific measurement model to be applied to the data.

According to this framework, the task-relevant schemas should be described first, followed by a pattern of tasks that would elicit evidence about these schemas, and finally by a scoring procedure for these tasks and a suitable measurement model to make statistical inferences about levels of acquisition of the assessed schemas. The following section describes possible implementations of the above general idea and examples of applying the rapid schema-based assessment to coordinate geometry tasks and arithmetic word problems. These two task areas differ in types of knowledge and levels of knowledge organization.

3.3 Rapid Diagnostic Assessment Methods

The idea of rapid schema-based assessment can be realized either as a first-step method or a rapid verification method. In the first method, learners are presented with a task for a limited time and asked to rapidly indicate their first step toward solution. Different first steps would indicate different levels of expertise. This method was validated in a series of studies using tasks in algebra, coordinate geometry, and arithmetic word problems. Results showed high levels of correlations between performance on the rapid tasks and detailed traditional measures of knowledge (Kalyuga 2006d, 2008; Kalyuga and Sweller 2004), with substantially reduced test times.

The rapid verification method is a version of the first-step procedure designed for the use in computer-based environments. Learners are actually presented with a series of possible steps (some of which are incorrect) at various stages of the solution

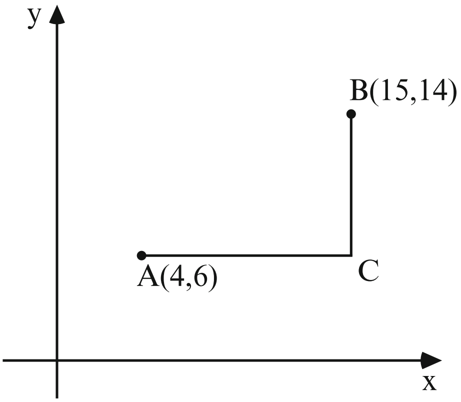


Fig. 3.1 A diagram for the basic task used in the coordinate geometry area (Adapted from Kalyuga and Sweller (2004). Copyright © 2004 by the American Psychological Association, Inc.)

procedure and asked to rapidly verify the correctness of these steps. Knowledge structures of more experienced learners would presumably allow them to verify suggested steps more successfully than novices. This method was validated using sentence comprehension tasks and tasks in kinematics (Kalyuga 2006a, c, 2008). Again, significant correlations were found between performance on the rapid verification tasks and extended traditional measures of expertise, with significantly reduced test times.

Since either of the above two forms of rapid assessment approach could be used with the same student model (model of expertise), task model, and evidence model, the following examples are concentrated on developing and implementing these components of the general design framework using areas of coordinate geometry and arithmetic word problems. According to this framework, a subgoal structure of the tasks and a sequence of corresponding solution steps should be established first. Then, for each step, representative subtasks could be designed and arranged in an appropriate ordered series to be presented to learners, each task for a limited time. The scoring procedure should distinguish student responses corresponding to different levels of expertise in the domain. To assess the level of acquisition of each schema, an appropriate measurement model should be fitted to the data.

3.3.1 Rapid Assessment of Expertise in Coordinate Geometry

3.3.1.1 Model of Expertise

A narrow task area selected for demonstrating the method could be described by the basic top-level task (Fig. 3.1) that includes a coordinate plane and two points A and B with given coordinates. Lines AC and BC are parallel to the *x*- and *y*-axes respectively. The task is to find the lengths of AC and BC. This task effectively requires finding the distance between projections of two points on a coordinate axis and using the knowledge that opposite sides of a rectangle have equal lengths. The schemas required for solving this task include:

- The schema for determining coordinates of a point as coordinates of projections of the point on *x*- and *y*-axes
- The schema for establishing equal opposite sides in a rectangle
- The schema for calculating the distance between two points on a coordinate axis (a number line) by subtracting the smaller coordinate value from the larger one (or left hand coordinate from the right hand coordinate, for *x*-axis; and lower coordinate from the upper coordinate, for *y*-axis)

Different levels of acquisition of these three schemas define the student model in this task area (Kalyuga 2006b). Solving the basic task requires the sequential applications of these schemas to corresponding subtasks. However, a learner who has some schemas at higher levels of acquisition (e.g., automated) could skip some intermediate stages of the solution that would be effectively encapsulated into a higher-level schema. For example, a student who has sufficient prior experience in finding coordinates of a point may find out the coordinates of the points immediately upon presentation of the task without drawing projection lines explicitly. Expert students with extensive experience in this area may immediately (as their first step) write a numerical expression for the length of AC as the difference between *x*-coordinates of points B and A.

3.3.1.2 Task Model

A pattern of tasks for a rapid assessment of expertise in this task area could have a hierarchical structure with three types of tasks in the pattern: a top-level basic task (requires schemas *a*, *b*, and *c*), a task corresponding to the second step in the solution of a basic-level task (requires schemas *b* and *c*), and a task corresponding to the final step in the solution of a basic-level task (requires schema *c*). To solve a basic-level task, a learner should acquire schemas necessary for solving each of these three tasks. Lack of any schema would interfere with the entire solution procedure.

Because completing a first step for each task leads directly to one of the following task levels, and each of these levels is represented by another task in the series, the first-step assessment method (or an equivalent rapid verification approach) would diagnose the entire set of schemas in this task area. Accordingly, the tasks should be sequenced according to the number of schemas that are required to solve each of them. For the top-level basic task, no additional details are provided on the diagram. For each of the lower-level tasks, progressively more additional details of partial solutions (e.g., indications of projecting lines and coordinates of the points on axes) are provided on the diagram. For instance, the third task in the series should present most of the details and require only calculation of the differences between the indicated coordinates of two points on each axis.

In task areas like coordinate geometry that use diagrammatic representations as essential components of tasks, each sequential step includes information presented at the previous stages of the solution process. The series of diagnostic tasks in the pattern is effectively a sequence of partially worked-out examples with gradually increasing levels of detail provided to learners. In other domains, it is possible to construct a different task pattern that is based on all possible relevant combinations of basic schemas (see an example for word problems in the next section).

3.3.1.3 Evidence Model

In a possible scoring procedure, for each step that requires application of a specific schema, two units are allocated for completing the step and one unit for an intermediate action (an unfinished solution step). If a procedure does not have an intermediate stage, one unit is allocated for completing the step. A zero score is allocated for a wrong answer and for not providing any answer. With a rapid verification method, the same scores are allocated for correct verifications of corresponding solution steps.

For example, for a lower-level task that requires applying only schema *c*, scores 2 and 1 are allocated respectively for providing or verifying a completed final answer (AC = 11; numbers correspond to Fig. 3.1 for illustrative purposes only, actual diagnostic tasks at different levels should vary in specific numerical parameters) and incomplete final answer (AC = 15 − 4).

In contrast, for a top-level task that requires application of all three schemas *a*, *b*, and *c*, scores 5 and 4 are allocated respectively for providing or verifying the above responses at the stages of application of the schema *c* corresponding to the final step and the step that immediately precedes it. A score 3 is allocated for providing or verifying an answer at the stage of application of the second schema *b* (indicating equal sides of a rectangle; there is no intermediate action for this schema). A score 2 is allocated for providing or verifying an answer at the stage of completed application of the first schema *a* (e.g., indicating projections and *x*-coordinates of points A and B). A score 1 is allocated for providing or verifying an intermediate (unfinished) step when applying the first schema (e.g., indicating only a projection line without the coordinate of a point). Thus, an additional score is allocated for each skipped intermediate step in the first-step response (or integrated into an advanced step in the rapid verification procedure)

The application of the first-step method in a paper-based format in a realistic class environment with 20 grade 9 students (Kalyuga and Sweller 2004) indicated a high level of correlation of 0.85 between learners' performance on the rapid test and traditional measures of knowledge of corresponding procedures and concepts, with the test time reduced by a factor of 2.5. The following instructions were provided to students:

> In each of the figures, A and B are two points on a coordinate plane. Lines AC and BC are parallel to the coordinate axes. Assume you need to find the lengths of AC and BC.

Some additional details (lines, coordinates) or partial solutions are provided on most figures. For each figure, spend no more than a few seconds to indicate your first step toward solution of the task.

Remember, you do not have to solve the whole task. All you have to do for each figure is to show only your first step toward the solution (e.g., it might be just writing a number or drawing a line on the diagram). If you do not know your answer, proceed to the next page.

Do not spend more than a few seconds for each figure, and do not go back to pages you have already inspected.

3.3.2 Rapid Assessment of Expertise in Solving Arithmetic Word Problems

3.3.2.1 Model of Expertise

The described model of expertise uses the analysis of schemas in this task area conducted by Marshall (1993, 1995a) that suggested five types of basic schemas. In order to simplify the illustration of the diagnostic method, four of these schemas are used: Change, Group, Vary, and Restate schemas (Kalyuga 2006d).

The Change schema (denoted as *C*-schema for convenience) applies to a situation in which there is a change over time in the value of a variable, for example, *After 5 students had left the class, 12 students remained. How many students were in the class initially?* Students who indicate as their first solution steps (or verify as correct steps) expressions like $X-5=12$, $5+12$, $12+5=17$, or *17* demonstrate evidence of the Change schema. Different first steps correspond to different levels of the schema acquisition. For example, experienced students may recognize a familiar situation right away and write (or verify) the final answer (*17*) immediately due to their automated schema and do not require much conscious processing in applying this schema.

The Group schema (*G*-schema) relates to situations in which a number of components are combined into a larger unit, for example, *John's homework contains 16 tasks. John completed 11 tasks in the afternoon. In the evening, he did the remaining tasks. How many tasks did John complete in the evening?* Students who write as their first steps or verify expressions $16=11+X$, $16-11$, $16-11=5$, or *5* demonstrate evidence of the Group schema (on different levels of acquisition).

The Vary schema (*V*-schema) relates to situations in which a systematic relationship exists between two variables: IF the amount of one variable decreases or increases, THEN the amount of the second variable changes in a certain way (*IF-THEN* relationship). The task *A train traveled 120 km in an hour. If the train continued to travel at the same speed, then how far would it travel in 4 h?* requires applying the Vary schema as it could be redescribed as *IF a train traveled 120 km in 1 h, THEN it will travel unknown amount of kilometers in 4 h*. Students who write as their first solution steps or verify statements like $1 * 4 \rightarrow 120 * 4$, $120 * 4 = 480$, or *480* demonstrate evidence of the Vary schema.

The Restate schema (*R*-schema) applies to situations where there is a known relationship between two variables (ratio-like situations such as *twice as*, *two more than*, etc.) and a restatement of this relationship using different values from those involved in the initial statement, for example, *Water is mixed with cement in the proportion 2 : 1?. How many units of water are required for 5 units of cement?* Students who write as their first solution steps or verify statements like $2:1=X:5$, $5 * 2$, or *10* would demonstrate evidence of the Restate schema.

As previously, the degree of schema acquisition is defined by the level of granularity of solution steps and the number of skipped steps. The levels of acquisition may range from a consciously controlled, slow, and articulated application of all possible fine-grained solution steps (a novice level) to a fluent automated performance with final answers obtained immediately after reading problem statements (an expert level).

The described schema-based model of student expertise is an attempt to impose a schematic structure on a relatively poorly structured task domain using a number of simplifying assumptions. For example, it is assumed that students have sufficient reading comprehension skills that would not introduce an interfering factor. Another assumption is that if a student starts solving a task by drawing a graphical representation, it could be possible to relate unambiguously this diagrammatic representation with a corresponding numerical solution step.

3.3.2.2 Task Pattern

Each of the above four basic tasks would require applying only one corresponding schema. There are $4\times4=16$ different tasks based on all possible combinations of two schemas. In these combinations, the order of schema applications is important, and repeated applications of the same schema are also allowed.

For example, the task *Paul is thinking of a number. When he adds 6 to the number and then subtracts 9, he would get 15. What is the number John is thinking of?* requires two sequential applications of the *C*-schema (CC-task). Applying the first Change schema could result in such responses as $N-9=15$, $9+15$, $9+15=24$, or *24*. The second Change schema could be used by the students who have completed the first operation, producing the following possible responses: $N+6=24$, $24-6$, $24-6=18$, or *18*. Some students could also combine two schemas and write $(15+9)-6$, $15+9-6$, etc.

The task *There are 15 boys in a class. The number of girls is 8 more than the number of boys. How many students in the class?* represents an example of the CG-task. The Change schema could be applied first with possible responses $15+8$, $15+8=23$, or *23*. Then, the Group schema could be used with possible responses $15+23$, $15+8+15$, $(15+8)+15$, or *38*. A GC task situation is different from the CG-task because it would require applying the Group schema first followed by the application of the Change schema, for example, *Two plates on a table contained respectively 4 and 7 apples. A third plate with apples was added making a total of 18 apples on the table. How many apples were on the third plate?*

Thus, all possible task situations that are based on applying one or two schemas could be represented by a pattern consisting of 20 tasks. Using a similar combinatorial approach, it is also possible to construct three-schema tasks, four-schema tasks, and so on. However, for three-schema tasks, it is unlikely that even highly experienced students would be able to skip first two operational steps and immediately indicate the final third operation or its result as their first step (or immediately verify the final answer). Therefore, a combinatorial pattern of 20 one- and two-schema tasks could be effectively used to collect data on student performance in arithmetic word problem solving.

3.3.2.3 Evidence Model

The scoring procedure should reflect different levels of schemas (if any) applied by students while making their first solution step or verifying a suggested step. If the response is based on an immediate next step corresponding to the first schema in the detailed solution sequence for the task, a score 1 should be allocated. If the response is one of the more advanced steps toward the solution (or the final answer), it should be allocated an additional score for each skipped step.

For example, for the above two-schema CG-task, responses at the level of the first schema (*C*-schema), $15+8$ or *23*, are scored as 1 or 2 respectively. Responses at the level of the second schema (*G*-schema) such as $23+15$, $15+8+15$, $(15+8)+15$ are allocated a score 3 (as an intermediate step for the second schema). Responding with (or verifying) the final answer (*38*) would attract a score 4 because three intermediate-level steps were skipped in this case.

In a rapid verification computer-based test, students could be presented the following instructions:

> On the following screens, you will see 20 arithmetic problems. You will be allowed a limited time to study each problem.
>
> For each task, several possible (correct or incorrect) solution steps will be presented one at a time. Spend no more than a few seconds to indicate if the provided solution step is correct or incorrect. Click on the "RIGHT" button if you think the step is CORRECT or the "WRONG" button if the step is INCORRECT. If you do not know the answer, click on the "DON'T KNOW" button.

The suggested approach was tested as the first-step technique in a realistic class environment (a paper-based format) with a sample of 55 grade 8 students (Kalyuga 2006d) and compared with a traditional test asking students to write complete solutions to 20 similar (although not identical) problems using a partial credit scoring procedure based on students' written solutions. The rapid test was 2.8 times faster, with a significant correlation of 0.72 between scores for both tests indicating a sufficient predictive validity of the rapid test.

The traditional classical test theory procedures are usually focused on one-dimensional overall performance indicators. If distinct schemas are defined in the models of student expertise, appropriate multidimensional measurement models could be used to assess each construct separately. In the arithmetic word problems area, two different multidimensional measurement approaches were applied to the data (Kalyuga 2006b, d). One approach was based on a multidimensional Rasch model (Adams et al. 1997). Another approach was based on Bayesian conditional probabilities estimations using the Markov chain Monte Carlo (MCMC) estimation procedure (Gelman et al. 1995).

In the multidimensional Rasch model, a student's position in the four-dimensional space was defined by a set of four parameters corresponding to four schemas. The ConQuest software for the partial credit model was used to carry out the multidimensional analysis (Wu et al. 1998). Model fit estimates generated by the software indicated acceptable ranges of values for most items. For each student, values of the knowledge variables for each schema dimension and corresponding error variances were determined.

The Bayesian conditional probability model is based on a certain assumption about probabilities $P(X \mid S)$ of observing a set of scores X for 20 tasks if the four-dimensional set S of a student's knowledge parameters (according to the student model) is known. If some prior hypothetical distribution $P(S)$ of these variables in the population of interest is defined, it is possible to apply the Bayes theorem to calculate the probability distribution for student parameters conditional on observed test scores, $P(S \mid X) \sim P(X \mid S)\, P(S)$. Then, the updated probability distribution could be used as a prior distribution for the next step of updating in the iterative process. For a prior distribution $P(S)$, the same categorical distribution for all students and for all four schematic dimensions was defined. The WinBUGS computer program (Bayesian inference Using Gibbs Sampling) was used to estimate posterior distributions conditional on the response data obtained in the experiment (Spiegelhalter et al. 2003). For each student, posterior means and standard deviations for parameters of each schema were estimated.

Although rough and simplified multidimensional methods were used, both models worked reasonably well and produced well-correlated (average correlation of 0.77) estimates of the parameters of students' schemas. Even though these results show that multidimensional measurement models could be used for making statistical inferences about learners' schematic knowledge structures, their application is not always practically plausible in small-scale formative assessments or during training sessions in adaptive instructional systems.

For each learner, a simple data summary using total scores for each schema based on the learner responses to the tasks that involve the corresponding schemas could do equally well. For two-schema items, the score for the first schema could be identical to the entire item score, while the second schema could be scored 1 if the item score is 3, or 2 if the item score is 4. For each of the four schemas, eight tasks contributed to the schema's score (e.g., tasks C, CC, CG, CV, CR, GC, VC, and RC contributed to the *C*-schema score; the last six items in this set also contributed to other dimensions). The summary scores for each schema dimension correlated significantly (between 0.80 and 0.96) with the parameters for levels of acquisition of corresponding schemas estimated by the two multidimensional measurement models.

3.4 Toward Rapid Dynamic Assessment for Learning

The rapid diagnostic methods could be related to dynamic assessment (Bransford and Schwartz 1999; Grigorenko and Sternberg 1998; Sternberg and Grigorenko 2001, 2002). Dynamic assessment is aimed at determining a learner's current stage of development at which he or she can solve a task if a certain level of guidance or help is provided, for example, by showing previous solution steps or hints. For example, if a student fails an item, she could be provided with a hint. If it does not help, another more detailed hint could be presented and the process repeated.

In rapid assessment methods, learners are presented with tasks reflecting various stages of a solution procedure with a gradually changing number of previously completed steps (e.g., see the previously described task model for rapid assessment in coordinate geometry) for making their next step or for rapid verification. Such task sequences effectively represent a form of scaffolding that is used to determine the precise level of learner expertise. This approach also effectively determines the learner zone of proximal development for dynamic selection of learning tasks that are just above the current level of expertise. Integrating the rapid diagnostic assessment approach with dynamic assessment into what could be called rapid dynamic assessment represents an important current direction of research and development in this area.

If learners are presented with incomplete intermediate stages of the task solution and asked to indicate the next step toward solution, they need to recognize both problem states and the solution moves associated with those states. Learners who are more advanced in the domain should be better able to recognize intermediate problem states and retrieve appropriate solution steps than less knowledgeable learners. For example, when training apprentices of manufacturing companies in reading charts used for setting cutting machines (Kalyuga et al. 2000), replacing visual on-screen texts with corresponding auditory explanations was beneficial for novice learners (modality effect). However, when learners became more experienced in using these charts, the best way to present a new type of charts was to display just a diagram without any explanations (an example of the expertise reversal effect Kalyuga et al. (2003)).

An appropriately designed series of rapid dynamic assessment tasks may allow switching instructional formats at the most appropriate time for an individual trainee. Such tasks may include regularly presenting trainees with a series of partially completed procedures in using charts with different degrees of completeness and asking them to indicate their next step toward solution. At the lowest level of completeness, no solution cues or hints are indicated on the chart. At the next level, only some relevant details of the task statement are highlighted. At the following levels, more lines and other solution details are shown. In this way, levels of expertise can be rapidly determined. Less knowledgeable learners then could be presented with comprehensive auditory explanations. In contrast, more experienced trainees, for whom the auditory explanations might be redundant, would learn better from a diagram with limited or no explanations.

Dynamic tests enhance students' learning and, at the same time, provide more accurate measures of current skill levels than traditional static tests. Students learn when they are tested, and they are tested when they learn. Integration of learning and testing into dynamic assessment formats is a current trend in the educational assessment field. For example, Feng et al. (2009) integrated continuous assessment and tutoring in their web-based tutoring system ASSISTment that combined assistance and assessment. The immediate tutoring is provided following each assessment item that students could not solve on their own. In addition to traditional scores based solely on correctness of students' responses to test items, the system collects data on its interactions with students (e.g., time taken to come up with an answer, response accuracy, and speed, time taken to correct an answer if it is wrong, help-seeking behavior as the number of requested hints, and solution attempts on sub-steps) that reflect their effort in solving the test item with instructional assistance in the form of hints, guidance, etc.

If students fail an item, they are provided with a small "tutoring" session where they must answer a few questions that break the problem down into steps. Thus, each ASSISTment task includes an *original question* and a list of *scaffolding questions* to coach students who fail to answer the original one. By analyzing these students' performance on the scaffolding questions, the system provides fine-grained diagnostic information. The system helps students to work through difficult problems by breaking them into sub-steps and meanwhile collecting data on different aspects of student performance (Feng et al. 2009). Thus, instruction is provided to students during the detailed evaluation of their knowledge and skills. As a result, a better evaluation of student abilities and prediction of their future performance is achieved. Since the ASSISTment system automatically provides students with feedback, scaffolding questions, and hints, it provides a form of embedded dynamic assessment.

3.5 Conclusion

The general idea of the rapid diagnostic assessment is to determine the level of most advanced domain-specific schemas (if any) a learner is capable of activating immediately on presentation of a test task. This assessment approach essentially evaluates the degree to which the learners' working memory capacity has been expanded due to available schemas in long-term memory. If a more knowledgeable learner is facing a task in a familiar domain, the relevant schemas are rapidly activated allowing the encapsulation of many elements of information (e.g., detailed solution operations and steps) in working memory into a single element (e.g., a higher-level advanced solution step). Different rapid responses would reflect different levels of acquisition of corresponding schemas. Thus, the rapidness of such tests is not only a means of reducing testing time, but it is essential for capturing schemas that learners use in specific situations before they can apply lengthy random search processes and chains of reasoning.

The rapid test tasks could be either used as stand-alone diagnostic probes or presented in a specific sequence. In order to qualify as dynamic assessment tasks, they should be developed as a series with a gradually changing number of completed essential steps or with different levels of instructional support provided in other forms. The diagnostic power of this rapid dynamic assessment may approach that of laboratory-based concurrent verbal reports, however achieved on a considerably shorter time scale.

3.5.1 Future Developments

3.5.1.1 Establishing Generality of the Tool

The examples and studies described in this chapter were limited to relatively narrow task areas associated with well-structured problems. In relatively poorly specified domains that involve problems with multiple possible routes to solutions, the rapid verification method could be potentially applied by selecting only a limited number of situations representing different possible paths and levels of solution steps (including both appropriate and unsuitable steps) for rapid verification. The generality and limits of usability of rapid assessment, especially in poorly structured domains, need to be investigated in further research.

In addition to domain-specific schemas, understanding verbally presented problems may also depend on reading comprehension skills and factual knowledge used in specific problem contexts. Therefore, while such tests could be usable with relatively advanced learners (e.g., secondary or high school students) for whom such factors may not influence results, their suitability for less advanced learners (e.g., primary school students) whose responses may depend on a wider range of factors needs to be further investigated.

3.5.1.2 Using Rapid Assessment in Adaptive Learning Environments

Rapid assessment methods have been applied in adaptive computer-based tutorials for high school students in solving linear algebra equations (Kalyuga and Sweller 2004, 2005) and vector addition motion problems in kinematics (Kalyuga 2006a). The levels of provided instructional guidance in tutorials were based on rapid measures of learner expertise. At the beginning of each session, the initial rapid test was used to select the level of support. For learners with lower levels of expertise, based on the rapid test, additional worked-out examples were provided. For learners with higher levels of expertise, less worked examples and more problem-solving exercises were provided. During the session, rapid tests were used to select the optimal learning pathway. Based on those tests, each learner was either allowed to proceed to the next stage with a lower level of guidance or required to repeat the same stage and then take the rapid test again. At each subsequent stage of the tutoring session, a lower level of instructional guidance was provided to learners, and a higher level of the rapid diagnostic tasks was used at the end of the stage.

The adaptive tutorials resulted in higher learning outcomes than similar nonadaptive tutorials in which learners either studied all tasks that were included in the corresponding stages of the training session of their yoked participants or were required to study the whole set of tasks available in the tutorial. The described studies provided preliminary evidence for the usability of the rapid assessment methods in adaptive instruction. Similar rapid test-based approaches could be used in other domains (including relatively less structured subject areas) for initial selection of the appropriate formats of learning materials according to levels of users' prior knowledge in the domain, monitoring their progress during learning, and real-time selection of the appropriate learning tasks and instructional formats.

An important direction for further improvements of adaptive learning environments is using rapid dynamic assessment methods (rather than stand-alone rapid tests embedded into the learning sessions) that allow a full and seamless integration of learning and assessment. Rapid dynamic assessment methods could also be used for enhancing assessment oriented toward self-directed learning (Mok 2010) by providing students with real-time evaluation of their current progress in a task domain. To further improve self-directed learning, learner-controlled adaptive environments that provide learner-tailored guidance need to be developed and experimentally tested in future research studies.

References

Adams, R., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Baddeley, A. (1997). *Human memory: Theory and practice*. East Sussex: Psychology Press.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 61–101). Washington, DC: American Educational Research Association.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Mind, brain, experience, and school*. Washington, DC: National Academy Press.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55–81.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanation: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

de Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton.

Embretson, S. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Mahwah: Erlbaum.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*, 725–747.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211–245.

Ericsson, K. A., & Simon, H. A. (1993). Verbal reports as data. *Psychological Review, 87*, 215–251.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction, 19*, 243–266.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124*, 75–111.

Kalyuga, S. (2006a). Assessment of learners' organized knowledge structures in adaptive learning environments. *Applied Cognitive Psychology, 20*, 333–342.

Kalyuga, S. (2006b). *Instructing and testing advanced learners: A cognitive load approach*. New York: Nova.

Kalyuga, S. (2006c). Rapid assessment of learners' proficiency: A cognitive load approach. *Educational Psychology, 26*, 613–627.

Kalyuga, S. (2006d). Rapid cognitive assessment of learners' knowledge structures. *Learning and Instruction, 16*, 1–11.

Kalyuga, S. (2008). When less is more in cognitive diagnosis: A rapid online method for diagnostic learner task-specific expertise. *Journal of Educational Psychology, 100*, 603–612.

Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology, 96*, 558–568.

Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development, 53*, 83–93.

Kalyuga, S., Chandler, P., & Sweller, J. (2000). Incorporating learner experience into the design of multimedia instruction. *Journal of Educational Psychology, 92*, 126–136.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). Expertise reversal effect. *Educational Psychologist, 38*, 23–31.

Larkin, J., McDermott, J., Simon, D., & Simon, H. (1980). Models of competence in solving physics problems. *Cognitive Science, 4*, 317–348.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction, 13*, 251–283.

Marshall, S. P. (1993). Assessing schema knowledge. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 155–179). Mahwah: Lawrence Erlbaum Associates.

Marshall, S. (1995a). *Schemas in problem solving*. New York: Cambridge University Press.

Marshall, S. (1995b). Some suggestions for alternative assessments. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 431–453). Hillsdale: Lawrence Erlbaum Associates.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363–389.

Mok, M. M. C. (2010). *Self-directed learning oriented assessment: Assessment that informs learning and empowers the learner*. Hong Kong: Pace Publishing.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 307–353). Washington, DC: AERA.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Research Council's Committee on the Foundations of Assessment. Washington, DC: National Academy Press.

Peterson, L., & Peterson, M. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58*, 193–198.

Pirolli, P., & Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review, 105*, 58–82.

Russell, M., O'Dwyer, L., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods, 41*(2), 414–424.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Mahwah: Lawrence Erlbaum Associates.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (pp. 263–331). New York: Macmillan.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual* (Version 1.4) [Computer software]. Cambridge: MRC Biostatistics Unit. Retrieved from http://www.mrc-bsu.cam.ac.ik/bugs

Sternberg, R. J., & Grigorenko, E. L. (2001). All testing is dynamic testing. *Issues in Education, 7*, 137–170.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge: Cambridge University Press.

Sweller, J. (2004). Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture. *Instructional Science, 32*, 9–31.

Sweller, J., van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–487). Hillsdale: Lawrence Erlbaum Associates.

Van Merriënboer, J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*, 147–177.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalized item response modeling software [computer program]*. Camberwell: ACER Press.

Chapter 4
Standardized Diagnostic Assessment Design and Analysis: Key Ideas from Modern Measurement Theory

Hye-Jeong Choi, André A. Rupp, and Min Pan

4.1 Introduction

Traditional standardized standards-based assessments created by professional agencies and partially standardized standards-based assessments made by teachers for assessments at the end of a unit, chapter, or term can be reliable indicators of general states of proficiency for groups of students. In short, they serve general monitoring and accountability purposes in selected key domains such as reading, mathematics, and science rather well. However, as Linn (1986) emphasized, they typically have very little or no instructional uses:

> a test that reliably rank orders students in terms of global test scores provides a teacher with relatively little information about the nature of a student's weaknesses, errors, or gaps. For example, the knowledge that a student scores, say, in the 10th percentile on a standardized arithmetic test suggests the student has a general weakness in the area of arithmetic relative to his or her peers. However, such a score does not, by itself, indicate the source of the problem or what should be done to improve the student's level of achievement; that is, it lacks diagnostic information. (p. 1158)

The seemingly increasing dissatisfaction in the field of education with the structure and potential uses of standardized standards-based assessments for guiding and

H-J. Choi
Office of Research and Department of Psychology,
University of Nebraska-Lincoln, Athens, GA, USA
e-mail: hchoi3@unl.edu

A.A. Rupp (✉)
Department of Human Development and Quantitative Methodology (HDQM),
University of Maryland, College Park, MD, USA
e-mail: ruppandr@umd.edu

M. Pan
Department of Measurement, Statistics, and Evaluation,
University of Maryland, College Park, MD, USA
e-mail: minr.l@foxmail.com

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_4, © Springer Science+Business Media Dordrecht 2013

evaluating the students' fine-tuned knowledge state motivated the development of more *diagnostic assessments*. Diagnostic assessments play a key role in establishing an alignment between developmental theories about learning in a domain, curricular objectives as set forth by policy documents, teacher practice in the classroom, and actual learning gains made by students (e.g., Leighton and Gierl 2007, 2011).

4.1.1 Assessment Of, For, and As Learning

The current literature on modern educational measurement for diagnostic assessment purposes makes a distinction between assessments *of*, *for*, and *as* learning, which helps to differentiate the various layers of interpretations drawn from them and the diverse uses to which they are put (e.g., O'Reilly et al. 2008; Mok 2010).

The phrase *assessment of learning* suggests that one purpose of assessments is to identify the achievement of the students at the end of a learning cycle to obtain a rich and sufficiently detailed picture of the degree to which students have met their targeted learning objectives. The information gathered from an assessment can support summative interpretations that allow for overall comparisons of how individual students perform relative to their peers.

The phrase *assessment for learning* suggests that the purpose of an assessment can also be to monitor the continual, ongoing learning process in order to provide directive and supportive feedback in a scaffolding process. The information is collected to seek for answers as to what underlying mechanisms drive the problem-solving strategies enacted by the students so as to make the learning process most efficient, effective, and engaging for the students.

The phrase *assessment as learning* suggests that the purpose of assessment is to make students self-directed by improving their level of metacognition. The process of assessment thus induces the cultivation of a capacity for goal setting, self-monitoring of the learning process, self-assessment of achievement, self-motivation, and self-regulation to enhance further learning.

In terms of assessment for learning in particular, what many teachers seek to guide their day-to-day instructional practice are more fine-grained descriptions of students' proficiency profiles, which are necessary to designing effective instructional interventions that make students efficacious (i.e., efficient and effective) in the targeted domains. Teachers continually collect potentially diagnostic information in informal or partially standardized ways on a daily basis. For instance, teachers may ask questions regarding what concepts or strategies students have mastered and which ones they are still struggling with; they may ask specifically why some students do not understand a particular aspect of what they have taught in class, or they may inquire about whether it is necessary to create certain types of additional opportunities for practice in class. In short, teachers are constantly concerned with how they can construct classroom environments which fit individual student's current learning needs best.

4.1.2 Measurement Models for Diagnostic Assessment Data

Traditional measurement models that can support inferences from summative assessments for quantitative rank-order purposes include predominantly models from the fields of *classical test theory* (CTT) (e.g., Lord and Novick 1968; Crocker and Algina 2006) and *item response theory* (IRT) (e.g., de Ayala 2009; Yen and Fitzpatrick 2006) even though *factor-analytic* (FA) models (e.g., McDonald 1999) can serve these purposes as well. However, the score reports created on the basis of data calibrations with these models are, at best, only partially useful for supporting more formative interpretations for qualitative diagnostic purposes.

Typically, CTT, IRT, and FA models are applied to large-scale standardized standards-based assessments of learning whose operational construct is defined at a rather coarse level of cognitive grain size thus leading to relatively coarse descriptions of students' proficiency levels in the target domain. In contrast, *diagnostic classification models* (DCMs) (e.g., Rupp and Templin 2008; Rupp et al. 2010) are models that are particularly suitable for large-scale standardized assessments for learning whose operational construct is defined at a finer level of cognitive grain size thus supporting more nuanced descriptions about students' proficiency profiles.

In this chapter, we present a few key ideas that are relevant to developing cognitively diagnostic assessments for learning and scaling them with DCMs. Specifically, in the next section, we present a key framework for principled assessment design that can be employed in powerful ways for developing cognitive diagnostic assessments. In the section after that, we introduce a unified specification and estimation framework for DCMs and illustrate its utility for operationalizing different cognitive theories of responding. In the final main section, we present a real-data analysis of a small section of a newly developed diagnostic mathematics assessment to illustrate how DCMs can be used for calibrating the instrument and classifying the students into different proficiency profiles.

4.2 Evidence-Centered Design

Some form of applied cognitive theory (e.g., influenced by information-processing or socio-cognitive perspectives) is necessary to design any test whose items or tasks are supposed to reflect the essential knowledge, skills, and abilities that are to be measured (NRC 2001). Arguably, the explicit focus on fine-grained proficiency profiles for students that can inform learning processes in an assessment for learning sense puts the explication and operationalization of applied cognitive theories at the forefront of diagnostic assessment design. In this chapter, we focus on an important design framework called *evidence-centered design* (ECD).

The ECD (Mislevy et al. 2003, 2004) framework provides a formal structure for *evidence-based reasoning* that provides guidance to interdisciplinary teams of experts who are charged with developing a wide range of assessments for a wide range of purposes. Despite its generality, its power for structuring assessment development, implementation, and score reporting is arguably most evident for assessments that involve *complex performance-based tasks*. The reason for this is that the number of decisions about designing tasks with appropriate constraints, identifying suitable task products, identifying individual pieces of evidence and scoring them, aggregating these scores with the help of modern statistical models, and reporting these scores back to students and stakeholders are much larger and arguably more complex in these contexts than in assessments that employ more selected-response formats.

The core purpose of diagnostic assessment development from an ECD framework perspective is the development of coherent *evidentiary arguments* in an *assessment narrative* about students that can serve as assessment of and assessment for learning, depending on the desired primary purpose of a particular assessment. The structure of the evidentiary arguments that are used in the assessment narrative can be described with the aid of terminology first introduced by Toulmin (1958).

An evidentiary argument is constructed through a series of logically connected *claims or propositions* that are supported by data through *warrants* and *backing* and can be subjected to *alternative explanations*. In diagnostic assessments, data consist of students' observed responses to particular tasks and the salient features of those tasks, claims concern examinees' proficiency as construed more generally, and warrants posit how responses in situations with the noted features depend on proficiency. Statistical models such as DCMs provide the mechanism for evaluating and synthesizing the evidentiary value in a collection of typically overlapping, often conflicting, and sometimes interdependent observations.

In concrete terms, the ECD framework allows one to distinguish the different structural elements and the required pieces of evidence in narratives such as the following:

> Jamie has most likely mastered basic addition (*claim*), because she has answered correctly a mathematical problem about adding up prices in a supermarket (*data*). It is most likely that she did this because she applied all of the individual addition steps correctly (*backing*) and the task was designed to force her to do that (*backing*). She may have used her background knowledge to estimate the final price of her shopping cart (*alternative explanation*), but that is unlikely given that the final price is exactly correct (*refusal*).

The ECD framework specifies five different assessment design components, which are shown in Fig. 4.1 below.

Guided by the theory-driven process of analyzing and modeling the key facets of expertise in a domain, the core elements in the ECD framework include (1) the *student models*, which formalize the postulated proficiency structures for different tasks, (2) the *task models*, which formalize which aspects of task performance are coded in what manner, and (3) the *evidence models*, which are the psychometric models linking those two elements. These three core components are complemented by (4)

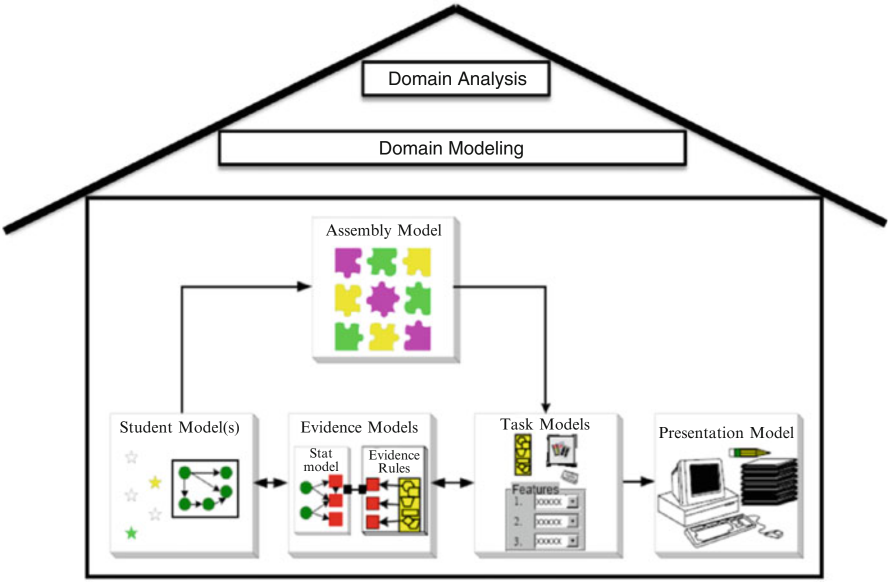


Fig. 4.1 The ECD model (Adapted from Mislevy et al. 2004)

the *assembly model*, which formalizes how these three elements are linked in the assessment, and (5) the *presentation model*, which formalizes how the assessment tasks are being presented.

Specifically, the *student model* is motivated by the learning theory that underlies the diagnostic assessment system. It specifies the relevant variables or aspects of learning that we want to assess at a grain size that suits the purpose of the diagnostic assessment. As many of the characteristics of learning that we want to assess are not directly observable, the student model provides a probabilistic or proxy model for making claims about the state, structure, and development of a more complex underlying system. This might concern a trait or a behavioral disposition in a traditional assessment. In more innovative diagnostic assessments in education such as a game or simulation, it could instead concern the models or strategies a student seems to employ in various situations, or the character or interconnectivity of his or her skills when dealing with certain kinds of situations in a discipline.

To make claims about learning as reflected through changes in the attributes in the student model, we thus have to develop a pair of *evidence models*. The *evaluation component* of the evidence model specifies the salient features of whatever the student says, does, or creates in the task situation, as well as the rules for scoring, rating, or otherwise categorizing the salient features of the assessment. The *probability or statistical component* of the evidence model specifies the rules by which the evidence collected in the evaluation is used to make assertions about the student model. This means that a suitable statistical model such as a DCM needs to be selected for summarizing observed information contained in indicator variables via statistically created, and typically latent, variables. The statistical model provides the machinery for updating beliefs about student model variables in light of this information. Taken together, evidence models provide a chain of inferential reasoning from observable performance to changes that we believe are significant in a student's cognitive, social, emotional, moral, or other forms of development.

The *task model* provides a set of specifications for the environment in which the student will say, do, or produce something. That is, the task model specifies the conditions and forms under which data are collected, and the variables in a task model are motivated by the nature of the interpretations the assessment is meant to support. Data collected in such models are not restricted to traditional formal, structured, pencil-and-paper assessments and can include information about the context, the student's actions, and the student's past history or particular relation to the setting.

The *assembly model* describes how these different components are combined for answering particular questions about learning in a given assessment situation. Using the analogy of *reusable design templates* within a task bank, the assembly model describes which task model, evidence model, and student model components are linked for a particular assessment or subsections of an assessment. The idea of a reusable design template is similar to the idea of automatic task generation within the general cognitive design system (e.g., Embretson, 1998) framework. However, rather than striving for an automatic generation, the ECD framework strives for principled construction under constraints that will result in tasks that are comparable to one another, both substantively and statistically.

Similarly, the *presentation model* describes whether modes of task and product presentation change across different parts of the assessment and what the expected implications of these changes are. In practice, ECD models for a given assessment are constructed jointly and refined iteratively because the full meaning of any model only emerges from its interrelationship with other components.

ECD has been successfully applied in different fields. *PADI*, *ECDLarge* and *NetPASS* are comprehensive ongoing assessment projects that are based on ECD. Specifically, PADI aims at developing assessments of science inquiry that combine new developments in cognitive psychology, science inquiry, as well as measurement theories and techniques (e.g., Mislevy and Riconscente 2005; see also http://padi.sri.com/index.html). ECDLarge is a successor to the PADI project that focuses on the application of the ECD framework to the development of large-scale assessments (see http://ecd.sri.com/index.html for more information). The NetPASS project is concerned, in part, with developing an authoring tool and simulation-based learning and assessment environment to train network engineers within the context of Cisco Networking Academy Program (e.g., Levy and Mislevy 2004; Mislevy et al. 2003; Rupp et al. in press; West et al. 2009; see also http://cisco.netacad.net/public/index.html). The set of applications cited here, taken together, illustrate the power of the ECD framework for developing a wide range of assessments that can support a wide range of inferences including fine-grained diagnostic feedback for formative assessment purposes as well as more coarse-grained feedback for summative accountability purposes.

The previous presentation is not meant to suggest that individual teachers have to think about the ECD framework during their day-to-day practice. However, we believe that teachers may find the language, conceptualization, and key assessment principles embedded within the ECD framework quite accessible and useful for shaping their own professional understanding. The ECD framework can also be very powerful for professional development purposes at the district or state level because it provides a coherent frame for structuring evidentiary arguments about students in a common language. This is essential for developing effective diagnostic assessment systems where experts from different disciplines have to work together efficaciously.

Importantly, the ECD framework underscores, but does not overemphasize, the importance of the statistical models that are used in the evidence model component. Statistical models such as DCMs are tools for reasoning about patterns of behavior of students based on data patterns with differential weighting. However, the choice of how the behavioral patterns are modeled and, thus, which real-life elements are represented in a statistical model, is squarely in the hands of the diagnostic assessment developer. The next section now discusses DCMs as a particular class of modern measurement models that can be useful for analyzing data from standardized diagnostic assessments.

4.3 Diagnostic Classification Models

Before beginning our discussion of DCMs, we want to reiterate that many modeling choices are driven by substantive considerations about the structure of desired evidence-based assessment narratives for students. That is, based on the desired level of precision at which a student characteristic is to be measured and interpretations are to be given as well as the real-life constraints imposed by the informational richness of the available data, diagnostic assessment designers have to decide which characteristics should be represented via variables in the DCM that is chosen for analysis. They need to decide which pieces of information are extracted from the complex performance of students and how these pieces of information are coded so that they can be used as input into the statistical models. The choice or construction of any statistical model thus emerges from a careful consideration of students, learning, situations, and theory; it does not or should not determine what interpretations should be or what observations must be limited to.

In this section, we introduce DCMs as a particular class of statistical models that can be useful for standardized diagnostic assessment processes. Specifically, we first discuss key terminology, then describe a unified specification and estimation framework for DCMs, and finally illustrate, using real data from a newly developed diagnostic assessment of elementary school mathematics, how one can estimate DCMs with a commercially available software program.

Table 4.1 Exemplary Q-matrix

Item		Addition	Subtraction	Multiplication	Division
1	$2+3-1$	1	1	0	0
2	4/2	0	0	0	1
3	$5\times3-4$	0	1	1	0
4	$8+12$	1	0	0	0

4.3.1 Attributes, Attribute Profiles, and Q-matrices

The term *attribute* generically refers to unobservable (i.e., latent) characteristics of students. In DCMs, we will operationalize these characteristics using unobservable (i.e., latent) variables. We use the values on these latent variables to reason backwards about students' mastery states on the attributes of interests based on students' observed response patterns to diagnostic assessment items. The resulting pattern of attribute mastery states are known as *attribute profiles* in the literature; under an effective diagnostic assessment design, attribute profiles carry reliable information for making meaningful instructional decisions.

Once the targeted attributes and potential attribute profiles are determined based on an appropriate applied cognitive theory, the next step is to specify which attributes are measured by each individual assessment item (i.e., which attributes are required by the students to obtain a maximum score on an item). The relationship between attributes and items is formally captured in a two-dimensional table known as a *Q-matrix* (Tatsuoka 1990). In general, rows of the table correspond to items, columns of the table correspond to attributes, and entries in the table are typically binary (i.e., "0" or "1"), indicating which attributes are measured by which items.

There are a number of ways of constructing Q-matrices. In educational testing, Q-matrices may be constructed based on theories about learning in the domain triangulated by experts' judgment, empirical research, think-aloud protocols, factor analyses of existing tests, and other means of empirical validation (Buck and Tatsuoka 1998; Gierl et al. 2005). To illustrate the structure of a Q-matrix in practice, we use an example scenario where five items measure four attributes in basic arithmetic ability; this matrix is shown in Table 4.1.

According to this Q-matrix in Table 4.1, item 2 and item 4 only measure one attribute, while item 1 and item 3 measure two attributes. Expressed reversely, only mastery of one attribute is required for item 2 and item 4 to get the maximum score on these items, while mastery of two attributes is required for the other two items.

Consequently, the attribute profile (i.e., the mastery state on all attributes measured by the diagnostic assessment) of each student can be represented in the same way using binary indicators where "1" indicates that a student has mastered an attribute, and "0" indicates that he or she has not. For instance, if a student has mastered only the first two attributes among the four attributes above, his or her attribute profile can be represented as [1,1,0,0].

Given the Q-matrix ($\boldsymbol{Q}$) and a student's attribute profile ($\boldsymbol{\alpha}$), an idealized response pattern (i.e., a response pattern that would be observed if the student responded without error) can be predicted through simple matrix algebra as follows:

$$\mathbf{Q} \times \boldsymbol{\alpha} = \begin{bmatrix} 1100 \\ 0001 \\ 0110 \\ 1000 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0/1 \\ 1 \end{bmatrix}.$$

In this example, the student should respond correctly to item 1 and item 4 but not to item 2. It is not clear, however, whether or not this student would respond correctly to item 3 as he or she has only mastered one out of the two required attributes. Different DCMs are designed to operationalize different relationships between the mastery states on individual attributes and the probabilities of a certain response while allowing for imperfect responding due to random errors.

4.3.2 A Definition of DCMs

DCMs are statistical models that were developed to respond to the desire of diagnostic assessment developers to classify students in terms of their mastery states on individual attributes that constitute of their attribute profiles (for overviews see, e.g., DiBello et al. 2007; Rupp and Templin 2008; Rupp et al. 2010; Templin 2004). Formally,

> Diagnostic classification models (DCMs) are probabilistic confirmatory multidimensional latent variable models. Their loading structure / Q-matrix is typically complex to reflect within-item multidimensionality, but may also be simple. DCMs are suitable for modeling observable reponse variables (i.e., dichotomous, polytomous) and contain unobservable latent categorical predictor variables (i.e., dichotomous, polytomous). The predictor variables are combined in compensatory and non-compensatory ways to generate latent classes. DCMs enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. (Rupp et al. 2010, p. 108)

The literature is replete with DCMs that differ in the number of parameters that they contain for items and attributes and the types and numbers of restrictions they place on these parameters; in other words, the flexibility with which they can handle various data structures. Rather than listing all of these models here, we refer to the overview sources cited earlier for detailed descriptions of these models. More importantly, current theory and practice has evolved to the point where many DCMs can now be parameterized as special cases of more general modeling families.

The three most common families in the literature are the *log-linear cognitive diagnosis model* (LCDM) framework by Henson et al. (2009), the *general diagnostic model* (GDM) framework by von Davier (2005, 2010), and the generalized deterministic inputs, noisy "and" gate (G-DINA) model by de la Torre (2009). For the purposes of this chapter, we will use the LCDM framework and refer to the chapter by de la Torre (Chap. 5, this volume) for an overview of the G-DINA model framework.

4.3.3 The LCDM Framework

As the GDM and G-DINA frameworks, the LDCM framework is a unified framework for the specification and estimation of DCMs. Its development was based on finite mixture models (e.g., McLachlan and Peel 2000), log-linear models (e.g., Agresti 2010), and generalized linear and latent mixed models (e.g., Skrondal and Rabe-Hesketh 2004). In the following, we will focus on the simplest case of an LCDM, which concerns binary item scores (i.e., “1” for a correct response and “0” for an incorrect response), binary attribute mastery states (i.e., “1” for a mastered attribute and “0” for a non-mastered attribute), and binary Q-matrix entries (i.e., “1” for an attribute that is measured by an item and “0” otherwise); extensions are relatively easily specified and estimated.

4.3.3.1 Model Specification

In the LCDM, the probability of a correct response as a function of attribute mastery states is defined as

$$\mathrm{P}\left(\mathrm{Y}_{\mathrm{ij}} = 1 \middle| \boldsymbol{\alpha}_{\mathrm{i}}, \mathbf{q}_{\mathrm{j}}\right) = \frac{\exp\left[\lambda_{0\mathrm{j}} + \boldsymbol{\lambda}'_{\mathrm{j}}\ \mathrm{h}\left(\boldsymbol{\alpha}_{\mathrm{i}}, \mathbf{q}_{\mathrm{j}}\right)\right]}{1 + \exp\left[\lambda_{0\mathrm{j}} + \boldsymbol{\lambda}'_{\mathrm{j}}\ \mathrm{h}\left(\boldsymbol{\alpha}_{\mathrm{i}}, \mathbf{q}_{\mathrm{j}}\right)\right]}, \tag{4.1}$$

where i and j denote student and item, respectively; λ_{0j} is an intercept and $\boldsymbol{\lambda}_j$ represents a vector of coefficient indicating the effects of attribute mastery on the response probability for item j, and $h(\boldsymbol{\alpha}_i, \mathbf{q}_i)$ is a set of linear combinations of the attribute mastery indicators $\boldsymbol{\alpha}_i$ and the Q-matrix entries $\mathbf{q}_j$. Specifically, the kernel of the above expression has the following general form:

$$\lambda_{0j} + \boldsymbol{\lambda}_j h\left(\boldsymbol{\alpha}_i, \mathbf{q}_j\right) = \lambda_{0j} + \sum_{u=1}^{k} \lambda_{ju}\left(\alpha_k q_{ju}\right) + \sum_{u=1}^{k}\sum_{v>u} \lambda_{juv}\left(\alpha_u \alpha_v q_{ju} q_{jv}\right) + \cdots \tag{4.2}$$

which is similar to the structure of factorial analysis of variance (ANOVA) models.

The intercept can be interpreted as a *guessing parameter* because it reflects the probability of providing a correct response for those students who have not mastered any attributes – this is the lowest probability for any attribute profile. The λ_{ju} parameters represent the main effects of each attribute on the response probability for item j, and the λ_{juv} parameters represent the two-way interaction effects of the combination of the mastery states of attributes u and v on the response probability for item j; higher-order parameters are defined likewise with aligned meanings. In other words, the specification of the kernel follows the specification of factorial ANOVA models with intercept, main-effect, and interaction-effect parameters.

Depending on how many attributes are included in the item, the LCDM can include main effects for each attribute, two-way and three-way interactions among attributes, and so forth. Simulation studies (Kunina-Habenicht, Rupp, and Wilhelm, 2012; Choi et al. 2010) have shown that interaction-effect parameters require very large sample sizes for reliable estimation, however, so that main-effect parameter specifications are probably most appropriate for most practical contexts.

To illustrate the general expression for the LCDM with a concrete example, consider the Q-matrix from Table 4.1. Since item 1 measures attribute 1 and attribute 2, $q_{11} = q_{12} = 1$, while $q_{13} = q_{14} = 0$. Consequently, the probability of a correct response for item 1 takes the form

$$P\left(Y_{i1} = 1 \,|\, \boldsymbol{\alpha}_i, \mathbf{q} = (1,1,0,0)\right) = \frac{\exp\left(\lambda_{10} + \lambda_{11}\alpha_1 + \lambda_{12}\alpha_2 + \lambda_{112}\alpha_1\alpha_2\right)}{1 + \exp\left(\lambda_{10} + \lambda_{11}\alpha_1 + \lambda_{12}\alpha_2 + \lambda_{112}\alpha_1\alpha_2\right)}, \quad (4.3)$$

with the exact probability values for each attribute profile (i.e., each combination of attribute mastery states for attribute 1 and attribute 2) depending on the values of the item parameters λ_{10}, λ_{11}, λ_{12}, and λ_{112}, which need to be estimated in practice from the student response data.

4.3.3.2 Illustrative Special Cases

As the response probability for this item is influenced by the mastery states on two attributes, we can ask several questions: What is the response probability for students who have mastered only one attribute out of two? Does mastering attribute 1 have a bigger impact on the response probability than mastering attribute 2? Is there an additional effect on the response probability for mastering both attributes once one of them has already been mastered?

These questions can be answered empirically either by specifying the most general DCM in Eq. 4.2 and inspecting the values of the resulting parameter estimates a posteriori or by specifying specific DCMs that reflect different hypotheses in alignment with these three questions a priori. To illustrate the flexibility of the LCDM framework, we discuss particular DCMs that would result from such a priori specifications in the following.

For the first scenario, if the DCM is supposed to reflect the assumption that both attributes need to be mastered to provide a correct response, then Eq. 4.3 can be modified as follows:

$$P\left(Y_{i1} = 1 \,|\, \boldsymbol{\alpha}_i, \mathbf{q} = (1,1,0,0)\right) = \frac{\exp\left(\lambda_{10} + (0)\alpha_1 + (0)\alpha_2 + \lambda_{112}\alpha_1\alpha_2\right)}{1 + \exp\left(\lambda_{10} + (0)\alpha_1 + (0)\alpha_2 + \lambda_{112}\alpha_1\alpha_2\right)}, \quad (4.4)$$

Here, the main effects for attribute 1 and attribute 2 are set to 0, and only the intercept and interaction effect take on non-zero values. Thus, the response probabilities for this item are identical for students who have not mastered any of the two or only one of the two measured attributes. This model is referred to as the *deterministic input*, *noisy "and" gate* (DINA) model in the literature and substantively reflects a situation where the mastery of a subset of attributes cannot compensate for the lack of mastery of any other attribute(s) that is not mastered by a student but measured by an item (e.g., Junker and Sijtsma 2001; de la Torre 2009). In substantive terms for our simple example, this model reflects the assumption that students are not likely to solve item 1 if they have not mastered both addition and subtraction.

For the second scenario, consider the case where an item can be solved when only one of several attributes has been mastered. For example, suppose that students are asked to determine the interior angle of a regular pentagon. Some students may draw a picture to determine how many triangles there are in a pentagon. Once they figure out that there are three triangles inside the pentagon, the answer becomes 180 * 3 = 540 because the interior angle of a triangle is 180. Others may solve the same question using the analytic knowledge that for any regular polygon, the sum of the interior angles = 180(n − 2) where n is the number of sides. Since a pentagon has five sides, 180(5 − 2) = 540. If both strategies were coded as attributes that this item measured, then mastering both attributes does not increase the probability of a correct response.

For this situation, Eq. 4.3 can be modified as follows:

$$P\left(Y_{i1} = 1 \mid \boldsymbol{\alpha}_i, \mathbf{q} = (1,1,0,0)\right) = \frac{\exp\left(\lambda_{10} + \lambda_1\alpha_1 + \lambda_1\alpha_2 + (-\lambda_1)\alpha_1\alpha_2\right)}{1+\exp\left(\lambda_{10} + \lambda_1\alpha_1 + \lambda_1\alpha_2 + (-\lambda_1)\alpha_1\alpha_2\right)}, \quad (4.5)$$

where the probability of getting a correct answer for those who possess the knowledge about triangles, those who have mastered analytic knowledge, or those who know both is the exactly same. This model is referred to as the *deterministic input*, *noisy* "or"*gate* (DINO) model in the literature and reflects the assumption that mastery of subset of attribute can compensate for the lack of mastery of other attribute(s) (e.g., Templin and Henson 2006).

For the third scenario, consider the case where the probability of getting a correct response to an item increases as the number of mastered attributes increases. For example, suppose that a reading comprehension item with a passage regarding physics is presented to students. The impact of understanding the meaning of a certain vocabulary in the text and knowledge of syntactic structure may be additive on the probability of students' correct answer.

In this case, Eq. 4.3 can be modified as follows:

$$P\left(Y_{i1} = 1 \mid \boldsymbol{\alpha}_i, \mathbf{q}_j\right) = \frac{\exp\left(\lambda_{10} + \lambda_{11}\alpha_1 + \lambda_{12}\alpha_2 + (0)\alpha_1\alpha_2\right)}{1+\exp\left(\lambda_{10} + \lambda_{11}\alpha_1 + \lambda_{12}\alpha_2 + (0)\alpha_1\alpha_2\right)}, \quad (4.6)$$

where the interaction effect sets to zero, indicating no additional effect of mastering both attributes. This model is referred to in the literature as the compensatory reparameterized unified model (C-RUM) (e.g., Hartz 2002; Roussos et al. 2007) and also reflects the assumption that mastery of a particular attribute can compensate for the lack of mastery of any other attribute, albeit not as strongly as in the DINO model for scenario two above.

4.3.4 Estimating DCMs via the LCDM Framework

To date, there exist no specific software programs that are designed to specify and estimate DCMs within a user-friendly GUI environment. In the past, researchers have typically written their own estimation codes. For instance, the commercially available *Arpeggio* program (www.assess.com) was originally developed specifically for the RUM/Fusion model and requires sophisticated knowledge of Bayesian estimation for reliable use, the code for the G-DINA model was written in the programming language Ox (http://www.doornik.com/) and is still under development, and the program MDLTM for the GDM (von Davier 2006) originally relied on a syntax interface and is available as a research license only.

However, since DCMs are special cases of restricted latent class models, they can be estimated within any commercial program for latent class models that allows for the imposition of parameter constraints if a unified framework like the LCDM is used. For example, Choi et al. (2010), Templin et al. (2011), Kunina-Habenicht et al. (2010), and Rupp et al. (2010) have demonstrated how DCMs can be specified and estimated in M*plus*. In the following section, we present an additional example based on the data from Kunina et al. (2010).

4.4 Illustrative Extended Example

4.4.1 Data Description and Q-matrix

The *diagnostic mathematics assessment* (DMA) that is the focus of this example was developed to provide information on basic arithmetic ability for students in the 3rd and 4th grades in Germany (Kunina-Habenicht et al. 2009; 2010). Test items were constructed to measure several basic arithmetic skills such as addition, subtraction, multiplication, division, executing inverse operation, executing carry over, solving word problems, and converting measurement units. The original item pool consisted of 70 items and was administered to a sample of 2,032 4th grade students in different schools in Germany in 2008 using a complex booklet design (Frey et al. 2009). For illustration purposes, we analyzed only a subset of 20 items, which reflected the structure of the Q-matrix of the original item pool.

Even though several fine-grained skills were originally defined and used in the item development process, Kunina-Habenicht et al. (2010) found that a Q-matrix with four attributes was most strongly supported when various FA models and DCMs were used for data analysis. The four resulting attributes were addition/subtraction (A/S), multiplication/division (M/D), modeling (model), and converting units (units); Table 4.2 shows the Q-matrix for our example using the same attribute definitions. As shown in Table 4.2, items 1–10 measure one attribute, while items 11–20 measure two attributes.

Table 4.2 Q-matrix of diagnostic mathematics assessment (DMA)

Item	A/S	M/D	Model	Units
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	0	1	0	0
6	0	1	0	0
7	0	1	0	0
8	0	1	0	0
9	0	0	1	0
10	0	0	1	0
11	0	1	1	0
12	0	1	1	0
13	0	1	1	0
14	0	1	1	0
15	1	0	0	1
16	1	0	0	1
17	1	0	0	1
18	1	0	0	1
19	1	0	0	1
20	1	0	0	1

4.4.2 Model Selection and Item Parameter Estimation

For illustration purposes, we fit the four different DCMs to this data set that we discussed in the previous section, namely, the full LCDM, the DINA, the DINO, and the C-RUM. Recall that, for items that measure two attributes, the full LCDM model includes both main-effect parameters and the interaction-effect parameter; the DINA model contains only the two-way interaction-effect parameter; the DINO model contains both main-effect parameters and a negative two-way interaction-effect parameter, all constrained to equality; and the C-RUM contains only main-effect parameters. Thus, the full LCDM is the most flexible model, while the DINA model is the most restrictive model with the remaining two models representing special intermediate cases. All models were estimated in Mplus 6.0 (Muthén and Muthén 1998–2010).

After fitting the four competing models, relative model fit indices were used to determine the best-fitting model. We used *Akaike's information criterion* (AIC) (Akaike 1974) and Schwarz's (1978) *Bayesian information criterion* (BIC) that were provided in the output files. As is typical in practical applications, AIC and BIC did not always agree about the best-fitting model because they penalize differentially strong for the parametric complexity of the fitted models and sample size. As shown in Table 4.3, the AIC suggested that the C-RUM was the best-fitting model, while the BIC suggested that the DINA was the best-fitting model; according to the AIC, the full model is a close competitor to the C-RUM.

Table 4.3 Results of fit indices for model selection

	DINA	DINO	C-RUM	FULL
AIC	19352.16	19359.94	**19314.87**	19316.85
BIC	**19649.06**	19656.84	19665.75	19721.71
Number of parameters	55	55	65	75

Boldfaced entries indicate model with the smallest information criterion value

Table 4.4 Item parameter estimate from two models

	DINA						C-RUM				
		Main effect			Interaction effect			Main effect			
Item	Intercept	A/S	M/D	Model	(M/D)× (Model)	(A/S)× (Units)	Intercept	A/S	M/D	Model	Units
1	−0.74	2.28					−0.60	2.33			
2	−1.09	2.40					−0.99	2.53			
3	0.09	1.87					0.17	1.98			
4	0.42	1.98					0.51	2.07			
5	−2.08		2.67				−1.95		2.69		
6	0.37		1.92				0.42		2.04		
7	−0.26		2.85				−0.19		3.18		
8	−0.94		3.00				−0.81		3.14		
9	−1.85			2.04			−1.73			1.97	
10	−1.46			2.01			−1.40			2.05	
11	**−1.69**				**1.91**		**−1.90**		**0.43**	**1.80**	
12	−2.18				2.72		−2.46		0.98	2.16	
13	−1.18				1.79		−1.39		0.34	1.81	
14	−2.76				2.63		−3.00		1.74	1.25	
15	−0.67					1.92	−1.05	1.19			1.70
16	−1.92					2.73	−2.62	1.81			2.52
17	−0.70					2.09	−1.11	1.40			1.65
18	−0.32					1.82	−0.62	0.85			1.94
19	1.16					1.12	0.98	0.99			0.46
20	−2.17					2.17	−2.63	1.84			1.15

The item used for illustrative computations is shown in boldface

Which models one chooses does not matter for items 1–10 because those items measure only one attribute, but it matters for items 11–20 because they measure two attributes. To see the impact of choosing either the DINA or the C-RUM models for those items, we present the estimated model parameters for all 20 items in Table 4.4.

Since parameter estimates are on the logit scale and it is easier to think in terms of response probabilities, it is insightful to look at the difference in response probabilities for students with different attribute profiles under different models. Due to space limitations, we present here the corresponding response probabilities for item 11 in Table 4.5. As only the two attributes M/D and model were required for this item, only the mastery states for these two attributes influence the resulting response probabilities.

Table 4.5 Probability of a correct answer for item 11

Attribute				Model	
A/S	**M/D**	**Model**	Units	DINA	C-RUM
0	**0**	**0**	0	0.16	0.13
0	**0**	**0**	1	0.16	0.13
0	**0**	**1**	0	0.16	0.48
0	**0**	**1**	1	0.16	0.48
0	**1**	**0**	0	0.16	0.19
0	**1**	**0**	1	0.16	0.19
0	1	1	0	0.55	0.58
0	1	1	1	0.55	0.58
1	0	0	0	0.16	0.13
1	0	0	1	0.16	0.13
1	0	1	0	0.16	0.48
1	0	1	1	0.16	0.48
1	1	0	0	0.16	0.19
1	1	0	1	0.16	0.19
1	1	1	0	0.55	0.58
1	1	1	1	0.55	0.58

Latent classes with identical probabilities are shown in identical shades of grey

As can be shown in Fig. 4.2, these response probabilities were computed as follows. The response probability for students with different attribute profiles under the DINA model for item 11 is

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.69\right)}{1+\exp\left(-1.69\right)}=0.16,$$

for those who have not mastered any or only one of the two measured attributes, while the response probability for those who have mastered both measured attributes is

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.69+1.91\right)}{1+\exp\left(-1.69+1.91\right)}=0.55.$$

The response probability for students with different attribute profiles under the C-RUM model for item 11 is

$$P(Y_{11}=1)=\frac{\exp(-1.90)}{1+\exp(-1.90)}=0.13.,$$

for those who have not mastered either measured attribute,

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.90+0.43\right)}{1+\exp\left(-1.90+0.43\right)}=0.19,$$

for those who have mastered only one M/D,

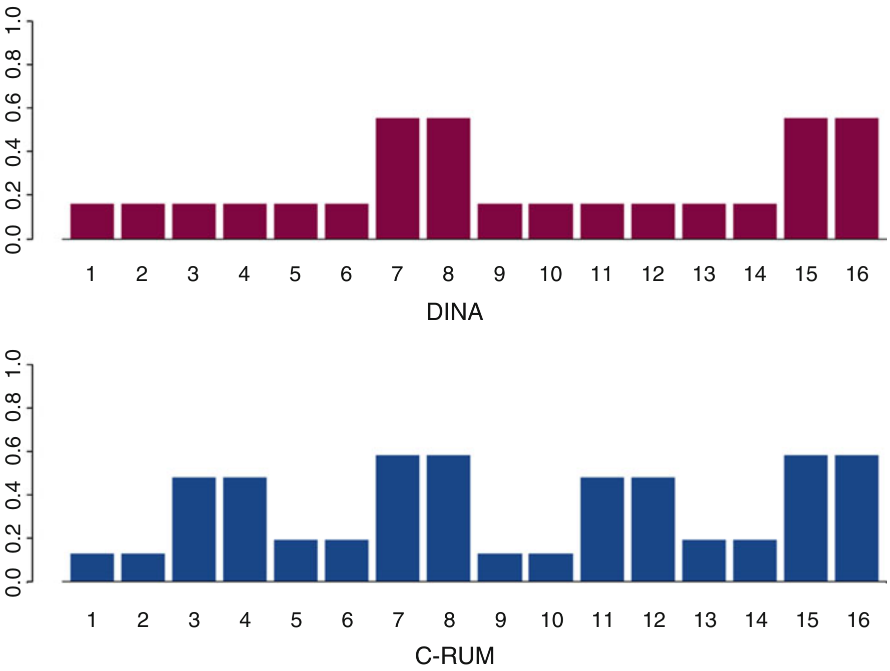


Fig. 4.2 Probability of a correct answer from each attribute profile for item 11

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.90+1.8\right)}{1+\exp\left(-1.90+1.8\right)}=0.48,$$

for those who have mastered only model and for those who have mastered both measured attributes,

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.90+0.43+1.8\right)}{1+\exp\left(-1.90+0.43+1.8\right)}=0.58.$$

These probability computations illustrate nicely how the C-RUM allows for a finer differentiation between students with different attribute profiles than the DINA model in terms of their resulting response probabilities. It is also worth noting that, for both models, the response probabilities for students who have not mastered any attributes are non-zero because the estimates of the intercept parameters are non-zero.

4.4.3 Reporting Attribute Profiles for Groups of Students

The primary purpose of DCMs is to classify students into one of a number of prespecified attribute profiles that correspond to sequences of mastery states on the attributes measured by the diagnostic assessment. Table 4.6 and Fig. 4.3 illustrate

Table 4.6 Distribution of attribute profiles

A/S	M/D	Model	Units	Proportion (%)
0	0	0	0	30.4
0	0	0	1	6.2
0	0	1	1	3.4
0	1	1	1	2.7
1	0	0	0	5.7
1	0	1	0	4.4
1	1	0	0	11.2
1	1	1	0	6.8
1	1	1	1	27.7

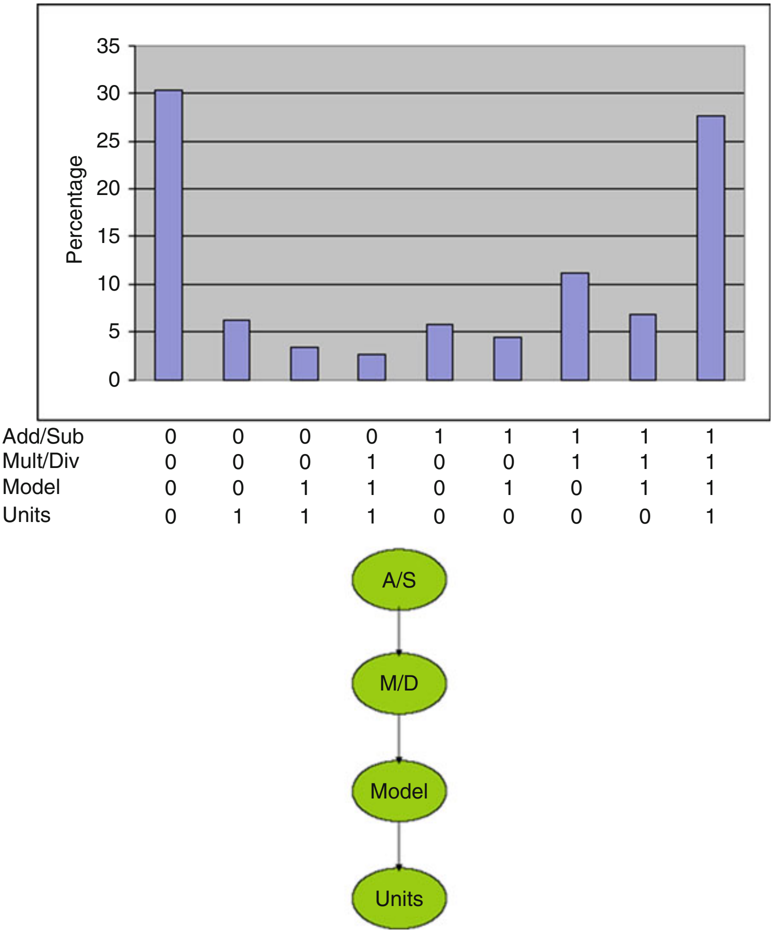


Fig. 4.3 Attribute profiles in sample (*left*) and inferred relationship among attributes (*right*)

how one could display the distribution of attribute profiles for the DMA in our example. Note that with four attributes that are defined in terms of mastery and nonmastery, there exist a total of 16 possible attribute profiles; however, empirically, only nine attribute profiles were populated for these data. Figure 4.2 clearly shows that students predominantly belonged to the two attribute profiles that reflected the lack of mastery of all attributes (30%) and the mastery of all attributes (28%). Moreover, 11% of students were classified as having mastered the first two attributes (A/S and M/D), and about 7% of students were classified as having mastered the first three attributes.

These results gently suggest what is known in the literature as a *linear attribute hierarchy* where the basic arithmetic skills (addition, subtraction, multiplication, division) seem to be mastered before the modeling and unit knowledge skills. However, it needs to be remembered that such inferences are tentative at best because (a) the current data are cross-sectional and not longitudinal in nature, making developmental claims inappropriate, (b) several attribute patterns have similarly low membership probabilities associated with them, and (c) no additional validation results are presented here.

The item parameters and distribution of attribute profiles can be interesting for those who are in charge of test development and require summative statements of students' proficiencies in the assessment of learning sense, while reporting about each student's attribute profile may be more useful for teachers, students, and parents to support assessment for learning.

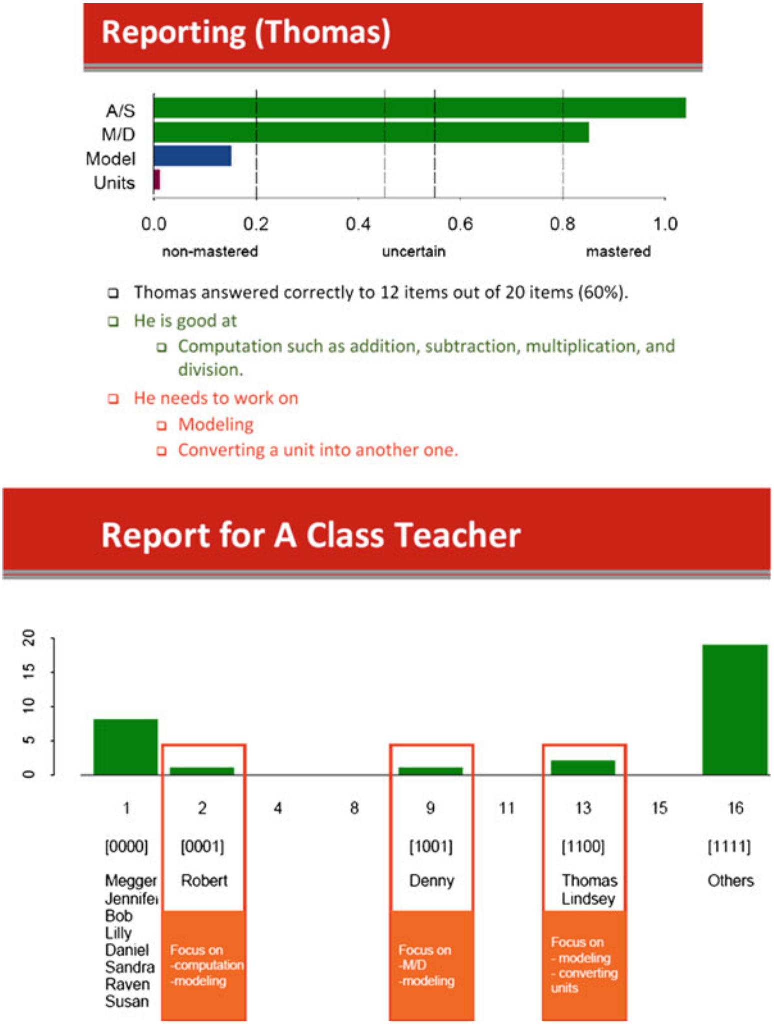
4.4.4 Reporting Attribute Profiles for Individual Students

To illustrate how report cards for individual students could be constructed, we show here the attribute profiles for selected students in Table 4.7. First, for each student, each column indicates the probability that a student should be classified as having each of the nine empirically observed attribute profiles, while the last four columns show the probabilities that each student possesses each of the four attributes that are measured by the test separately. For example, the first student is classified as having mastered attributes A/S and M/D but neither model nor units. This can be seen in the high probabilities of mastery for the first two attributes, which are 0.92 and 0.85, respectively, and the low probabilities of mastery for the last two attributes, which are.15 and.01, respectively. It can also be seen in the fact that his or her probability for the attribute profile [1,1,0,0] is considerably higher at.72 than the probability for any of the other eight attribute profiles.

At the same time, note how there can be challenges in reliably classifying individual students. The second student has a probability of mastery of .55 for the first attribute but is nevertheless classified as having mastered none of the attributes in the profile with a probability of .33. This probability is rather low, however, compared to the highest probability for the first, third, and fourth students and is relatively close to the probability for the attribute profile where only the first attribute

Table 4.7 Sample probabilities for attribute profiles and individual attributes

	Attribute profiles									Attributes			
ID	[0,0,0,0]	[0,0,0,1]	[0,0,1,1]	[0,1,1,1]	[1,0,0,0]	[1,0,1,0]	[1,1,0,0]	[1,1,1,0]	[1,1,1,1]	A/S	M/D	Model	Units
1	.07	.00	.00	.00	.06	.02	**.72**	.12	.00	.92	.85	.15	.01
2	**.33**	.08	.04	.01	.20	.13	.12	.05	.05	.55	.22	.28	.18
3	.00	.00	.01	.00	.02	.04	.05	.07	**.78**	.98	.91	.93	.82
4	.02	.00	.00	.00	.06	.02	**.75**	.13	.02	.98	.90	.17	.02



Reporting (Thomas)

- Thomas answered correctly to 12 items out of 20 items (60%).
- He is good at
  - Computation such as addition, subtraction, multiplication, and division.
- He needs to work on
  - Modeling
  - Converting a unit into another one.

Report for A Class Teacher

Fig. 4.4 Exemplary report card for each student (*top*) and for each class (*bottom*)

is mastered. In practice, it would not be advisable to use this student's classification for high-stakes decision-making, but it may still be useful to suggest to the student additional practice on all attributes with a particular emphasis on the last three.

Based on the classification probabilities shown in Table 4.7, one can create *diagnostic report cards* for each individual student; Fig. 4.4 shows a sample report card for the first student and a class, respectively.

This card shows a total score that expresses how well a student, fictitiously named Thomas, did on the assessment overall and also his mastery states for each attribute that can inform him of his strengths and weaknesses in particular areas if he is taught how to read this information well.

4.5 Conclusions

Developments in the areas of diagnostic assessment design, from a procedural perspective, and DCMs, from a statistical perspective, have the potential to lead to well-aligned large-scale diagnostic assessment systems that can yield more fine-tuned and more instructionally relevant information about students' strengths and weaknesses. In particular, this can be useful as assessment for learning as well as assessment as learning. Nevertheless, it is important to note a variety of caveats.

Substantively, what is crucially needed is a focus on long-term investigations of student progress similar to innovative work in performance-based science assessment (e.g., Thadani et al. 2009). Since education is an ongoing process in class and monitoring students' growth is one of the primary tasks of teachers, diagnostic assessment needs to be carried out with a longitudinal perspective of an assessment-intervention cycle.

Statistically, because of the complexity of the desired diagnostic inferences and the resulting parametric complexity of DCMs, the design requirements for diagnostic assessments are high. On the one hand, it is crucial that every effort be put into place to ensure that calibrations of resulting response data yield reliable profiles on multiple attributes (i.e., separable statistical dimensions). This requires longer assessments in general because sufficient information is needed for each attribute to achieve a reliable statistical classification with DCMs. However, the amount of required statistical information is somewhat smaller than when traditional models from multidimensional IRT or FA are used due to the discrete nature of classifications. On the other hand, this requires data from hundreds or thousands of students per assessment item because item parameters need to be estimated reliably in preoperational settings. Once diagnostic assessments have been calibrated with DCMs, however, it is much easier to score future generations of students with these assessments.

In the end, DCMs are just statistical tools that serve a larger purpose of creating a defensible evidence-based assessment narrative about students. Since the specification of DCMs is still relatively tedious, a wider implementation of these models will probably also not take place unless more user-friendly software is made available. We also want to underscore that they are also not the only models that can be used for diagnostic assessment purposes as the special issue of the *Journal of Educational Measurement* in 2007 demonstrated. For example, multidimensional models from IRT (e.g., Reckase 2009) or FA (e.g., McDonald 2009), as well as cluster analysis methods (e.g., Gan et al. 2007; Steinley 2006), may provide reasonable alternatives even though they result in multiple continuous scales rather than discrete attribute profiles. IRT and FA models in particular have been in use much longer than DCMs and are, thus, generally more strongly trusted by interdisciplinary specialists. Cluster analysis models have a similarly long history in the social and behavioral sciences and are computationally more efficient than DCMs. Thus, they represent attractive modeling alternatives for day-to-day implementations of diagnostic assessments (see Nugent et al. 2009, 2010).

Acknowledgements We would like to thank Olga Kunina-Habenicht for giving us access to the data that was used for the example in the last section in this chapter. The work of Dr. Kunina-Habenicht, including the design, implementation, and analysis, was funded, in part, by grant number RU-424/3–1 from the *German Research Foundation* (*Deutsche Forschungsgemeinschaft, DFG*) in the *Priority Research Program* entitled "Models of Competencies."

References

Agresti, A. (2010). *Categorical data analysis* (2nd ed.). New York: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157.

Choi, H-. J., Templin, J. L., Cohen, A. S., & Atwood, C. H., (2010, April). *The impact of model misspecification on estimation accuracy in diagnostic classification models (DCMs)*. Paper presented at the annual meeting of the National Council for Measurement and Education, Denver, CO.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Pacific Grove: Wadsworth.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 979–1030). Amsterdam: Elsevier.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380–396.

Frey, A., Hartig, J., & Rupp, A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*, 39–53.

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: Theory, algorithms, and applications*. Alexandria: American Statistical Association.

Gierl, M. J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT* (Research Rep. No. 2005–2011). New York: College Examination Board.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Kunina-Habenicht, O., Rupp, A., & Wilhelm, O. (2010, May). *Modelling the latent structure of a diagnostic mathematics assessment within a general log-linear modelling framework*. Presented at the annual meeting of the National Council for Measurement in Education (NCME), Denver, Colorado.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*, 59–81.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*, 64–70.

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press.

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. New York: Cambridge University Press.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4*, 333–369.

Linn, R. L. (1986). Testing and assessment in education: Policy issues. *American Psychologist, 41*, 1153–1160.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: Erlbaum.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Mislevy, R. J., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Rep. 9). Menlo Park: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Technical Rep. 632). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Mok, M. M. C. (2010). *Self-directed learning oriented assessment: Assessment that informs learning & empowers the learner*. Hong Kong: Pace Publications Ltd.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus* (Version 6) [Computer software]. Los Angeles: Muthén & Muthén.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Nugent, R., Ayers, E., & Dean, N. (2009). Conditional subspace clustering of skill mastery: Identifying skills that separate students. In *Proceedings from the 2nd international conference on educational data mining* (pp. 101–110). Retrieved July 19, 2010, from www.educationaldatamining.org/EDM2009/

Nugent, R., Dean, N., & Ayers, E. (2010). Skill set profile clustering: The empty K-means algorithm with automatic specification of starting cluster centers. In *Proceedings from the 3rd international conference on educational data mining* (pp. 151–160). Retrieved July 19, 2010, from http://educationaldatamining.org/EDM2010/

O'Reilly, T. P., Sheehan, K. M., & Bauer, M. I. (2008, March). *Cognitively based assessments of, for, and as learning: Bridging the gap between research and practice*. Presented at the annual meeting of the American Educational Research Association, New York.

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). Cambridge: Cambridge University Press.

Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.

Rupp, A. A., Levy, R., DiCerbo, K., Sweet, S., et al. (in press). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. Journal of Educational Data Mining.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Steinley, D. (2006). *K*-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*, 1–34.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale: Erlbaum.

Templin, J. L. (2004). *Generalized linear mixed proficiency models*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychology Methods, 11*(3), 287–305.

Templin, J. L., Henson, R. A., & Douglas, J. (2011). *General theory and estimation of cognitive diagnosis models: Using Mplus to derive model estimates*.

Thadani, V., Stevens, R. H., & Tao, A. (2009). Measuring complex features of science instruction: Developing tools to investigate the link between teaching and learning. *The Journal of the Learning Sciences, 18*, 285–322.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

von Davier, M. (2005). *A general diagnostic model applied to language testing data.* (ETS Research Rep. No. RR-05–16). Princeton: Educational Testing Service.

von Davier, M. (2006). Multidimensional latent trait modelling (MDLTM) [Software program]. Princeton: Educational Testing Service.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., DiCerbo, K. E., Crawford, A., Choi, Y., & Behrens, J. (2009, June). *A Bayes net approach to modeling learning progressions and task performances*. Paper presented at the Learning Progressions in Science (LeaPS) conference, Iowa City, IA.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport: American Council on Education.

Chapter 5
Application of the DINA Model Framework to Enhance Assessment and Learning

Jimmy de la Torre

5.1 Introduction

Assessment should go beyond ascertaining and auditing the status of student learning – it should also be tapped as a tool for improving learning and performance (Stiggins 2002; Wiggins 1998). However, because most large-scale educational assessments are based on a framework that assumes a single underlying trait (i.e., proficiency in a domain), these assessments provide overall scores that are primarily informative in scaling and rank-ordering students along a unidimensional continuum. However, because single, overall scores are coarser by nature, they are of limited value in practical instructional settings in that they lack the finer-grained information necessary for diagnosing students' specific strengths and weaknesses and for informing teachers how classroom instruction can be adjusted to better target students' needs. Assessments that can help inform classroom instruction and learning must provide information that is "interpretative, diagnostic, highly informative, and potentially prescriptive" (Pellegrino et al. 1999, p. 335). With this in mind, this chapter will introduce an alternative psychometric framework (i.e., cognitive diagnosis modeling) that can serve as a basis for developing and analyzing educational assessments that can provide diagnostically relevant scores. In this chapter, the cognitive diagnostic framework will be contrasted with the unidimensional item response theory (IRT) framework. One cognitive diagnosis model (CDM), the *deterministic, input, noisy* "and" *gate* (DINA) model, will be highlighted. In addition to the original formulation, several extensions of the DINA model and their corresponding assumptions will be discussed. The different approaches and models will be illustrated using a mixed fraction subtraction problem. Details pertaining to

J. de la Torre (✉)
Graduate School of Education, Rutgers,
The State University of New Jersey, Newark, USA
e-mail: j.delatorre@rutgers.edu

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_5, © Springer Science+Business Media Dordrecht 2013

the model specification, estimation, and other technical issues will be briefly covered. This chapter concludes with a discussion of the relevant issues pertaining to diagnosis modeling and classification.

5.2 IRT vs. CDM

In contrasting the IRT and CDM frameworks, we start by examining the mixed fraction problem: $2\frac{4}{12}-\frac{7}{12}$. In IRT, it is assumed that a single latent trait θ (i.e., mixed fraction subtraction proficiency) underlies a student's performance in this domain. Students with higher proficiencies are expected to have higher probabilities of success on this item and the remaining items on the test. Figure 5.1 shows two hypothetical students, 1 and 2, with proficiencies $\theta_1=-0.8$ and $\theta_1=-0.4$. Using a hypothetical item characteristic curve, it shows that $P(\theta)$, the probability of answering the item correctly, is 0.2 for Student 1, whereas $P(\theta)$ is 0.6 for Student 2.

In CDM, instead of a *single*, *continuous* latent trait, performance in a domain is assumed to be a function of *multiple, discrete* latent traits generically referred to as attributes. The generic term *attribute* can encompass skill, thinking process, and problem-solving strategy. An attribute vector is typically denoted as α and has the elements $\alpha_1,\alpha_2,\ldots,\alpha_K$, where K is the total number of attributes. For the mixed fraction subtraction domain, $K=5$ *attributes* have been identified, namely, (1) borrowing from the whole number to fraction, (2) performing basic fraction subtraction,

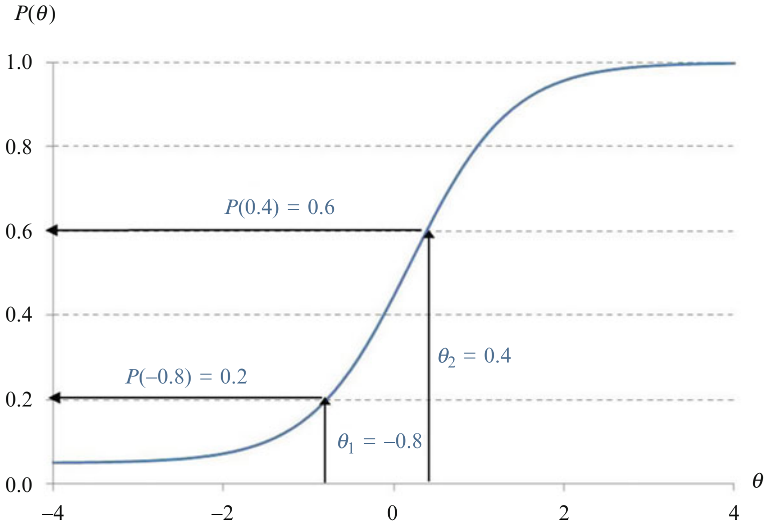


Fig. 5.1 A hypothetical item characteristic curve and probabilities of success of two students

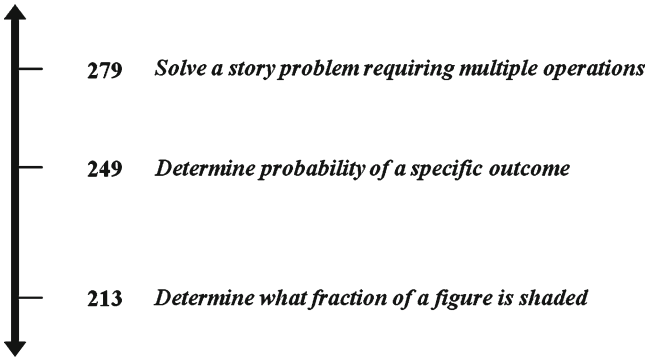

279 Solve a story problem requiring multiple operations
249 Determine probability of a specific outcome
213 Determine what fraction of a figure is shaded

Fig. 5.2 Descriptions of NAEP 4th grade mathematics problems at three scale points

(3) reducing/simplifying, (4) separating whole from fraction, and (5) converting whole to fraction (Tatsuoka 1990; Mislevy 1995). A successful performance on the problem $2\frac{4}{12}-\frac{7}{12}$ requires a series of successful implementations of the first three attributes (i.e., α_1, α_2 and α_3), the required attributes for the problem. Instead of a single score, CDM-based assessments generate and report a score profile of length K for each student detailing which attributes the students have mastered or not mastered. The finer-grained and interpretative natures of attributes make them more suitable for diagnostic and prescriptive purposes.

However, as noted by de la Torre and Karelitz (2009), unidimensional IRT models have also been used for diagnostic ends. In IRT, both the student proficiency and item difficulty can be located on the same scale. Using an item map, exemplars or problem descriptions are associated with the different points of the proficiency continuum to allow students and teachers to identify the types of problems students of differing proficiencies can do. Given in Fig. 5.2 are three scale scores and descriptions of three corresponding problems from the 2007 National Assessment of Educational Progress (NAEP) 4th grade mathematics assessment. Based on the NAEP item map for the 2007 mathematics assessments (Lee et al. 2007), the three scores represent the points along the continuum where problems closest to the cut points of the basic (213), proficient (249), and advanced (282) proficiency levels can be found. In addition to the expectations of what type of problems students at different proficiency levels should be able to do, the item map also indicates the types of problems a student with a particular proficiency score has and has not mastered.

Although exemplars and problem descriptions in item maps can provide richer information, their diagnostic applications can also be challenging in that the location of an item (i.e., its difficulty) is a coarse summary of the different features that make an item easy or difficult. That is, unless the specific item features are teased apart, it would be unclear which aspects of the domain a student is struggling with. Although teasing out a problem can provide additional information, this practice

What is the closest approximation of $\sqrt{30}$?

Set A	Set B
(a) 4.5	(a) 5.4
(b) 5.0	(b) 5.5
(c) 5.5	(c) 5.6
(d) 6.0	(d) 5.7

Fig. 5.3 A multiple-choice problem with a single stem and two sets of options of different difficulties

can be counterproductive if attention is given to idiosyncratic features of the item rather than the features that it shares with other items. Without consistent information across several items, isolated item features are too unreliable to be used as a source of diagnostic information. However, it is not clear whether students and teachers are in a position to identify features that cut across multiple items solely based on exemplars. It should also be noted that, in addition to the specific topic being examined, an item's difficulty can also be affected by how the problem is posed. Given in Fig. 5.3 is a problem on approximating the root of a number. Two sets of options, A and B, are provided for the same stem. By examining the option sets, it can be surmised with reasonable certainty that using Set B will result in a problem with a higher level of difficulty. Thus, focusing on the superficial features of a problem rather than the problem in its entirety can also be highly misleading.

In some applications, it is reasonable to assume that attributes have a certain cognitive structure. Attribute structures can assume various forms, and one such structure is the hierarchical linear structure (Leighton and Gierl 2007a; Leighton et al. 2004). In the hierarchical linear structure, mastery of simpler attributes is a prerequisite to the mastery of more complex attributes. De la Torre and Karelitz (2009) claim that a correspondence between unidimensional IRT and CDM can be established when the attributes have a hierarchical linear structure. However, such correspondence may not exist in situations where a more complex cognitive structure (e.g., divergent, convergent) is involved. For this reason, unidimensional IRT models have limited utility, if at all, in applications where diagnostic information about multiple, disparate dimensions are of interest. But as de la Torre and Karelitz noted, even in situations where IRT and CDM can be considered interchangeable, using a model that corresponds to the underlying process can produce better results.

5.3 The DINA Model

A psychometric model is needed to relate the observable assessment performance to the posited latent traits, and a CDM would be appropriate if a model of the students' cognitive processes exists, and inferences are to be made across several dimensions. One such model is the *deterministic inputs, noisy "and"* gate (DINA; Junker and Sijtsma 2001) model. The DINA model can be considered the simplest of the existing

CDMs that is appropriate for educational assessment data. Like most CDMs, applying the DINA model requires a binary matrix called the Q-matrix (Tatsuoka 1983). The Q-matrix, which has *J* rows corresponding to the number of items and *K* columns corresponding to the number of attributes, indicates which attributes are needed for each item. For example, the row corresponding to the problem $2\frac{4}{12}-\frac{7}{12}$ should have three 1s followed by two 0s to indicate that the first three attributes are needed to answer the item correctly, whereas the last two attributes are irrelevant with respect to this item. In addition to identifying the relevant attributes, CDM applications that involve the Q-matrix also presuppose that a mapping between the items and the attributes they measure can be established. When the attribute specifications are used as a blueprint for test construction, the Q-matrix can play an important role in developing cognitively diagnostic assessments (Leighton et al. 2004; Junker 1999).

There are two distinct components to the DINA model. The first component pertains to the *deterministic* aspect of the model. Based on the Q-matrix attribute specifications, the DINA model creates two distinct groups that vary item by item. One group, η_1, consists of individuals who possess all the required attributes for the item, and another group, η_0, consists of individuals who lack one or more of the required attributes. The process is deemed deterministic in that, given an attribute specification for an item, individuals are always assigned to the same groups. The "and" part of the model arises from the group assignment process, which is conjunctive in nature in that all the required attributes need to be simultaneously present for an individual to be classified in group η_1. Individuals in group η_1 are expected to answer the item correctly, whereas individuals in group η_0 are not. However, the second component of the model, the stochastic (i.e., *noisy*) aspect, allows for the possibility that individuals in group η_1 can slip and answer the item incorrectly and that individual in group η_0 can guess the correct answer to the problem. The amount of noise in the model is determined by the size of slips and guesses, which represent the two parameters of the DINA model. Incidentally, the sizes of the slip and guessing parameters can be used to determine the discrimination of an item (de la Torre 2008). Finally, the simplicity of the DINA model stems from the fact that the model has only two parameters per item regardless of the number of prescribed attributes or the total number of attributes. Given in Fig. 5.4 is a hypothetical item requiring three attributes. From the three required attributes, eight attribute combinations can be distinguished. Only one of these attribute combinations (i.e., 111) will be classified in group η_1; the rest will be classified in group η_0. Individuals in group η_1 have 0.90 probability of answering the item correctly (i.e., they will slip 10% of the time). In contrast, individuals in group η_0 will be able to guess and answer the item correctly 20% of the time. The difference in the probabilities of success between the two group, $P(\eta_1)-P(\eta_0)=0.75$, indicates that the item is highly discriminating in that a correct response is more likely to have come from individuals who have mastered the three required attributes, whereas an incorrect response from individuals who have not mastered one or more of the required attributes.

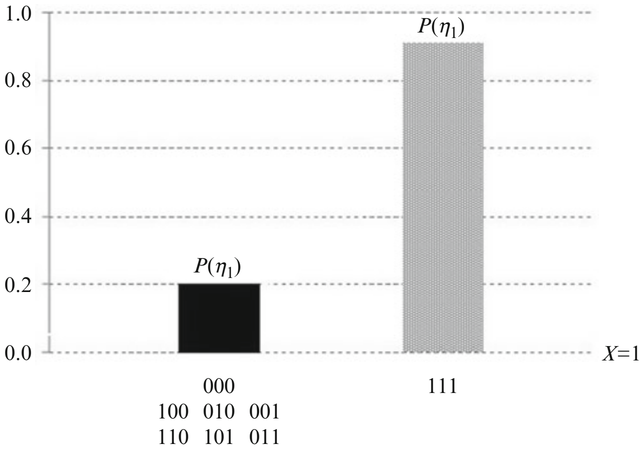


Fig. 5.4 DINA model probabilities of success for a hypothetical item requiring three attributes

5.4 Extensions of the DINA Models

As can be seen from Fig. 5.4, the DINA model assumes that individuals in group η_0 have the same probability of success regardless of the number of attribute deficiencies. Thus, someone who has not mastered any of the required attribute is as likely to succeed on the item as someone who has mastered all but one of the required attributes. In applications where student responses are scored as right when all the steps are correctly applied, this can be considered a reasonable assumption. However, such an assumption may be deemed unreasonable when the probability of guessing can vary as a function of the subset of required attributes mastered, or when partial credit can be given. Regardless, scoring responses as either right or wrong when additional information is available is suboptimal.

5.4.1 MC-DINA Model

In most applications of CDMs, responses from multiple-choice (MC) assessments are treated as dichotomous data (e.g., de la Torre 2006; Tatsuoka et al. 2004). Because this approach ignores the diagnostic insights about student difficulties and alternative conceptions that can be found in the distractors, it is considered suboptimal (Haertel and Wiley 1993; Nitko 2001; Sadler 1998). To address this concern, de la Torre (2009a) proposed a CDM framework that allows MC data to be used more optimally for diagnostic purposes. The framework includes a component prescribing how MC distractors must be constructed, and the MC-DINA model, a CDM

$$2\frac{4}{12}-\frac{7}{12}=$$

(a) $2\frac{3}{12}$	(b) $2\frac{1}{4}$	(c) $1\frac{9}{12}$	(d) $1\frac{3}{4}$
α_2	α_2, α_3	α_1, α_2	α_1, α_2 α_3

Fig. 5.5 An example of a mixed fraction subtraction problem with coded options and required attributes associated with the each option

	DINA Model	
	Group η_0 : Incorrect	Group η_1 : Correct
MC-DINA Model	(a) 010	(d) 111
	(b) 011	
	(c) 110	
	000, 001, 100, 100	

Fig. 5.6 Attribute pattern group membership based on the DINA and MC-DINA models

specifically designed to capitalize on the additional information generated by cognitively coding the MC options. Figure 5.5 provides an example of a cognitive-based MC problem. In addition to the key or correct response (i.e., "d"), which requires the first three attributes, the distractors are also coded to reflect the options students with different mastery patterns are likely to choose. For example, students who have mastered "α_2 basic fraction subtraction," but not "α_1 borrowing from the whole number to fraction" and "α_3 reducing/simplifying," are expected to choose option "a." By designing and coding the distractors to correspond to specific attribute patterns, the choice of distractors reveals not only what the students know but also what they do not know. Consequently, data collected from such an assessment contain more diagnostically relevant information.

Instead of looking at the DINA model in terms of the slip and guessing parameters, the model can be alternatively viewed as a CDM that associates the groups η_0 and η_1 with the incorrect and correct responses, respectively. From this perspective, the MC-DINA model extends the DINA by further differentiating individuals in group η_0 and associating them with the different coded distractors. Figure 5.6 shows that the cognitively based distractors in Fig. 5.5 allow for the individuals in group η_0 to be split into four groups, thus, creating a total of five, instead of just two groups. The additional distinctions created within group η_0 allow the MC-DINA model to better classify students according to which attributes they have and have not mastered. In a simulation study carried out by de la Torre (2009a), it was found that, by cognitively coding the distractors and analyzing the data using the MC-DINA model, attribute and attribute-vector classifications can be improved by at least 6% and 20%, respectively, relative to the classifications based on the DINA model analysis (i.e., analysis of dichotomized data).

In practice, because coding all the distractors may not always be feasible, the MC-DINA model was designed to also handle data where only a subset of the distractors are coded. Moreover, de la Torre (2009a) noted that even if only the key is coded, but the distractors are kept distinct from each other (i.e., they are not all assigned a score of zero), the MC-DINA model remains applicable. In such applications, the MC-DINA model is equivalent to the CDM for nominal data proposed by Templin et al. (2008). Lastly, the MC-DINA model reduces to the DINA model if no distinctions are made between the distractors.

5.4.2 PC-DINA Model

From the cognitive perspective, MC items are often viewed as reflecting comparatively low-level cognitive processing, whereas constructed response items are more likely to evoke higher-level processing. Constructed response format, either of the short or extended type, can reduce the probability that students will correctly guess the answer to a problem (Nitko 2001). Depending on the scoring key or rubric, student responses can be scored on a scale from 0 to M_j, where M_j is the maximum score that a student can receive for Item *j*. When $M_j = 1$, the item is scored as either right or wrong; when $M_j > 1$, students can receive partial credit for their answers that demonstrate partial or incomplete knowledge. Thus, instead of reducing the responses into only two categories, partial credit creates additional nuances in the responses. These additional nuances can contain extra information that can potentially result in better classification of the students based on their mastery and nonmastery of the attributes.

Although typically associated with constructed response items, partial credit can also be used with MC items. Specifically, instead of assigning a score of zero to all the distractors, some of them can be given partial credit. Although assigning partial credit creates distinction among the distractors, this practice is not the same as the cognitive-coding described above in that the distractors are not associated with any particular attribute patterns.

To accommodate scoring that involves $M_j \geq 1$ in the context of cognitive diagnosis modeling, de la Torre (2010) proposed a generalization of the DINA, the *partial-credit DINA* (PC-DINA) model. As with the DINA model, the PC-DINA model also categorizes the individuals into groups η_0 and η_1. In its conventional formulation, the DINA model gives the probability of obtaining the correct response conditional on the individual's group membership, as in, $P\left(X = 1 \middle| \eta\right)$. Implicit in the model is the complementary conditional probability of obtaining an incorrect response,

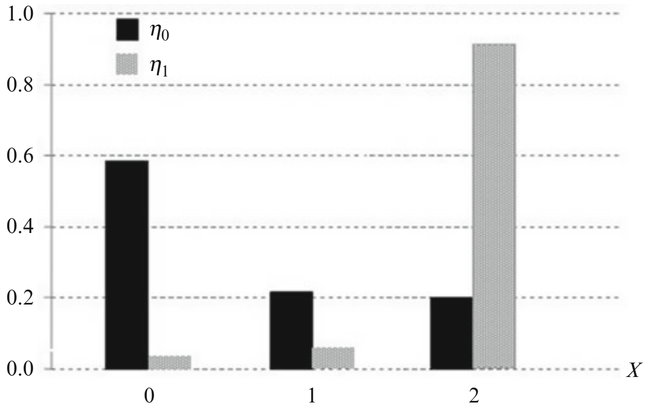


Fig. 5.7 Conditional probabilities based on the PC-DINA model when $M_j = 2$

which trivially is $P(X=0|\eta) = 1 - P(X=1|\eta)$. Viewed from such a perspective, the PC-DINA model is a straightforward extension of the DINA model that gives the conditional probabilities associated with the $M_j + 1$ score categories. Given in Fig. 5.7 are the conditional probability distributions for groups η_0 and η_1 based on the PC-DINA model for a hypothetical item with $M_{j^*} = 2$. The conditional probabilities indicate that individuals in group η_1 have a high probability of obtaining a score of 2 (about 0.9) and a very low probability of obtaining a score of 0 or 1 (a combined probability of about 0.1). In contrast, individuals in group η_0 have a relatively low probability of obtaining a score of 2 (about 0.2 only) and a moderately high probability obtaining a score of 0 or 1 (a combined total of about 0.8). When only the maximum score (i.e., 2) is deemed correct, the conditional probabilities for a score of 2 are identical to the DINA model probabilities given in Fig. 5.4.

Implementing partial-credit scoring is more resource-intensive and time-consuming. It requires that appropriate investments be made to produce reliable scoring keys and to train human scorers to be more objective. In addition to potential unreliability, involving human scorers also means that a longer lag between the test administration and score reporting can be expected. Given the higher cost of implementation, what additional diagnostic information can be gained from using partial-credit scoring? In a simulation study conducted by de la Torre (2010) where M_j was fixed at 2, he compared the PC-DINA model attribute classification accuracy based on polytomous data against that of the DINA model based on dichotomized data. The results of this study indicate that the improvement in the classification accuracy can range from 0.01 to 0.04 at the attribute level and from 0.05 to 0.010 at the attribute-vector level, where larger differences occurred when the items are less discriminating. These results suggest that, assuming partial-credit scoring can be done reliably, using more scoring categories beyond right/wrong can improve the diagnostic usefulness of assessments.

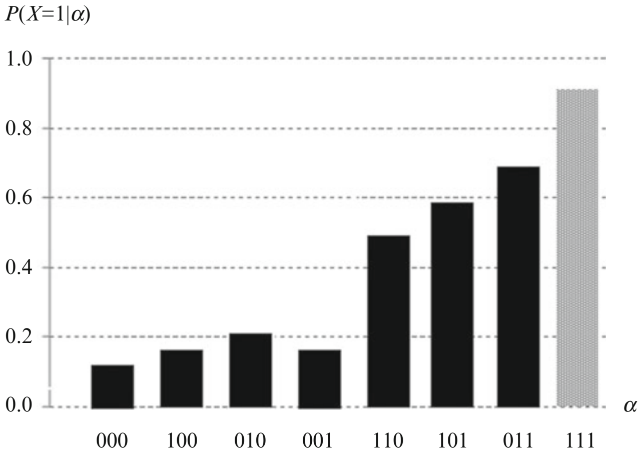


Fig. 5.8 Probabilities of success associated with different attribute patterns

5.4.3 Generalized DINA Model

As noted earlier, the conjunctive component of the DINA model assumes that individuals in group η_0, irrespective of the number and nature of their deficiencies, have the same probability of answering an item correctly. However, there are applications where this might be too strong an assumption. For example, given several options, individuals who have most of the required attributes for an item might have a higher chance of guessing the correct answer compared to, say, those who lack all the required attributes. Figure 5.8 gives the probabilities of success for a hypothetical three-attribute item associated with different attribute patterns. This figure illustrates that, although the highest probability of success is associated with the pattern where all the required attributes are present, individuals who lack at least one of the required attributes do not have the same probability of success. Specifically, individuals who lack only one of the required attributes are expected to outperform those who lack two or more of the required attributes on this item.

Relaxing the conjunctive assumption of the DINA model allows for the possibility that individuals who have not mastered all the required attributes for an item may have varying probabilities of success. This variation in the success probabilities is what is captured by the *generalized DINA* (G-DINA; de la Torre 2011) model. Given the set of required attributes for an item, the G-DINA model assigns a success probability for each of the possible combinations of attribute mastery and nonmastery. With finer distinctions between the success probabilities, the specific differential contribution of mastering a single attribute to item performance and the interaction effect due to mastering several attributes at the same time can be examined. For example, the item $\frac{11}{8}-\frac{1}{8}$ requires attributes α_2, performing basic fraction subtraction, and α_3, reducing/simplifying. The G-DINA model estimates indicate that students who have not mastered any of the required attributes have a .011 probability of answering the item correctly; mastering α_3 only does not improve this success probability, but mastering α_2 only increases the success probability to 0.59; however, for an optimal success rate of 0.97, both the required attributes need to be mastered. The example indicates that it might be advantageous for students to master learning basic fraction subtraction before learning reduction or simplification, but both attributes need to be eventually mastered if students are to do well on this type of problem.

Although motivated by the DINA model, de la Torre (2011) has shown that the G-DINA model subsumes a wider class of CDMs that include the *deterministic input, noisy* "or" *gate model* (Templin and Henson 2006), *reduced reparametrized unified model* (Hartz 2002), *linear logistic model* (Hagenaars 1990, 1993; Maris 1999), and the *additive CDM*. More than a model, the G-DINA is a framework consisting of a component for estimating the models it subsumes and a component that tests whether reduced or constrained models can be used in place of the general model. By performing the estimation and testing of reduced models at the item level (i.e., one item at a time), these operations can be carried out more efficiently. Taken together, with the general formulation of the G-DINA model, it is possible to conduct cognitive diagnosis modeling without making an a priori commitment to a particular CDM; with its estimation and testing components, it is possible for multiple, possibly disparate CDMs to be used within the same assessment. Another advantage of using the G-DINA model was demonstrated by de la Torre and Chiu (2010). They used the G-DINA model to propose a general procedure for empirically validating the attribute specifications in the Q-matrix. This procedure extends the method that previously applies only to the DINA model (de la Torre 2008). In addition to a mathematical proof, their simulation study, which involved data generated using different CDMs and various types and numbers of Q-matrix misspecifications, shows that the method can identify incorrectly specified attributes from those that have been correctly specified.

5.4.4 Other Extensions of the DINA Model

In some applications, a problem can be solved in more than one way. For example, Mislevy (1996; also de la Torre and Douglas 2008) used two strategies in analyzing mixed fraction subtraction data. In the first strategy, fraction subtraction is carried out using mixed numbers and involves the five mixed fraction subtraction attributes described above. In the second strategy, mixed numbers are first converted to improper fractions before the subtraction operation is performed. In addition to α_2, α_3, and α_5, the second strategy also involves two additional attributes, α_6, converting mixed number to fraction, and α_7, column borrowing in subtraction. Aside from a larger set of attributes, modeling multiple strategies with CDMs also requires that a separate Q-matrix be constructed for each strategy.

To accommodate multiple strategies, de la Torre and Douglas (2008) proposed the *multiple-strategy DINA* (MS-DINA model), which is a straightforward extension of the DINA model. As with the DINA model, the MS-DINA model creates the groups η_0 and η_1, although the groups are defined differently. In the MS-DINA model, group η_1 for an item consists of individuals who have the set of attributes required by at least one of the strategies, whereas group η_0 consists of individuals who do not fully satisfy the attribute requirements of any of the strategies. The MS-DINA model preserves the conjunctive assumption within each strategy in that individuals are expected to be able to solve an item correctly using a particular strategy if and only if they have all the attributes required by that strategy. Like the DINA model, the parameters of the MS-DINA model are probabilities of success for groups η_0 and η_1. Thus, despite the increase in the number of required attributes per item and the total number of attributes being measured, both the DINA and MS-DINA models have the same number of parameters for each item – two. The simplicity of the MS-DINA model is based on the implicit assumption that individuals can switch from one strategy to another and that the application of the different strategies is equally difficult. As such, the MS-DINA model approach differs from the approach used by Mislevy (1996) in that the latter involves mixture modeling where individuals primarily use a single strategy and the probabilities of success can vary across strategies.

The increasing popularity of computer-based assessments has made the large-scale collection of response time or latency, a type of continuous response, more practicable. In conventional educational testing settings, response time has been used as a source of collateral information (van der Linden et al. 2010) to detect aberrant response patterns (van der Linden and van Krimpen-Stoop 2003) and to control differential speededness in computer adaptive testing (van der Linden et al. 1999). De la Torre and Liu (2008) proposed the *continuous DINA* (C-DINA) model partly to take advantage of response time or latency in the context of cognitive diagnostic modeling. Except for the response type (i.e., continuous vs. dichotomous), the C-DINA model is similar to the DINA model, in that, given the attribute specification for an item (i.e., one row of the Q-matrix), both models give the probability distribution of the item response conditional on the group membership. However, the two models are not interchangeable – using dichotomized response when continuous response is appropriate can result in poorer attribute classification, particularly when the items are not as diagnostically informative. For most CDM applications, attribute classification will primarily be based on the correctness of the response. However, when response latency is also available, incorporating it into the modeling process can provide ancillary information that can improve the correct classification rate. Finally, from a more integrative perspective, it might be helpful to note that the PC-DINA model is an intermediate case situated between the DINA and C-DINA models.

5.5 Some Technical Details

5.5.1 Estimation

A CDM links the observable response to the underlying latent attributes and gives the conditional probability of a correct response given an attribute pattern. However, for the latent variable model to be complete, the probability distribution of the attributes needs to also be specified. Given K attributes, there are 2^K attribute patterns corresponding to all the attribute mastery and nonmastery combinations. The saturated model, which represents the most general model for the attribute distribution, imposes no constraints on the structure of the attribute patterns. This model contains $2^K - 1$ number of parameters. In conjunction with the marginalized maximum likelihood (MML) estimation, the saturated model has been used to estimate the parameters of the DINA, MC-DINA, PC-DINA, C-DINA, and G-DINA models (de la Torre 2008, 2009a, b, 2010; de la Torre and Liu 2008). Simulation studies indicate that, for up to a reasonably large K, MML estimation is an efficient and accurate method of estimating these DINA-based model parameters.

The number of attribute patterns 2^K grows exponentially with the number of attributes. Thus, when K is large, using the saturated model to keep track of all the attribute patterns may become too computationally expensive, if not impossible altogether. This is particularly a salient issue with MML estimation because it requires marginalizing (i.e., summing over) the 2^K attribute patterns for each iteration. A simplification of the attribute structure is needed if CDM estimates were to be obtained. One such simplification was proposed by de la Torre and Douglas (2004). In their model, they assume that mastery of the attribute is related to a higher-order, more broadly defined construct, ω, such that those with a higher ω have a greater likelihood of mastering the required attributes. This assumption is reasonable in situations where assessments can be viewed as also tapping some general proficiency to which the attributes are related. By assuming that the elements of α are conditionally independent given ω, the higher-order formulation dramatically reduces the complexity of the saturated model. Instead of 2^K–1 parameters, the higher-order model typically requires as few as $K+1$ parameters. The DINA and MS-DINA models and a constrained case of the R-RUM (i.e., the *noisy input, deterministic* "and" *gate* model; Junker and Sijtsma 2001) have been specified using the higher-order formulation, and the parameters of the models have been estimated using Markov chain Monte Carlo algorithm (de la Torre and Douglas 2004, 2008). Other approaches to simplifying the attribute distribution include assuming that the attributes are independent of each other (de la Torre and Douglas 2004; Maris 1999), resulted from dichotomizing elements of a multivariate normal variable (Hartz 2002), and follow a particular hierarchical structure (Leighton et al. 2004).

5.5.2 Software

At present, no estimation package has been specifically designed to cover all the DINA models presented in this chapter. However, two codes for obtaining the MML estimates of the DINA and G-DINA models based on a saturated attribute distribution are available. The codes are written in Ox (Doornik 2003) and run using the console version of Ox which can be downloaded free of charge for academic research, study, and teaching purposes. In addition to model parameter estimates and standard errors, the codes also provide attribute classification and item- and test-level fit statistics. In their current forms, both codes are limited to analyzing complete data matrices and a maximum of $K = 15$ attributes. A separate Ox code that can be used for Q-matrix validation is also available. This code is based on the general procedure for empirically validating the attribute specifications in the Q-matrix developed by de la Torre and Chiu (2010). All the Ox codes can be made available by contacting the author.

In addition to the specific codes, general statistical packages that can perform latent class analysis such as Latent GOLD (Vermunt and Magidson 2005) and Mplus (Muthén and Muthén 2010) can also be adapted to estimate some of the DINA-based and other CDM models, including those that have a higher-order formulation. However, because these software packages are not specifically designed for cognitive diagnosis modeling purposes, they can be less efficient and possibly constrained in some respects compared to custom-built estimation programs. On the other hand, as noted by Templin et al. (2009), using existing software packages allows users to take advantage of the advanced features that are inherent in these programs. Examples of these features include facility and flexibility in accommodating different response types, handling missing data, incorporating covariates, and dealing with time-series data.

5.6 Discussion and Conclusion

Emphasized in this chapter are different CDMs, particularly those that belong to the DINA model family. However, psychometric models are only a component of cognitive diagnosis modeling. Equally important to highlight, but perhaps elsewhere, is the role of cognitive models. Depending on one's persuasion, cognitive models can range from simply identifying the relevant domain attributes and how they relate to the items to specifying the relationships between these attributes. The DINA models described above do not explicitly require that attribute relationships be specified, but when substantive theories are available, structuring the attributes and eliminating some attribute patterns can greatly reduce the complexity of the models. When attribute structures are undergirded by appropriate cognitive or learning theories, they can improve the accuracy of the item parameter estimation and attribute classification; however, when incorrect attribute structures are imposed, they can lead to highly misleading results (de la Torre et al. 2010). For this reason, the importance of incorporating attribute validation, including Q-matrix validation, as an integral component of cognitive diagnosis modeling cannot be overemphasized.

One of the impediments to fully realizing the diagnostic potential of CDMs is the dearth of assessments that are developed using a cognitive diagnosis framework. More often than not, CDMs are retrofitted to existing assessments that measure a single dominant proficiency. Such assessments are highly unidimensional and can only provide very limited diagnostic information, in that the students' locations along the proficiency continuum can only provide a good summary of what they can and cannot do. If richer inferences spanning several dimensions are of interest, retrofitting

CDMs to unidimensional tests needs to be minimized; instead, investments need to be made to develop diagnostic assessments that are grounded on appropriate cognitive and learning theories. Only in conjunction with appropriately constructed assessments can the use of CDMs be optimized.

With their various formulations, researchers and practitioners have the option to choose from a collection of DINA models. An obvious factor to consider in making this choice is the type of response that will be analyzed. For example, the DINA model is suitable for dichotomous data, whereas the MC-DINA and PC-DINA models are more suitable for polytomous data. However, an appropriate choice of model also depends on other practical considerations. One such consideration is the sample size. More complex models are more informative only to the extent that they can be accurately estimated. Consequently, when sample size is relatively small, simpler models might be a better choice even though the data type warrants a more complex model. Another consideration is how the scores will be used. Diagnostic scores can influence classroom learning and instruction when they are reported in a timely manner. Scoring some of the more complex DINA models (i.e., PC-DINA model) can be time-consuming. Resorting to simpler scoring procedures (e.g., right or wrong) and, hence, simpler models, although suboptimal in a statistical sense, might be necessary if punctual reporting is of paramount importance.

Finally, it should be noted that DINA models presented in this chapter represent only a subset of currently available CDMs. In turn, the use of CDMs is but one of the several approaches in the field of diagnostic modeling and classification. For more detailed treatments of other CDMs and approaches, the readers are referred to some of the most recent works in this area. These works include those by Leighton and Gierl (2007b), Rupp and Templin (2008), Rupp et al. (2010), Tatsuoka (2009), and the Winter 2007 special issue of the *Journal of Educational Measurement*.

Acknowledgment This material is based upon work supported by the National Science Foundation under Grant No. DRL-0744486. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

de la Torre, J. (2006, June). *Skills profile comparisons at the state level: An application and extension of cognitive diagnosis modeling in NAEP*. Paper presented at the international meeting of the Psychometric Society, Montreal, Canada.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343–362.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*, 163–183.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.

de la Torre, J. (2010, July). *The partial-credit DINA model*. Paper presented at the international meeting of the Psychometric Society, Athens, GA.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

de la Torre, J., & Chiu, C. Y. (2010, April). *General empirical method of Q-matrix validation*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

de la Torre, J., & Douglas, J. (2008). Model evaluation and selection in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595–624.

de la Torre, J., & Karelitz, T. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure. *Journal of Educational Measurement, 46*, 450–469.

de la Torre, J., & Liu, Y. (2008, March). *A cognitive diagnosis model for continuous response*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*, 227–249.

Doornik, J. A. (2003). *Object-oriented matrix programming using Ox* (Version 3.1). [Computer software]. London: Timberlake Consultants Press.

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359–384). Hillsdale: Erlbaum.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Newbury Park: Sage.

Hagenaars, J. A. (1993). *Log-linear models with latent variables*. Newbury Park: Sage.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Junker, B. W. (1999, November 30). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*. Paper prepared for the Committee on the Foundations of Assessment, National Research Council.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Lee, J., Grigg, W., & Dion, G. (2007). *The Nation's report card: Mathematics 2007* (NCES 2007–494). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Available online at: http://nces.ed.gov/nationsreportcard/itemmaps/?subj=Mathematics&year=2007&grade=4

Leighton, J. P., & Gierl, M. J. (2007a). *Cognitive diagnostic assessment for education: Theory and application*. Cambridge: Cambridge University Press.

Leighton, J. P., & Gierl, M. J. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*, 3–16.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41*, 205–236.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43–71). Hillsdale: Erlbaum.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Education al Measurement, 33*, 379–416.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (Version 6) [Computer software and manual]. Los Angeles: Muthén & Muthén.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Columbus: Merrill Prentice Hall.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 307–353). Washington, DC: American Educational Research Association.

Rupp, A., & Templin, J. (2008). Unique characteristics of cognitive diagnosis models: A comprehensive review of the current state-of-the-art. Measurement: Interdisciplinary Research & Perspectives, 6, 219–262.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35*, 265–296.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*, 758–765.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge Academic.

Tatsuoka, K. K., Corter, J., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*, 901–906.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Templin, J., Henson, R., Rupp, A., & Jang, E. (2008, March). *Cognitive diagnosis models for nominal response data*. Paper presentation at the annual meeting of the National Council on Measurement in Education Society, New York, NY.

Templin, J., Henson, R., Douglas, J., & Hoffman, L. (2009, April). *Estimating a family of diagnostic classification models with Mplus*. Paper presentation at the annual meeting of the American Educational Research Association, San Diego, CA.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika, 68*, 251–265.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*, 195–210.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*, 327–347.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced* [Computer software and manual]. Belmont: Statistical Innovations Inc.

Wiggins, G. (1998). *Educative assessment: Designing assessment to inform and improve performance*. San Francisco: Jossey-Bass.

Chapter 6
Theory of Self-Directed Learning-Oriented Assessment: A Non-technical Introduction to the Theoretical Foundations and Methodologies of Cognitive Diagnostic Assessment

William John Boone, John R. Staver, and Melissa Seward Yale

6.1 Introduction

This chapter will be an effort to share 20 years of the first author's personal joy and mental growth as the result of using Rasch theory and Rasch Winsteps software for hundreds of research projects. Rasch measurement has allowed us to solve many real-world educational and psychological problems in classrooms, schools, and school districts. Thinking and evaluating data within the context of Rasch theory can be conducted at many levels. In all cases, no matter where researchers start, great strides can be made. In this chapter, we will share many of the teaching techniques we have employed to explain Rasch to diverse audiences – such as medical researchers, faculty in schools of business, teachers, and test developers. The techniques are generally non-mathematical and applied.

Many books and papers have been written about applying Rasch measurement. These documents each have unique strengths and weaknesses. Sometimes an entirely new and useful way of understanding Rasch is presented. Also, sometimes written text may be of little use to a specific reader. Our effort herein is to provide concepts and techniques that are easy to read, digest, and apply immediately to

W.J. Boone (✉)
Department of Educational Psychology, Miami University, Oxford, OH, USA
e-mail: boonewj@muohio.edu

J.R. Staver
Department of Curriculum and Instruction, College of Education, Purdue University, West Lafayette, IN, USA
e-mail: jstaver@purdue.edu

M.S. Yale
Department of Educational Studies, Purdue University, West Lafayette, IN, USA
e-mail: myale@purdue.edu

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_6, © Springer Science+Business Media Dordrecht 2013

problems in cognitive diagnostic assessment. Moreover, we hope to provide new perspectives about the benefits of using Rasch measurement even for veteran users of Rasch measurement. Parts of this chapter are based upon the forthcoming book (Boone et al. in preparation) also to be published by Springer.

6.2 Theoretical Premises of Rasch

Our goal is to target Rasch in light of one specific topic (e.g., self-directed learning-oriented assessment, cognitive diagnostic assessment). For those with additional interests in Rasch, the classic book *Best Test Design* (Wright and Stone 1979) should be consulted. That work discusses a wide range of Rasch techniques in great detail, techniques that can be used for test development. One of the major strengths of Rasch is its requirement that one must think before one leaps. One does not develop a test, compute a KR-20, item discrimination indices, and then decide what is good or bad about a test. When using Rasch measurement, one begins by thinking about what is being assessed. We try to think through examples of what it means to be at the one end of a continuum and what it means to be at the other end of a continuum.

We begin our brief introduction by considering a mathematics test that might be authored for assessing 15-year-old students. The goal of the test is to compute a student measure which then can be used to better understand student mastery of a district math curriculum. In Rasch measurement, we often draw a horizontal line, and at one end of the line, we might write "More Complex Mathematics Items," and at the other end of the line, we might write "Less Complex Mathematics Items" (Fig. 6.1).

Once the teacher, principal, or test developer has completed this simple task, a more difficult step is required. That step is to predict the location of sample items along the line. Locating items along the line requires thinking and reflecting in the context of theory. In this example, what have teachers observed are the most difficult mathematics concepts for students to master? Thought and reflection may suggest potentially complex mathematics items. Reflection may also help a teacher recall concepts and operations that appear to be odd in some manner. For example, some poorly performing students may unexpectedly master certain concepts and operations, while high-performing students might struggle with these same concepts and operations.

Such concepts and operations are important to monitor in some manner, but do they lend themselves to inclusion in a test? Use of theory also involves reviewing

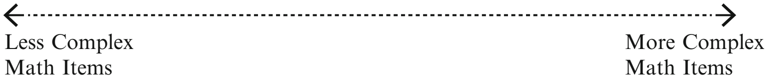


Fig. 6.1 A graphic to show mathematics as a single trait

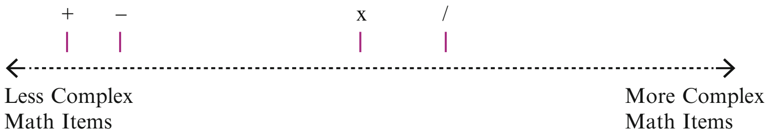


Fig. 6.2 The locations of four test items and the mathematical operation of each item have been added above the line of the math trait

mathematics educators' hypotheses regarding student learning in mathematics. For instance, there might be strong experimental evidence that addition is easier than subtraction. There might be evidence that multiplication of single digits is harder than subtraction of single digits. These and other techniques help the test developer mark the location of potential concepts and operations along the line of Fig. 6.1.

Figure 6.2 presents a simple mapping of mathematical operations along a line. When developing an entire test, predicting item locations is certainly a complex task; however, developing an entire diagnostic assessment should follow, not precede, this task. Why is this step important? By locating test items, test developers are forced to make sure that – as we say in English – we do not mix apples and oranges. Also, since a limit exists regarding the number of items that can be presented on a test – due to issues such as time and test taker fatigue – the procedures detailed in Figs. 6.1 and 6.2 help avoid presenting items that mark the same point on the line. For instance, most teachers would agree, if a 15-year-old student is taking a test, it makes little sense to present the test item "4 + 4 = ?" and to present the test item "5 + 5 = ?". The 15-year-old student who correctly answers the first item will more than likely correctly answer the second item. The 15-year-old student who misses the first item will likely miss the second item as well. As a result, at least for students at this level of mathematics ability, no additional information is gained about student understanding through the inclusion of both items.

When test developers use theory to predict an ordering of items along a trait, they should be reminded of an added nuance: If test developers wish to use the performance of the test taker on all test items to measure and compare students, then test items must indeed mark a level of performance along a single trait. Furthermore, if one notes the hardest item a student has correctly answered, one would predict that easier items should be correctly answered as well. Of course, this prediction will not always be correct, as nothing is perfect in the real world, but this general pattern should be observed.

6.3 Characteristics of Rasch Models

Continuing our introduction to Rasch measurement, we find it helpful to introduce some additional components. Rasch theory is a way of thinking, which is expressed through mathematical equations. The core of Rasch thinking, however, does not

$$\ln[\mathrm{P(ni)}/1 - \mathrm{P(ni)}] = \mathrm{Bn} - \mathrm{Di}$$

Fig. 6.3 An equation describing the Rasch model that can be used to evaluate data from a test whose items can be scored as right (1) or wrong (0)

need a mathematical equation to understand. The core is when one wishes to measure with a set of items, accurate measurement can take place only when all items define a single trait. Most researchers are familiar with the work of Jean Piaget, and we often use Piaget's work as an example to explain Rasch. Piaget hypothesized and tested the idea that humans must first pass through what he called a preoperational stage before they could "conserve." With Rasch measurement, the performance of a student on each item in a set of items should describe what that student knows and does not know with regard to the trait measured by the set of items. If we want to use a student's raw score to compare one student with another student, then there must be a general pattern in the test items from easy to hard. For example, if there is a haphazard pattern of correct and incorrect answers for Bob and Janet (Bob answered 15 correctly out of 30 on a vastly different set of items than Janet, who also answered 15 correctly out of 30 items), then one cannot confidently use a summary of the students' performance (number of items answered correctly) to compare Bob and Janet.

Figure 6.3 presents the mathematical expression of the Rasch model that is used to evaluate multiple-choice test data. The first important aspect of the model to note is its use of probabilities (the symbol "P" denotes probability). Why is using probabilities noteworthy? First, nothing is certain in this world of ours. For example, a highly performing high school student should get the easiest item on a middle school mathematics test correct. But one could never be 100% sure that this highly performing high school student will correctly answer the easiest item. Probabilities are used to evaluate and understand data in many fields of science. The use of probability is an important distinction of the Rasch model in comparison to other techniques used for the analysis of test data.

Now, let us turn our attention to the part of the equation with the symbols "n", "i", "B", and "D". What do these symbols mean and how do they help us comprehend the Rasch model? The term "Bn" represents the location of any person along the line of the trait. So, if we measure John along the trait of mathematics ability, the value of Bn represents John's location along the trait. For example, let us pretend that a low value of mathematics ability is 100 scale score units, and a very high value of mathematics ability is 800 scale score units. If John has "an ability level with respect to the trait" of 650, then John has a B of 650. Figure 6.4 displays John along the mathematics trait presented in Figs. 6.1 and 6.2. It is important to remember that the "n" of the term "Bn" refers to any person. Thus, "Bn" could represent Kim who has an ability level of 484. The "n" just means that we can think of different people. So, when you read the symbol "Bn", remember that the symbols B and n work together to represent the ability level of any person along a trait. The "B" means we are considering a person's ability, and the "n" is shorthand for a person. In this case, Bn can be BJohn or BKim!

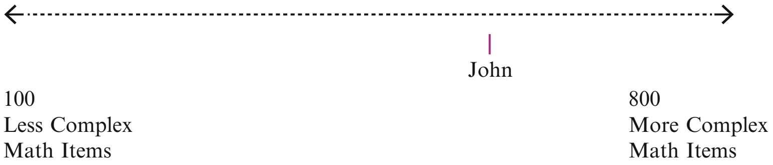


Fig. 6.4 The trait can be used to express the location of a test taker, and the scale of the trait is the same for items and persons

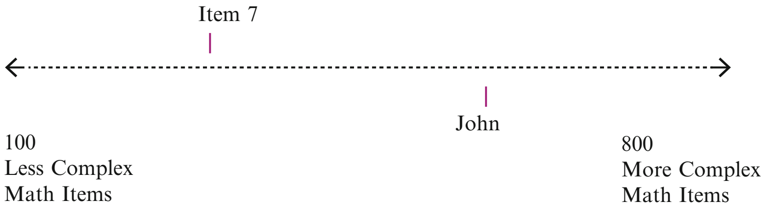


Fig. 6.5 A mathematics trait with the location of one item and one test taker plotted

What do the two symbols in the term "Di" represent? The term "Di" is shorthand for the "Difficulty" of any item of a test, in this case a mathematics test item. Just as the term Bn represents the ability of any test taker, the term Di represents the difficulty of any test item. Figure 6.5 presents the location of an item – let us call it item 7 of our mathematics test – along a single trait. Readers should note something very important, namely, that the location of the test item and the location of John are presented on the same scale. This means that John's mathematics ability and the level of difficulty of item 7 are expressed in the same units of measurement! This point generalizes, in that all students' mathematics abilities and the level of difficulty of all test items use the same units of measurement. This is a critically important aspect of the Rasch model. Later in this chapter, readers will see a number of advantages of being able to express items and persons in identical units of measurement along the same trait.

We are now ready to put all the pieces together. Referring back to the equation of Fig. 6.3, the equation states that the probability (Pni) of John answering item 7 correctly is dependent solely upon John's ability (Bn) with regard to the trait and the difficulty (Di) of the specific item John is attempting to answer. In the case of Fig. 6.5, we observe that John has a higher ability level (Bn) than the difficulty of the item (Di). So, we can assume that the probability of John correctly answering the item is above.5.

We will not discuss how the equation of Fig. 6.3 was derived; however, it is important for readers to appreciate three points. First, the Rasch model uses probabilities. Second, the model expresses the relationship between a person correctly answering an item, the difficulty of an item, and a person's ability level. Third, this relationship between each person and each item is solely dependent upon the difficulty of the test item and the ability level of the person.

6.4 Potential Benefits of Rasch Measurement

Many reasons exist for test developers and teachers to use Rasch measurement. Rasch measurement helps one think more critically about the items of a test. By using Rasch measurement, test developers and teachers can design tests that distinguish differences between students easier, faster, and cheaper.

Rasch measurement and the use of Winsteps allow one to quickly construct Wright maps, sometimes called person-item maps. These plots present the location of items and persons along the trait and provide the location of each student's ability measure.

Figure 6.6 presents an edited Wright map to simplify our discussion. Only four test items are presented. The mean test performance computed by Winsteps is presented for two schools, A and B. Let us now imagine that a school district conducted a statistical analysis of the mathematics test data (perhaps a *t*-test) and found a statistically significant difference in the performance of school A and school B. The results suggest that school B students on average performed better on the mathematics test than school A students. Evaluation of the effect size also suggests a meaningful difference. *But, what is the meaning of performing better?* Prior to Rasch analysis and Wright maps, teachers and administrators would not be able to explain the meaning of the difference between schools A and B, but now we can.

Our first step is to carefully draw a vertical line up from the location of each school. Now, we can see the mathematics concepts that were apparently mastered by each school. School A has a greater than 50–50 chance of correctly answering the items concerning addition (+) and subtraction (−). School B has a greater than 50–50 chance of correctly answering items involving addition (+), subtraction (−), and multiplication (×). Both schools need to work on division (/) because that item has a less than 50–50 chance of being correctly answered by the typical student at both school A and school B. The most important information, however, is the segment of the horizontal line of the trait that lies between the vertical line for school A and school B. This is the conceptual meaning of the statistical difference in the performance of the two schools. School B has mastered multiplication. Teachers at school A should spend time helping their students master multiplication.

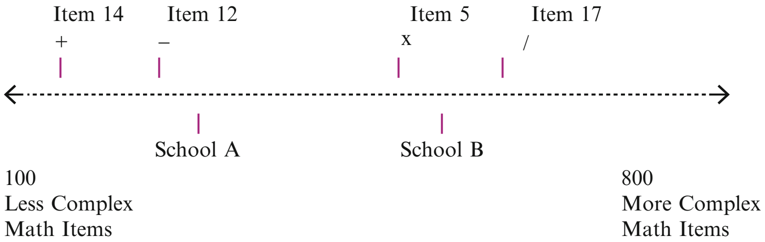


Fig. 6.6 The location of four test items and the mean performance of two schools (A and B) plotted along a single trait

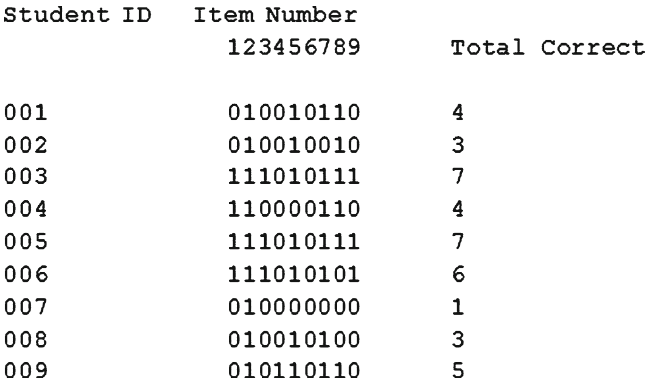
6.5 Method Itself

Earlier herein, we presented ideas central to understanding and using the Rasch model. Conceptualizing the variable, the basic mathematics of the model and the interplay of items and persons were presented. To further readers' understanding of the Rasch model, we now outline in broad strokes how data are used to compute item difficulty, where each item is located along the trait, and person ability, where each person is located along the trait.

A first step in understanding the method is to imagine what data collected as part of a test will look like when entered in a spreadsheet. Figure 6.7 presents such data. The first column frequently includes a student ID, and subsequent columns often indicate whether or not the student correctly answered the item. In this example, student 001 missed item 1 of the test – a "0" indicates this miss – and the same student correctly answered item 2 – a "1" indicates this correct answer. To compute the number of items correctly answered by person 001, one simply adds up the number of "1"s. Researchers will sometimes indicate a raw score total, the number of items correctly answered, in their spreadsheet. It is important to note that parametric statistical tests should not be performed on raw score totals; rather, such tests should be performed on person ability measures that are computed through application of the Rasch model with Winsteps. Computing a raw score total, however, does help us understand the development and application of the Rasch model.

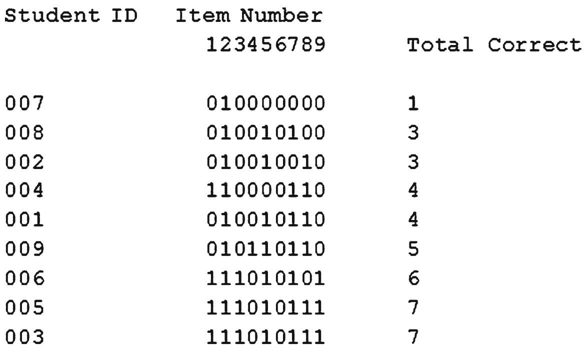
Let us now organize the rows of persons by the raw score total, the number of items they correctly answered. It does not matter if the best-performing students are on the top or bottom row. Why don't we place the lowest performers on the top line and work down to the best students? Readers will note, of course, that all we have done is move entire rows not columns. The answers presented by students in Fig. 6.8 below are in the same order as what was previously displayed in Fig. 6.7.

To help us better recognize what Rasch measurement does, now, organize the vertical columns for items from easiest item to hardest item. If we do this by hand,



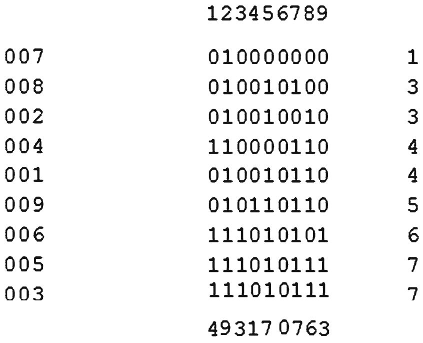
Student ID	Item Number	
	123456789	Total Correct
001	010010110	4
002	010010010	3
003	111010111	7
004	110000110	4
005	111010111	7
006	111010101	6
007	010000000	1
008	010010100	3
009	010110110	5

Fig. 6.7 Initial organization of student test data in a spreadsheet



Student ID	Item Number 123456789	Total Correct
007	010000000	1
008	010010100	3
002	010010010	3
004	110000110	4
001	010010110	4
009	010110110	5
006	111010101	6
005	111010111	7
003	111010111	7

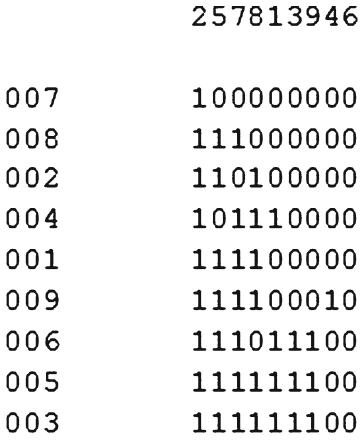
Fig. 6.8 Reorganization of student test data in a spreadsheet in which rows are ordered by lowest-performing students (007) to highest-performing students (005, 003)



	123456789	
007	010000000	1
008	010010100	3
002	010010010	3
004	110000110	4
001	010010110	4
009	010110110	5
006	111010101	6
005	111010111	7
003	111010111	7
	49317 0763	

Fig. 6.9 Row order of data maintained from Fig. 6.8, with lowest performer as the first row of data and a total correct for each row

we can help ourselves by first adding up the total correct in each column. We will keep the persons organized in rows from lowest performers to highest performers. In Fig. 6.9 below, we present the same data as in Fig. 6.8 immediately above, but the total number of students correctly answering each item is provided. For example, only four (4) people (ID 004, 006, 005, 003) correctly answered item 1. Therefore, we provide a “4” below the column of student answers for item 1. When doing this work by hand, we find it helpful to double check our work by noting the number of students who missed (0) an item. Item 1 was missed by students 007, 008, 002, 001, and 009. This makes sense because nine (9) students took the test; we computed that four (4) students correctly answered the item. If our calculations are correct, we should observe that five (5) students missed item 1, and our second calculation does indeed indicate that five students missed the item.



```
                 257813946

007              100000000
008              111000000
002              110100000
004              101110000
001              111100000
009              111100010
006              111011100
005              111111100
003              111111100
```

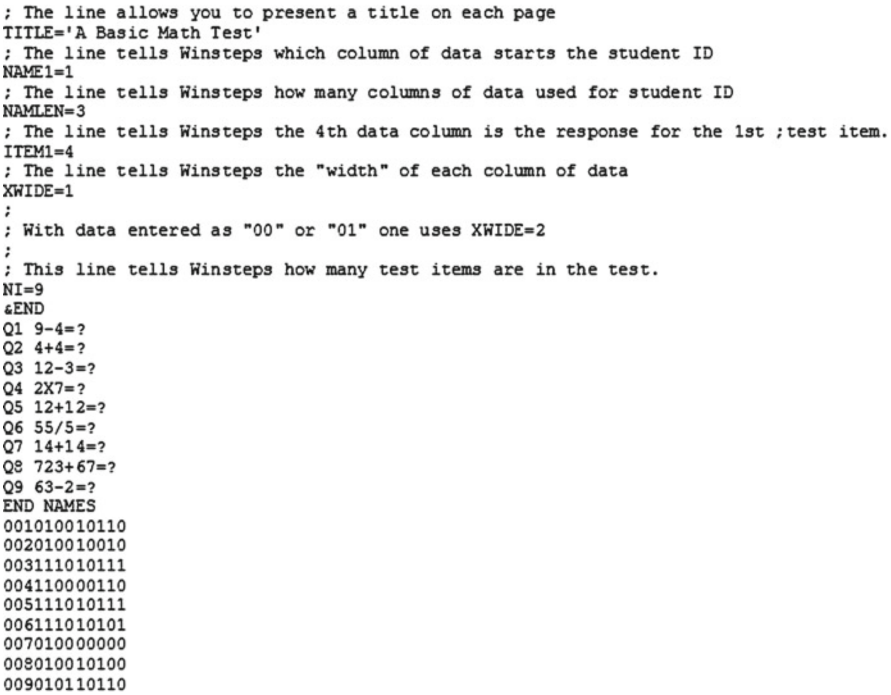
Fig. 6.10 Row order of data maintained from Figs. 6.8 and 6.9, but columns have been ordered from easiest test item (item 2) to hardest test item (item 6)

Conducting this procedure for all nine items helps us see that item 6 was the hardest item; there is a "0" below the column of numbers for that item because no one correctly answered that item. Review of the table of data also reveals that item 2 was the easiest item because a raw score of 9 is reported for item 2; all 9 test takers correctly answered this item.

A final step in organizing our data and understanding a core aspect of Rasch measurement is to order the rows of data for each respondent by lowest performer (top) to the best performer (bottom) and to also order the columns of data from the easiest item (left most column) to the hardest item (right most column). The hardest item is the item most often answered incorrectly for the entire group of respondents. The easiest item is the item that was answered correctly by the largest number of test takers. The data are already organized from lowest performer to highest performer, so all one must do to organize the data is move the entire columns of data. Figure 6.10 presents the reorganized data. The data column for item 2 (the easiest item) is presented on the far left. The column immediately to the right of the data for item 2 is the data for item 5. One could have just as well presented the data for item 7 (for item 7 was just as easy for respondents) as item 5.

Organizing data in this manner is not just an exercise to test one's ability to add and move columns. The general pattern of 0s and 1s in Fig. 6.11 can be used to help one better understand concepts of the Rasch model. In Fig. 6.10 above a diagonal line roughly separates the 0s and 1s. One can see a pattern when data are organized by total raw score for each respondent and by item difficulty. This pattern should make sense, in that lower-performing students would be expected to miss easier items and higher-performing students would be expected to miss harder items while correctly answering easier items.

When we explain Rasch to our students, we ask them to construct such a matrix of data, and we explain that the Rasch model can be, in part, viewed as an endeavor to use the mathematical model and the test data to make sense out of the pattern of



```
; The line allows you to present a title on each page
TITLE='A Basic Math Test'
; The line tells Winsteps which column of data starts the student ID
NAME1=1
; The line tells Winsteps how many columns of data used for student ID
NAMLEN=3
; The line tells Winsteps the 4th data column is the response for the 1st ;test item.
ITEM1=4
; The line tells Winsteps the "width" of each column of data
XWIDE=1
;
; With data entered as "00" or "01" one uses XWIDE=2
;
; This line tells Winsteps how many test items are in the test.
NI=9
&END
Q1 9-4=?
Q2 4+4=?
Q3 12-3=?
Q4 2X7=?
Q5 12+12=?
Q6 55/5=?
Q7 14+14=?
Q8 723+67=?
Q9 63-2=?
END NAMES
001010010110
002010010010
003111010111
004110000110
005111010111
006111010101
007010000000
008010010100
009010110110
```

Fig. 6.11 A sample Rasch Winsteps control file

0s and 1s. The details of how the mathematics and the software make sense of the data are beyond the goal of this chapter. For beginning Rasch users, the important issue to note is that when a test involves a single trait, there should be a pattern of 0s and 1s, and the presence of this pattern can be used to ultimately compute Rasch person measures and Rasch item measures which then can be used for parametric statistical analyses.

6.6 Information on Software

A wide range of Rasch software exists. Some large statistical packages contain modules that facilitate Rasch analysis research; however, specific software programs have been developed to conduct Rasch analysis. The software that we use and recommend to readers is Winsteps. This software is Windows-based and easy to use. For the beginner, any software package can be scary to use; however, Winsteps is an exception in that detailed guidance is provided in a software manual. Additionally, the author, Mike Linacre, provides almost immediate feedback when a user sends a question to him by email. Ease of use, support, detailed understandable documentation, and capability make Winsteps our choice for Rasch software. At a more specific level, Winsteps can read an Excel, SPSS, STATA, or SAS file. One can begin a Rasch analysis of a data set with only a few key strokes. The Winsteps manual furnishes a wealth of guidance for beginners. Invaluable guidance is also provided in the text authored by Bond and Fox (Bond and Fox 2007). This book provides free Winsteps software and ready to run data files, but not all the tables are included in the free software. Finally, it is important to note that Winsteps can handle data sets up to 10,000,000 persons and 30,000 items!

6.7 Examples (Part 1 of 2)

To help readers better understand and use Rasch measurement and Winsteps software to develop a cognitive diagnostic assessment, we provide a sample code in Fig. 6.11 that can be used for a simple Winsteps analysis. Readers will see that the sample code uses the data presented in earlier portions of this chapter. The Winsteps code used to run a Rasch analysis is quite simple. Moreover, one can add additional lines of code to the Winsteps file, but the point of this chapter is to help readers better understand how easy it is to author Winsteps code to develop a cognitive diagnostic assessment instrument and to evaluate data collected with such devices.

First, notice one can start a line with the symbol ";", and when one does so, the program does not read the line at all. In Winsteps jargon, the file provided below (which includes lines to tell Winsteps how to read the data and which lines include the data) is referred to as a "control file."

The control file presented above is simple, and it could be typed in by hand if necessary. Briefly, what action does each line do? First, any line that starts with a semicolon is NOT read by the program. Usually, Winsteps users type comments into their control file to remind themselves of what they have done and why they have done it. The line that starts with the word TITLE specifies a particular phrase that is printed on the output of the analysis. This is a good way to keep track of a specific analysis a number of data sets are analyzed. The line NAME1 = 1 tells the program that the first piece of data identifying a person ID is in column 1 of the data set. NAMLEN = 3 tells the program that the person ID information is 3 columns wide. In this case, the person ID starts in column 1 and ends in column 3. ITEM1 = 4 tells the program that the 4th column of data is the first item in the data. XWIDE = 1 tells the program that one column of data is used to indicate each answer. If the data had been entered as "00" for incorrect and "01" for correct, then XWIDE = 2 would have been used. Finally, one sees two more lines, NI = 9 and &END. The first phrase tells Winsteps how many items are in the data set. In this case, there are nine items. Finally, the phrase &END tells the program that the final command line has been reached. There are two additional parts of the control file. Following the line &END, one sees that there are nine lines that describe each of the math items. Anything can be typed into a line. The important thing to remember is to type descriptions for each item you will read in your data file. You will want to make sure that you type only a one-line descriptor. When your description can be short, this will help you when you look at some of the results of your analysis. So, type a descriptor that quickly and succinctly describes your item. Following the nine-item descriptions, there is always a line with the following: END NAMES. This line tells the program that the end of the item names has been reached. Finally, we have one more type of line. This is our data. Please note that the ID indeed starts in the 1st column and is 3 columns wide. The 4th column of data is the start of the data, and there are a total of 9 columns of data.

To run this control file and conduct a Rasch analysis, first type the file into a Word document. Then save the file as a .txt file, not as a Word document. You can use Word to type in your document, but when you save the file, do not just click on save; that will save your file as a Word document. Instead, make sure to select "Save As". When you see the phrase "Save as type:", go to the drop down box for type and select "Plain Text". Individual Winsteps users often name their control files with the letters "cf" for control file. So, you might call this file "BasicMathTestcf.". Remember, when you look for this file on your computer, it will have this name and an extension (the part after the dot) of "txt". So, the control file name will appear on your computer as file BasicMathTestcf.txt.

We are now ready to run Winsteps. If you have downloaded Winsteps, you will be able to double-click on the Winsteps icon (you will see a nice "W") on your screen. If your data set is not huge, you can learn how to use Winsteps by downloading the free Ministeps program. Ministeps, supplied by Mike Linacre, does much of what Winsteps does, but there is a limit to how many items and persons can be evaluated.

To run our cognitive diagnostic assessment, double-click on the Winsteps program icon or the free Ministeps icon. You will see a gray box on your screen and the top two lines of the box read "Welcome to Winsteps!", "Would you like help setting up your analysis?". Since we have written our control file and have data in the file, we can just click on the square marked "No". This square is just to the left of the square in the box marked "Help". Once you click on the square marked "No", a white screen will appear, but you will see the following text:

Control file name? (e.g., exam1.txt). Press Enter for Dialog Box:

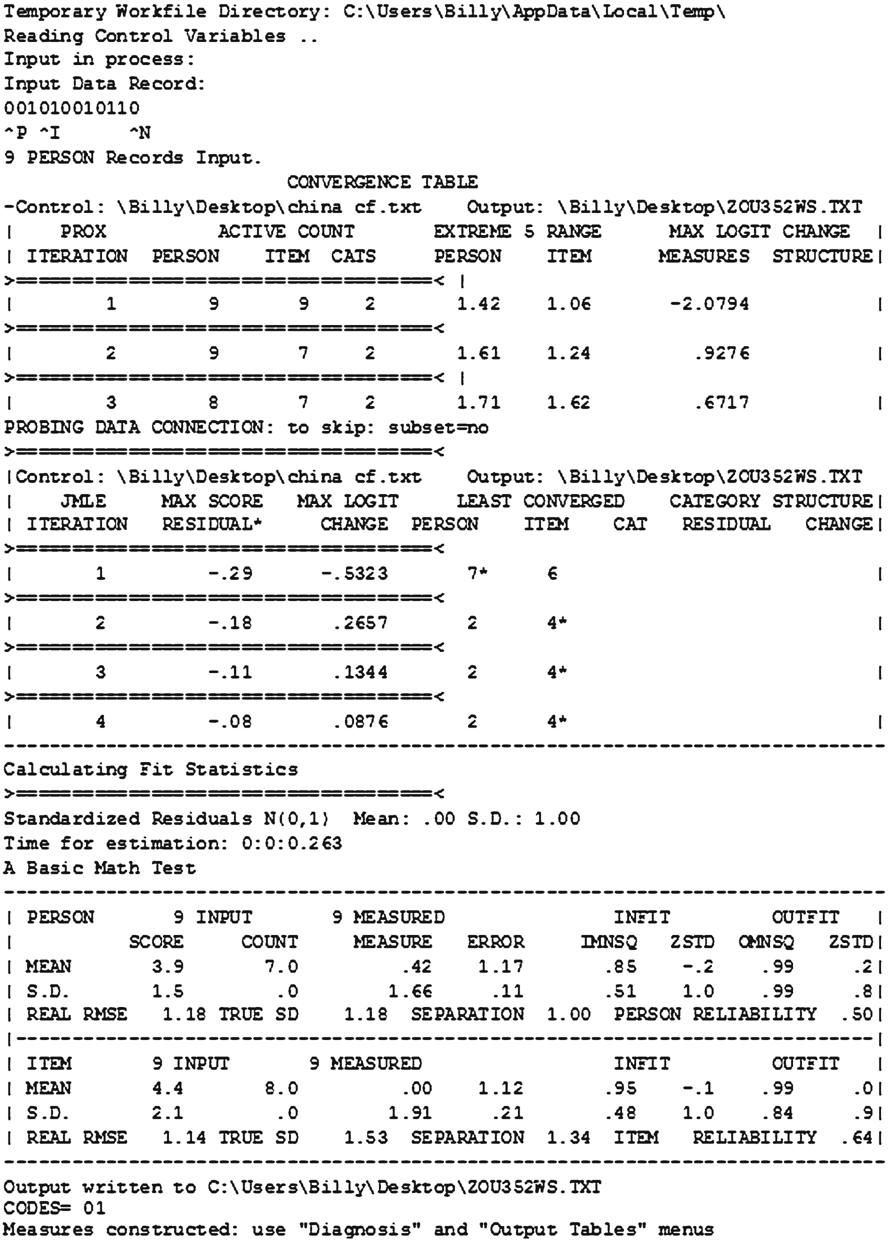
Now, push the "Enter" key on your keyboard. This allows you to tell Winsteps where the control file you want to look at is located. When you find the file on your computer, you then click on the Microsoft "open" button located at the lower right part of your screen. Once you click the "open" button, you will see the white screen again and the phrase:

Report output file name (or press Enter for temporary file, Ctrl + O for Dialog Box):

At this point, the program is asking you if you want the output dumped to a particular file. Since one can complete an analysis so easily using Windows, just push the "Enter" button on your keyboard. This tells the program that you do not want a specific output file with hundreds of printed pages. Once you depress the “Enter” key, you will see the following text on the screen:

Extra specifications (if any). Press Enter to analyze:

Let us not worry about this for now. Just hit the “Enter” key to start the Winsteps analysis! Immediately below is approximately what you will see on your screen when the analysis has been read in correctly.



```
Temporary Workfile Directory: C:\Users\Billy\AppData\Local\Temp\
Reading Control Variables ..
Input in process:
Input Data Record:
001010010110
^P ^I     ^N
9 PERSON Records Input.
                         CONVERGENCE TABLE
-Control: \Billy\Desktop\china cf.txt    Output: \Billy\Desktop\ZOU352WS.TXT
|    PROX          ACTIVE COUNT       EXTREME 5 RANGE      MAX LOGIT CHANGE  |
| ITERATION  PERSON    ITEM  CATS     PERSON    ITEM      MEASURES  STRUCTURE|
>===============================< |
|        1        9        9     2      1.42    1.06       -2.0794           |
>===============================<
|        2        9        7     2      1.61    1.24         .9276           |
>===============================< |
|        3        8        7     2      1.71    1.62         .6717           |
PROBING DATA CONNECTION: to skip: subset=no
>===============================<
|Control: \Billy\Desktop\china cf.txt    Output: \Billy\Desktop\ZOU352WS.TXT
|    JMLE     MAX SCORE   MAX LOGIT     LEAST CONVERGED     CATEGORY STRUCTURE|
| ITERATION   RESIDUAL*     CHANGE  PERSON     ITEM     CAT   RESIDUAL   CHANGE|
>===============================<
|       1         -.29       -.5323      7*     6                              |
>===============================<
|       2         -.18        .2657      2      4*                             |
>===============================<
|       3         -.11        .1344      2      4*                             |
>===============================<
|       4         -.08        .0876      2      4*                             |
------------------------------------------------------------------------------
Calculating Fit Statistics
>===============================<
Standardized Residuals N(0,1)  Mean: .00 S.D.: 1.00
Time for estimation: 0:0:0.263
A Basic Math Test
------------------------------------------------------------------------------
| PERSON       9 INPUT       9 MEASURED              INFIT         OUTFIT    |
|            SCORE      COUNT      MEASURE    ERROR      IMNSQ  ZSTD  OMNSQ  ZSTD|
| MEAN        3.9        7.0          .42     1.17        .85    -.2    .99    .2|
| S.D.        1.5         .0         1.66      .11        .51    1.0    .99    .8|
| REAL RMSE   1.18 TRUE SD    1.18  SEPARATION  1.00  PERSON RELIABILITY  .50|
|----------------------------------------------------------------------------|
| ITEM        9 INPUT       9 MEASURED               INFIT         OUTFIT    |
| MEAN        4.4        8.0          .00     1.12        .95    -.1    .99    .0|
| S.D.        2.1         .0         1.91      .21        .48    1.0    .84    .9|
| REAL RMSE   1.14 TRUE SD    1.53  SEPARATION  1.34  ITEM   RELIABILITY  .64|
------------------------------------------------------------------------------
Output written to C:\Users\Billy\Desktop\ZOU352WS.TXT
CODES= 01
Measures constructed: use "Diagnosis" and "Output Tables" menus
```

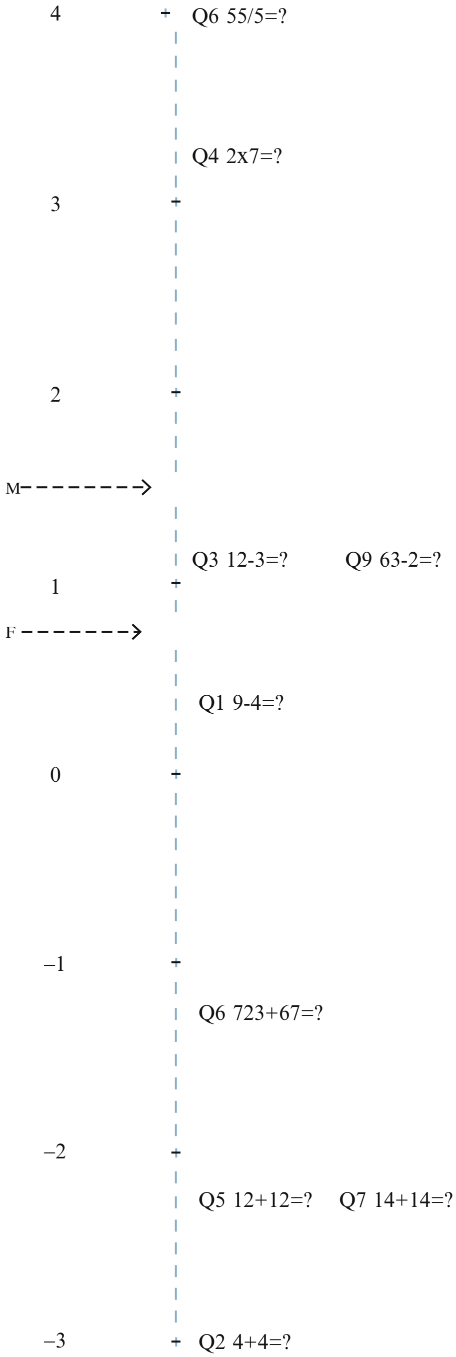
6.8 Examples (Part 1 of 2)

Unfortunately, we cannot present even 1% of what is possible in terms of cognitive diagnostic assessment using Rasch measurement. However, a single, relatively simple step can help one begin learning how to use the results of a Rasch analysis. After completing your Winsteps run with the data set, look at the top of your screen for a horizontal gray bar. The leftmost word will be “File”, which is followed by the word “Edit”. Now find the word “Output Tables”, and click on that button. Now select Table 12, which is displayed in Fig. 6.12 immediately below. This table is identified with the phrase “12. ITEM: map”. Click on this phrase and you will see a plot that looks very similar to some of our figures, except that the plot is vertical. When Wright maps are made, it does not matter if the maps are vertical or horizontal. Some prefer to look at Wright maps as if they were reading a time line (e.g., dinosaurs at one end, humans at the other). But others like to read Wright maps as if they were gazing at a thermometer.

How can this Wright map, which contains only a few items, be used for cognitive diagnostic assessment? First, the Rasch logit scale, the range of numbers on the far left from −3 to 4, is an equal-interval scale. If needed, we can convert this scale from −3 to 4 to a positive scale, say 0–7. Also, we can convert to numbers that teachers, parents, and students are more familiar with, but for our purpose herein, the important points are that a low logit value (e.g., −3) means an easy item and a low-performing student, and a higher logit value (e.g., 3) means a harder item and better-performing student. An equal-interval scale means that the quantitative amount of change in student performance and item difficulty is the same for any pair of equally separated points on the scale, say between −2 and −1 compared to 3 and 4. The pattern one sees from easiest item (Q2) to the hardest item (Q6) represents the hierarchy and spacing of item difficulty. For a teacher using this Wright map, this immediately suggests the order in which topics might be optimally presented to students. Another critical lesson from this Wright map has to do with the magnitude of the gaps between items. Since the Rasch item measures are plotted on an equal-interval scale, the gaps have immediate meaning to teachers. Returning to our point immediately above about the quantitative amount of change, let us consider the size of two gaps. Gap 1 is between item 1 and item 9 (as well as item 3). Gap 2 is between item 9 (and item 3) and item 4. A rough calculation to measure the two gap sizes shows that Gap 1 is about half the size of Gap 2. This means that when students learn mathematics, the growth from mastery of the topics of item 1 to mastery of item 9 topics represents about half the growth as a student moving from mastery of item 9 topics to item 4 topics.

This type of information was not available prior to the use of Rasch analysis. Understanding and appreciating the gaps is a struggle for many people. When we teach, we often discuss Olympic diving. Even though we can only do belly flops when we dive, it makes sense to us that some dives are more difficult than other dives. Our students also agree with this idea! Then we discuss the different types of dives one can accomplish. For the students and for us, it makes sense that the added

Fig. 6.12 A Wright map computed following Winsteps Rasch analysis of the control file presented in Fig. 6.11



skill to complete a double somersault compared to single summersault is probably different than the added skill needed to complete three somersaults compared to two. This is an intuitive example that our students understand.

A second Wright map application that connects to earlier parts of this chapter will hopefully provide a boost in readers' understanding. Let us pretend that the Wright map presented above was constructed using a larger number of students, perhaps 200 students who attend a local school. Let us further imagine that the mean scores of the male and female test takers were computed and that a *t*-test suggests a statistical difference between males and females. To bring meaning to this difference, a researcher would want to plot the location of the typical male (M) and the typical female (F). The hypothetical locations of male and female means are marked along the scale of the Wright map. These marks help one see two points. First, the males did better than the females. Second and more important, the substance of this key difference is that the females are typically not able to answer items 3 and 9. Knowledge of this difference then provides immediate diagnostic guidance to the teacher, who must then ask, why is this the case? What can I do to help my female students?

6.9 Pitfalls

There are always pitfalls to learning new techniques. Regarding Rasch measurement, we prefer to use the word caution as opposed to pitfalls. Let us discuss some cautions. First, Rasch measurement causes one to think. Of course, one can enter a data set into Winsteps and get an output. In Rasch measurement, however, we think about the variable we are going to measure before we conduct an analysis. In many cases, Rasch theory may not have been used to develop a test, and we are contacted to provide guidance after the fact. When using Rasch measurement, we always think before we leap. For example, thinking helps us decide if it makes sense to include all items of a test for the computation of a person measure. Thinking allows us to examine our results and more quickly conduct quality control of our data. Of course, it takes time to think, so sometimes Rasch measurement takes more time, but it is the only technique that results in useful measures that can be analyzed with parametric statistics.

A second caution is that some researchers argue that Rasch is useful only for large data sets. Given our work of 20 years with data sets ranging in size from approximately 25 students to 10,000 students, we have found that one can uncover trends in data even with very small sample sizes and that Rasch is VERY useful with small data sets.

A third caution focuses on an important misconception. Rasch is sometimes called an item response theory (IRT) model. This is because the mathematics of the Rasch model looks like the mathematics of IRT models. We encourage readers to NOT view Rasch as an IRT model. IRT models are altered to fit data; the Rasch model is not altered to fit data. In Rasch analysis, the Rasch model is viewed as a definition of measurement. If data fit the model, then the Rasch perspective is that measurement may be possible with that data set.

6.10 Future Developments

Throughout the world, Rasch measurement is being applied to solve measurement problems in a wide variety of fields. Continual refinement and expansion of Winsteps software suggests that the software will continue to facilitate fast, accurate analyses of complex data sets ranging from right/wrong tests to partial-credit tests, to survey data, and to data involving test takers-judges-tasks. The use of Rasch measurement in many fields and as a part of interdisciplinary research efforts seems to be ever expanding with no end in sight.

6.11 Conclusions

This chapter presents key ideas and examples of cognitive diagnostic assessment informed by Rasch measurement theory and the application of Rasch measurement. Using this chapter, teachers and all others who develop assessments can quickly and thoughtfully create assessments that lend themselves to the collection and analysis of data that can be used to inform decisions. Please feel free to contact the authors if you have any questions!

References

Boone, W., Staver, J., & Yale, M. (forthcoming and tentatively entitled: An Introduction to the Theory and Application of Rasch Measurement). Springer.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah: Erlbaum.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Chapter 7
Getting to the Core of Learning: Using Assessment for Self-Monitoring and Self-Regulation

Lorna Earl and Steven Katz

Pupils in modern society are living in confusing and unpredictable times, in which they must be equipped with skills that enable them to think for themselves and be self-initiating, self-modifying and self-directing. They must acquire the capacity to learn and change consciously, continuously and quickly, to anticipate what might happen and to continually search for more creative solutions. Learning for the twenty-first century involves much more than acquiring knowledge. It requires the capacity for 'reflective judgement' – the ability to make judgements and interpretations, less on the basis of 'right answers' than on the basis of 'good reasons' (King and Kitchener 1994).

Delors et al. (1996), in their powerful work for UNESCO *Learning: the treasure within*, identified four essential pillars of learning – learning to know, learning to do, learning to live together and learning to be – a testament to the growing need for informed, skilled and compassionate citizens who value truth, openness, creativity, interdependence, balance and love, as well as the search for personal and spiritual freedom in all areas of one's life. This image of learning means fundamental changes in orientations to teaching and learning in schools. It means that schools must become places that foster high-level learning for all students in all of these domains.

Assessment has the potential to be a key element in transforming schools into places of high-quality learning for all students. Why? Because assessment can be one of the most powerful processes that schools and teachers have to prepare students for the future in all of the domains in the UNESCO framework.

L. Earl (✉)
Aporia Consulting Ltd., Toronto, Canada

S. Katz
Psychological Foundations of Learning and Development at OISE, University of Toronto, Toronto, Canada
e-mail: steven.katz@utoronto.ca

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*, Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_7, © Springer Science+Business Media Dordrecht 2013

7.1 Assessment for Learning

Since the ground-breaking work of Terry Crooks (1988), Black and Wiliam (1998a, b), and the Assessment Reform Group (1999), assessment for learning has taken hold worldwide as a high-leverage approach to school improvement. In assessment for learning, we have a pedagogical approach that has the potential, at least, to influence student learning. But, as many authors have told us, assessment for learning is not a 'quick fix'. For teachers really to engage in assessment for learning requires a lot of new professional learning, and it requires changes in how teachers interact with their pupils, how they think about the material they teach and, most importantly, how they use assessment in their daily work. Much of this volume is focused on helping teachers understand assessment for learning better – both the theory on which it is based and the practical processes that make it work.

In this chapter, we aim to provide teachers with a deeper understanding of the ways that assessment can help pupils become thoughtful, self-monitoring and self-regulating learners. Assessment for learning is based on a complex set of ideas and theories and provides a model for teachers to use assessment to rethink, revise and refine their teaching. It also assists them in the provision of feedback and to focus on creating the conditions for pupils to become confident and competent self-assessors. In our experience with teachers who are engaging in assessment for learning, they are often preoccupied with using assessment to inform their teaching decisions and provide feedback to students. Sometimes the feedback gives students the raw materials for becoming better at self-assessment. However, teachers rarely think proactively about what they need to do to use assessment to promote student self-assessment and self-regulation so that students become adept at defining their own learning goals and monitoring their progress towards them.

This chapter is concerned with this second dimension of assessment for learning, emphasising the role of the pupil as the critical connector between assessment and learning. We have called this 'assessment as learning' (Earl 2003; Earl and Katz 2005) – the kind of assessment that recognises students as active, engaged and critical assessors who make sense of information, relate it to prior knowledge and use it for new learning. This is the regulatory process in metacognition, in which students personally monitor what they are learning and use the feedback from this monitoring to make adjustments, adaptations and even major changes in what they understand. When teachers focus on assessment as learning, they use classroom assessment as the vehicle for helping pupils develop, practise and become comfortable with reflection and with critical analysis of their own learning. Viewed this way, self-assessment and meaningful learning are inextricably linked.

In this chapter, we expand on this notion of assessment as learning by showing how it relates to current learning theory and by describing teachers' roles in developing reflection and self-regulation in their pupils.

7.2 Assessment and Learning

How People Learn: Brain, Mind, Experience and School, the seminal synthesis of literature in the cognitive and developmental sciences produced by the National Research Council in the USA (Bransford et al. 1999), identified three principles that underpin how people learn:

1. Students come to the classroom with preconceptions about how the world works. If their understanding is not engaged, they may fail to grasp the new concepts and information, or they may learn them for purposes of the test but revert to their preconceptions outside the classroom.
2. To develop competence in an area of inquiry, students must have a deep foundation of factual knowledge, understand facts and ideas in the context of a conceptual framework and organise knowledge in ways that facilitate retrieval and application.
3. A 'metacognitive' approach to instruction can help students learn to take control of their own learning by defining learning goals and monitoring their progress in achieving them.

These principles portray learning as an interactive process by which learners try to make sense of new information and integrate it into what they already know. Students are always thinking, and they are either challenging or reinforcing their thinking on a moment-by-moment basis.

Before teachers can plan for targeted teaching and classroom activities, they need to have a sense of what it is that pupils are thinking. What is it that they believe to be true? This process involves much more than 'Do they have the right or wrong answer?' It means making pupils' thinking visible and understanding the images and patterns that they have constructed in order to make sense of the world from their perspective (Earl 2003). It means using this information to provide scaffolding for the learner to create new connections and attach these connections to a conceptual framework that allows efficient and effective retrieval and use of the new information.

The following anecdote gives a vivid description of how this learning process happens and the critical role that assessment plays in the learning process. When she was about 5 years old, my niece Joanna (Jojo to the family) came up to me and announced that 'All cats are girls and all dogs are boys'.

When I asked her why she believed cats were girls and dogs were boys, she responded: 'Your cat Molly is a girl and she's little and smooth, girls are little and smooth, too. Cats are girls. The dog next door is a boy and he's big and rough, just like boys are big and rough. Dogs are boys'. Clearly, she had identified a problem, surveyed her environment, gathered data and formulated a hypothesis, and, when she tested it, it held – pretty sophisticated logic for a 5-year-old.

I pulled a book about dogs from my bookshelf and showed her a picture of a chihuahua, a dog that was little and smooth.

> 'What's this?' I asked.
> 'Dog', she replied.
> 'Girl or boy?'
> 'It's a boy, dogs are boys'.
> 'But it's little and smooth', I pointed out.
> 'Sometimes they can be little and smooth', said Jojo.
> I turned to a picture of an Irish setter, surrounded by puppies. She was perturbed.
> 'What's this?'
> 'Dog', she replied, with some hesitation.
> 'Boy or girl?'

After a long pause she said, 'Maybe it's the dad'. But she didn't look convinced and she quickly asked: 'Can dogs be girls, Aunt Lorna?'

This anecdote is a simple but vivid demonstration of the process of assessment, feedback, reflection and self-monitoring that we all use when we are trying to make sense of the world around us. Jojo had a conception (or hypothesis) about something in her world (the gender of cats and dogs). She had come to a conclusion based on her initial investigation that held with her experience. With the intervention of a teacher (Aunt Lorna) who used assessment (How do you know?) and created the conditions (the picture book) for her to compare her conceptions with other examples in the real world, she was able to see the gap between her understanding and other evidence. Once she had the new knowledge, she moved quickly to adjust her view and consider alternative perspectives.

This kind of assessment is at the core of helping pupils become aware of and take control of their own learning. And it is this kind of assessment that supports the type of learning that psychologists describe as conceptual change. Rather than transforming evidence that exists in the world to fit established mental structures (conceptions), the mental structures themselves shift (or accommodate) to take new evidence into account. Classroom assessment, in this view, promotes the learner's accommodation process. It is something best – and necessarily – accomplished by the learner herself since it is she who holds privileged access to the relevant beliefs, though as we saw above, the teacher's role is to help make them public (Katz 2000).

7.3 Assessment as Learning

Assessment as learning is premised on the need for all young people to become their own best self-assessors. Why? Because self-assessment is the third fundamental principle of how people learn (Bransford et al. 1999). Although the first two principles identified above are key ingredients of good pedagogy and enhanced learning, the third principle is the one that underpins self-awareness and life-long learning – creating the conditions to develop metacognitive awareness so that they have the skills and habits to monitor and regulate their own learning. Metacognition, as defined by Brown (1987), has two dimensions – 'knowledge of cognition' (knowledge about ourselves as learners and what influences our performance, knowledge about learning strategies and knowledge about when and why to use a strategy) and 'regulation of cognition' (planning – setting goals and activating relevant background knowledge; regulation – monitoring and self-testing; and evaluation – appraising the products and regulatory processes of learning).

Metacognition means that pupils must become reflective about their own learning, a skill that like all complex learning requires years of practice, concentration and coaching. It does not have a beginning and an end but rather continues to develop and to be honed across disciplines and contexts (Costa 2006). And it does not happen by chance. If pupils are to become metacognitive thinkers and problem solvers who can bring their talents and their knowledge to bear on their decisions and actions, they have to develop these skills of self-assessment and self-adjustment so that they can manage and control their own learning.

7.4 Helping Pupils Become Their Own Best Assessors

To become independent learners, students must develop a sophisticated combination of skills, attitudes and dispositions. Students become productive learners when they see that the results of their work are part of critical and constructive decision making. They need to learn to reflect on their own learning, to review their experiences of learning (What made sense and what did not? How does this fit with what I already know, or think I know?) and to apply what they have learned to their future learning.

Self-monitoring and self-regulation are complex and difficult skills that do not develop quickly or spontaneously. Teachers have the responsibility for fostering and cultivating these skills. The rest of this chapter is concerned with how teachers can foster the development of self-assessment and self-regulation in pupils.

7.4.1 Habits of Mind for Self-Regulated Thinking

A number of writers have referred to the 'habits of mind' that creative, critical and self-regulated thinkers use and that students (and many adults, for that matter) need to develop. These habits are ways of thinking that will enable them to learn on their own, whatever they want or need to know at any point in their lives (Marzano et al. 1993).

When people succeed or fail, they can explain their success or failure to themselves in various ways: effort, ability, mood, knowledge, luck, help, interest, clarity of instructions, unfair policies and so on. Some of these are controllable; others are not.

Attribution theory makes clear that to the extent that successes and failures are explained by (attributed to) controllable factors, adaptive motivational tendencies will follow (Weiner 2000). Self-assessment is the mechanism by which learners assign attributions to particular outcomes, and the teacher's role is to help pupils learn how to shift their attributions away from uncontrollable explanations (like ability) to controllable ones (like effort). A student who explains a poor result in a math test by appealing to a lack of ability will be more likely to repeat the same behaviour pattern and meet with the same result on a future occasion than one who attributes the outcome to having not studied the correct material. In the latter example, the subsequent behaviour pattern actually shifts so that the learner asks himself or herself the regulatory question 'Am I focusing on the right material?' at the outset.

Several authors have identified an 'inquiry habit of mind' as an essential component of profitable learning for individuals and groups (Newmann 1996; Costa and Kallick 2000; Earl and Lee 2000; Katz et al. 2002). If pupils are going to develop these 'habits of mind' and become inquiry-minded, they need to experience continuous, genuine success. They need to feel as if they are in an environment where it is safe to take chances and where feedback and support are readily available and challenging. This does not mean the absence of failure. It means using their habits of mind to identify misconceptions and inaccuracies and work with them towards a more complete and coherent understanding. Teachers have the responsibility of creating environments for pupils to become confident, competent self-assessors who monitor their own learning.

7.4.2 Lots of Examples of 'What Good Work Looks Like'

As Sadler (1989) suggested, pupils' ideas of quality can approach those of the teacher if they have good exemplification and support; this is what he refers to as 'guild knowledge'. This knowledge is a prerequisite for pupils, taking responsibility for their own learning and for setting their own targets, since success is only possible if the end results are clearly delineated. Knowing what good work looks like not only increases the learner's conceptual awareness and provides reference points to strive for but also enhances his or her metacognitive awareness of the progress of learning. With such insight and engagement, pupils become more proficient in monitoring their work continuously during production while developing sustainable learning and self-assessment skills. They develop a repertoire of approaches, such as editing and self-evaluating in addition to that of setting their own targets, since their needs become apparent as part of the procedure. If, as Sadler argued, self-assessment is essential to learning because students can only achieve a learning goal if they understand that goal and can assess what they need to do to reach it, the criteria for evaluating any learning achievements must be made transparent to students to enable them to have a clear overview both of the aims of their work and of what it means to complete it successfully.

Although curriculum guides and standards (such as the national curriculum, schemes of work and level descriptions) provide a skeleton image of the expectations for students, nothing is as powerful as multiple images of ‘what it looks like when experts do it’. Not only do pupils begin to see, hear and feel the expectations for the work at hand, they become acutely aware of the variations that can occur and the legitimacy of those variations. Once learners have a sense of where they are aiming, teachers can offer many intermediate examples of the stages along the way and how experts also struggle to meet their own expectations.

Many assessment methods have the potential to encourage reflection and review. What matters in assessment as learning is that the methods allow students to consider their own learning in relation to models, exemplars, criteria, rubrics, frameworks and checklists that provide images of successful learning. When pupils contribute to developing these models, they are even more likely to internalise them and develop a concrete image of what ‘good work looks like’.

7.4.3 Real Involvement and Responsibility

When teachers work to involve pupils and to promote their independence, they are really teaching pupils to be responsible for their own learning and giving them the tools to undertake it wisely and well (Stiggins 2001). How else are they likely to develop the self-regulatory skills that are the hallmark of experts? It is not likely, however, that pupils will become competent, realistic self-evaluators on their own. They need to be taught skills of self-assessment, have routine and challenging opportunities to practise and develop internal feedback or self-monitoring mechanisms to validate and to call into question their own judgements. They compare their progress towards an achievement goal and create an internal feedback loop for learning. The more control and choice that students have in thinking about their learning, the less likely they are to attribute their understanding (or lack of understanding) to external factors like teachers or subject matter. Instead, they become more responsible for their learning and have increased self-efficacy and resilience. For pupils to become independent learners, they need to develop a complicated combination of skills, attitudes and dispositions in order to set goals, organise their thinking, self-monitor and self-correct. Each of these skills can be learned by engaging pupils in these activities and helping them change their learning plans based on what they learn, over and over again during their years in school.

7.4.4 Targeted Feedback

Learning is enhanced when pupils see the effects of what they have tried and can envision alternative strategies to understand the material. Although assessment as learning is designed to develop independent learning, pupils cannot accomplish it without the guidance and direction that comes from detailed and relevant descriptive feedback from teachers to help them identify their learning needs and to develop autonomy and competence (Gipps et al. 2000; Clarke 2003). Students need feedback not just about the status of their learning but also about the degree to which they know when, where and how to use the knowledge they are learning (Bransford et al. 1999). Effective feedback challenges ideas, introduces additional information, offers alternative interpretations and creates conditions for self-reflection and the review of ideas. Pupils can apply these approaches for themselves to monitor their own learning, think about where they feel secure in their learning and where they feel confused or uncertain and decide on a learning plan. In so doing, pupils are encouraged to focus their attention on the task rather than on getting the answer right, and they develop ideas for adjusting, rethinking and articulating their understanding.

7.4.5 Discussion, Challenge and Reflection

As Vygotsky (1978) argued, the capacity to learn from others is fundamental to human intelligence. With help from someone more knowledgeable or skilled, the learner is able to achieve more than she or he could achieve alone. Ideas are not transported 'ready-made' into students' minds. Instead, as the Jojo story showed, new ideas emerge through careful consideration and reasoned analysis and just as important, through interaction with new ideas from the physical and social worlds. Learning is not private, and it is not silent. It may happen in individual minds, but it is constantly connected to the world outside and the people in that world. Peers and parents can be strong advocates and contributors to this process, not as judges, meeting out marks or favours, but as participants in the process of analysis, comparison, rethinking and reinforcing that makes up learning. Learning is a social activity. Teachers, peers and parents, when they understand their role, and the situation is structured to support the process, can be key players as learners grapple with 'what they believe to be true' in relation to the views, perspectives and challenges of others.

7.4.6 Practice, Practice, Practice

Independence in monitoring learning is not something that just occurs. It does not happen immediately, and there may be setbacks along the way. Even those with natural talent require a great deal of practice in order to develop their expertise. But practice is more than repetitive drills. Modern theories of learning and transfer retain the emphasis on practice, but they specify the kinds of practice that are important and take learner characteristics (e.g. existing knowledge and strategies) into account. Learning and transfer is most effective when people engage in 'deliberate practice' that includes active monitoring of their learning experiences (Bransford et al. 1999). When teachers involve pupils and promote independence, they are making their students responsible for their own learning and giving them the tools to undertake it wisely and well, by allowing them to experiment with new ideas, try them on, see how they fit, struggle with the misfits and come to grips with them. Effective problem solvers monitor their own mental progress as they plan for, execute and reflect on a learning task, and learners need opportunities to talk aloud overtly about what is going on inside their heads. This requires many opportunities to practise, reflect, challenge and try again.

7.4.7 An Environment of Emotional Safety

Becoming independent and responsible learners who embrace assessment as a positive part of the process is not something that comes easily. In fact, it is downright scary for many adults, let alone young people. It is no surprise that some (perhaps many) students do not wholeheartedly embrace the idea. The extent to which pupils are willing to engage in self-assessment is very much connected to their sense of self and their self-esteem. Persistence depends on expectations of success, even if it is not immediate. However, pupils who have had a history of, or fear, failure will adopt techniques to protect themselves, even if it means avoiding opportunities for learning. Pupils who define themselves by their ability are often dependent on high grades as a visible symbol of their worth and find the challenge of moving away from their positions of confidence rather like a free fall into the unknown. It is not enough to have a few safe moments or episodes of learning. These need to be the norm. Through detailed case studies of individual children throughout their primary schooling, Pollard and Filer (1999) demonstrate how these pupils continuously shaped their identities and actively evolved as they moved from one classroom context to the next. What this means is that each child's sense of self as a pupil can be enhanced or threatened by changes over time in their relationships, structural position in the classroom and relative success or failure. Their sense of self was particularly affected by their teachers' expectations, learning and teaching strategies, classroom organisation and criteria for evaluation.

If students are going to feel at ease with self-monitoring and self-regulation, they need to be comfortable with identifying different possibilities; they need to learn to look for misconceptions and inaccuracies in their own thinking and work towards a more complete and coherent understanding. Students (both those who have been successful – in a system that rewards safe answers – and those who are accustomed to failure) are often unwilling to confront challenges and take the risks associated with making their thinking visible. Teachers have the responsibility for creating environments in which students can become confident, competent self-assessors by providing emotional security and genuine opportunities for involvement, independence and responsibility.

Images and Points for Reflection

Changing assessment to capitalise on its power to enhance learning can be a fundamental shift in the preconceptions that teachers have about assessment – about what it is for, how it is connected to learning and how it works. In fact, shifting to routines in the classroom where assessment is used to help pupils monitor and regulate their own learning requires that teachers draw on their personal metacognitive skills and engage in a process of rethinking their assessment and teaching practices. Teachers, like students, may need help, feedback and reflection so that they can try out and adapt their newly acquired skills and knowledge in new environments. And they need images of how assessment can contribute to student reflection and self-regulation. We have included three examples to stimulate thinking about what using assessment for self-monitoring and self-regulation might look like and as a starting point for creating others.

7.4.8 Image 1 – Primary Mathematics

A primary teacher has been teaching the concept of two-digit addition with regrouping. She uses a worksheet that includes a range of items (such as single-digit additions without regrouping, single-digit with regrouping, double-digit additions with and without regrouping and tricky additions). These items enable her to become an investigator, making inferences and establishing hypotheses about what different pupils understand and what is still unclear or even inaccurate in their conceptions. After she analyses the pupils' work, she conducts a 'think aloud' with the class for each of the items, in which she models her thinking as she attempts each question. In this way, she provides them with insights about the correct approach as well as indicating the kinds of misconceptions and errors that might creep into someone's thinking. Finally, she does individual 'think alouds' with selected students, in which they tell her what they were thinking as they did particular questions (that she identified from their pattern of errors) so that she can help them see where their thinking needs some adjustment or practice. These targeted moments of reflection and rethinking on the part of individual pupils also provide information that forms the basis for the next stage in teaching and the grouping of pupils.

Points for Reflection

1. What content knowledge does this teacher need to construct this assessment?
2. What predictable patterns of errors would the teacher look for in analysing the students' work?
3. How has the teacher created opportunities for individual students to see their own thinking, reflect on it and make adjustments? What other strategies might she use now that she has additional information about their thinking?

7.4.9 Image 2 – Middle Years Social Studies

One of the history curriculum targets for pupils in key stage 3 is ‘organisation and communication’. This overarching objective includes several sub-items: recall, prioritise and select historical information; accurately select and use chronological conventions; and communicate knowledge and understanding of historical events. Within ‘recall, prioritise and select historical information’ alone, there are five additional sub-items: organising information; using a range of sources of information; finding relevant information; sorting, classifying and sequencing information; and comparing/contrasting information. The possibility of gaps in knowledge, underdeveloped skills, misunderstandings or misconceptions and confusion for pupils is massive. If teachers are serious about assessment for learning, every assessment task (and there will be many, both formal and informal) should provide insight into different pupils’ status in relation to organisation and communication of history and give pupils the reference points and the exemplars to allow them to reflect on their own thinking. Each assessment should explicitly focus on a subset of the skills, understanding, conventions, etc., that make up the overall curriculum expectation. And the teacher’s job is not just to score the assignments; rather she or he takes each assignment and, over time, constructs and continually adjusts the profile of learning and of teaching for each pupil in order to move their learning forward in effective and efficient ways.

Points for Reflection

1. What are the likely gaps in prior knowledge, areas of difficulty, misconceptions and challenges that students are likely to exhibit in relation to organising information; using a range of sources of information; finding relevant information; sorting, classifying and sequencing information; and comparing/contrasting information?
2. Design an assessment task for ‘recall, prioritise and select historical information’ that allows students to make decisions about their own knowledge and skill in relation to organising information; using a range of sources of information; finding relevant information; sorting, classifying and sequencing information; and comparing/contrasting information.

7.4.10 Image 3 – Middle Years Mathematics

At the beginning of the school year, a middle-school mathematics teacher uses a series of ‘games’ that he has devised to give him insights into his pupils’ knowledge and depth of understanding of concepts in the mathematics curriculum. One of these games uses a modified pool table to help him ascertain the pupils’ conceptions of algebraic relationships, either formal or intuitive. Pupils were given a graphic of a four-pocket pool table. They were told that the ball always leaves pocket A at a 45°

Table 7.1 Pool table dimensions and observations

Length	Width	Number of hits	Number of squares
6	4	5	12
3	5	8	

angle, rebounds off a wall at an equal angle to that at which the wall was struck and continues until it ends up in a pocket. Pupils counted the number of squares the ball passed through as well as the number of hits the ball made, the first and last hit being the starting and finishing pockets. They experimented with tables of various dimensions and recorded their observations on a chart (see Table 7.1).

As the pupils gathered data (with many more data combinations than we have included in the table), they began to make predictions about the number of hits, the number of squares and the destination of the ball, based on the patterns that they observed. Some moved to general statements of relationships like ‘You can tell the number of hits by adding the width and the length together and dividing by their greatest common factor’. Or, ‘The number of squares that the ball goes through is always the lowest common multiple of the width and the length’. Other students continued to count to reach the answers without seeing the relationships that existed.

During this task, the teacher wandered around the room observing and noting the thinking that was occurring for individual pupils. He stopped and asked questions, not about the answers that they were recording, but about the process that they were using. He prompted them to think about the patterns and to take a chance at making predictions. All the while, he was making notes on a pad that contained the names of the students and blank fields for writing his observations. From this information, he decided how to proceed in teaching the next series of lessons and how to group the class for the various teaching elements to come. For some, the work progressed very quickly to an introduction of formal notation of an algebraic equation to symbolise the general patterns that they had identified. For others, he used a number of patterning exercises to help them see the patterns that arose and formulate them in very concrete ways. He was very conscious of the importance of moving from concrete experience and direct consciousness of the phenomenon to the more abstract representation. The pool table task gave him a window into his pupils’ thinking and a starting place for planning teaching, resources, grouping, timing and pacing. When he moves on to another concept, all of these are likely to change. Once again, he will need to find out what the students see, what they think and what they understood before he decides what to do.

Points for Reflection

1. What mathematical content knowledge did this teacher need to have to design this task? To learn from this task?
2. What are the patterns of prior learning that he would be looking for in his students’ thinking?

7.5 Further Reading

The following resources may be useful for teachers in their study and implementation of classroom assessment with purpose in mind. This list is not exhaustive. Instead, it includes examples of books, articles, materials and web links that can be the starting point for individuals and groups to build their own personalised assessment resource compendia.

Active Learning Practice for Schools: Teaching for Understanding. Online, available at: learnweb.harvard.edu/alps/tfu/index.cfm (accessed 11 July 2007).

Airasian, P.W. (1999) *Assessment in the Classroom: a concise approach* (2nd edn), New York, NY: McGraw-Hill.

Arter, J. and Busick, K. (2001) *Practice with Student-Involved Classroom Assessment*, Portland, OR: Assessment Training Institute.

Arter, J. and McTighe, J. (2001) *Scoring Rubrics in the Classroom*, Thousand Oaks, CA: Corwin.

Association for Achievement and Improvement through Assessment. Online, available at: www.aaia.org.uk (accessed 11 July 2007).

Black, P. and Harrison, C. (2001) 'Feedback in questioning and marking: the science teacher's role in formative assessment', *School Science Review*, 82: 55–61.

Black, P., Harrison, C., Lee, C., Marshall, B. and Wiliam, D. (2003) *Assessment for Learning: putting it into practice*, Maidenhead: Open University Press.

Blythe, T., Allen, D. and Schieffelin, P.B. (1999) *Looking Together at Student Work: a companion guide to assessing student learning*, New York, NY: Teachers' College Press.

Earl, L. (2003) *Assessment as Learning: using classroom assessment to maximize student learning*, Thousand Oaks, CA: Corwin.

Gipps, C., McCallum, B. and Hargreaves, E. (2000) *What Makes a Good Primary School Teacher? Expert classroom strategies*, London: RoutledgeFalmer.

Gregory, G. and Kuzmich, L. (2004) *Data-Driven Differentiation in the Standards-Based Classroom*, Thousand Oaks, CA: Corwin.

Griffin, P., Smith, P. and Martin, L. (2003) *Profiles in English as a Second Language*, Clifton Hill, Victoria, BC: Robert Andersen and Associates.

Griffin, P., Smith, P. and Ridge, N. (2001) *The Literacy Profiles in Practice: toward authentic assessment*, Portsmouth, NH: Heinemann.

Joint Committee on Standards for Educational Evaluation (2000) *The Student Evaluation*, Kalamazoo, MI: The Evaluation Center, Western Michigan University.

Little, J.W., Gearhart, M., Curry, M. and Kafka, J. (2003) 'Looking at student work for teacher learning, teacher community, and school reform', *Phi Delta Kappan*, 85: 185–92.

National Research Council (1999) *How People Learn: bridging research and practice*, Committee on Learning Research and Educational Practice, Washington, DC: National Academy Press.

Rolheiser, C., Bower, B. and Stevahn, L. (2000) *The Portfolio Organizer: succeeding with portfolios in your classroom*, Alexandria, VA: Association for Supervision and Curriculum Development.

Acknowledgement This chapter is a reprint of Earl, L., & Katz, S. (2008). Getting to the core of learning: Using assessment for self-monitoring and self-regulation. In S. Swaffield (Ed.), Unlocking assessment (Chapter 6, pp. 90–104). Abingdon, England; New York: Routledge. Permission by Taylor & Francis/Routledge for reprint of this material is gratefully acknowledged.

References

Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box*. Cambridge: University of Cambridge School of Education.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*, 7–74.

Black, P., & Wiliam, D. (1998b). *Inside the black box*. London: King's College.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Brown, A. (1987). Metacognition, executive control, self-regulation, and mysterious mechanisms. In F. Weinert & R. Kluwe (Eds.), *Metacognition, motivation, and understanding*. Mahwah: Erlbaum.

Clarke, S. (2003). *Enriching feedback in the primary classroom*. London: Hodder and Stoughton.

Costa, A. (2006). *Developing minds: A resource book for teaching thinking* (3rd ed.). Alexandria: Association for Supervision and Curriculum Development.

Costa, A., & Kallick, B. (2000). *Activating and engaging habits of mind*. Alexandria: Association for Supervision and Curriculum Development.

Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438–481.

Delors, J., Al Mufti, I., Amagi, A., Carneiro, R., Chung, F., Geremek, B., Gorham, W., Kornhauser, A., Manley, M., Padrón Quero, M., Savané, M-A., Singh, K., Stavenhagen, R., Suhr, M. W., & Nanzhao, Z. (1996). *Learning: the treasure within.* Report to UNESCO of the International Commission on Education for the twenty-first century, Paris: UNESCO.

Earl, L. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks: Corwin.

Earl, L., & Katz, S. (2005). *Rethinking assessment with purpose in mind*, Western and Northern Canadian Protocol for Collaboration in Education.

Earl, L., & Lee, L. (2000). Learning, for a change: School improvement as capacity building. *Improving Schools, 3*, 30–38.

Gipps, C., McCallum, B., & Hargreaves, E. (2000). *What makes a good primary school teacher? Expert classroom strategies*. London: RoutledgeFalmer.

Katz, S. (2000). Competency, epistemology and pedagogy: Curriculum's Holy trinity. *Curriculum Journal, 11*, 133–144.

Katz, S., Sunderland, S., & Earl, L. (2002). Developing an evaluation habit of mind. *Canadian Journal of Program Evaluation, 17*, 103–119.

King, P. M., & Kitchener, K. S. (1994). *Developing reflective judgement: Understanding and prompting intellectual growth and critical thinking in adolescents and adults*. San Francisco: Jossey-Bass.

Marzano, R., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning mode*. Alexandria: Association for Supervision and Curriculum Development.

Newmann, F. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. San Francisco: Jossey-Bass.

Pollard, A., & Filer, A. (1999). *The social world of children's learning*. London: Cassell.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Stiggins, R. (2001). *Student-involved classroom assessment*. Upper Saddle River: Merrill Prentice Hall.

Vygotsky, L. S. (1978). *Mind in society: The development of the higher psychological processes*. Cambridge, MA: Harvard University Press (Originally published in 1930, Oxford University Press, New York edn.)

Weiner, B. (2000). Intrapersonal and interpersonal theories of motivation from an attributional perspective. *Educational Psychology Review, 12*, 1–14.

Chapter 8
Metacognitive Self-Confidence in School-Aged Children

Sabina Kleitman, Lazar Stankov, Carl Martin Allwood, Sarah Young, and Karina Kar Lee Mak

8.1 Introduction

This chapter focuses on the self-confidence construct that belongs to a broader area of metacognition. Metacognition refers to 'knowing about knowing' (Metcalfe and Shimamura 1994). An aspect of the second 'knowing' refers to one's understanding of different task-related factors, such as the state of one's knowledge and abilities. The first 'knowing' represents the awareness of this understanding.

Most theories distinguish between two major components of metacognition—knowledge about cognition and regulation of cognition (Nelson and Narens 1994; Schraw and Dennison 1994). Knowledge of cognition consists of different sets of beliefs one holds about oneself. They include (but are not limited to) beliefs about how effective one is as a learner and factors that influence one's own performance. This information may assist a person in the successful planning of his or her learning. For instance, if the learners are aware that they excel in understanding the logic behind rules but struggle with simple memorisation of the material, to optimise their performance, they should focus on acquiring understanding of the principles of the

S. Kleitman (✉) • S. Young • K.K.L. Mak
School of Psychology, University of Sydney, Sydney, NSW, Australia
e-mail: sabinak@psych.usyd.edu.au; syou5950@uni.sydney.edu.au; karinamak@gmail.com

L. Stankov
National Institute of Education, Jurong West, Singapore

Centre for positive Psychology and Education, School of Education, University of Western Sydney, Sydney, NSW, Australia
e-mail: lazondi@rocketmail.com

C.M. Allwood
Department of Psychology, University of Gothenburg, Göteborg, Sweden
e-mail: cma@psy.gu.se

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_8, © Springer Science+Business Media Dordrecht 2013

studied phenomenon rather than trying to memorise the outcomes of it. Reflections on who one is as a learner may also assist in realising what aspirations, expectations and evaluations to hold with respect to one's own performance. In this chapter, we tap into the processes of metacognitive knowledge by assessing the metacognitive beliefs of children. In particular, we focus on children's perception of competency in their fundamental cognitive abilities of memory and reasoning. We also assess children's academic self-efficacy beliefs.

Another focus of this chapter is the metacognitive experience of the feeling of confidence, which is a part of a broader domain of regulation of cognition—that is, self-monitoring of cognition (Efklides 2001, 2006; Schraw and Moshman 1995). Self-monitoring is defined as the ability to watch, check and appraise the quality of one's own cognitive work in the course of doing it (Schraw and Moshman 1995). Confidence judgments reflect this activity since they evoke subjective feelings of certainty that one experiences in connection with answering a question or regulating one's actions (Allwood and Granhag 1999; Koriat and Goldsmith 1996; Stankov 1999). The level of confidence informs the learner about the quality of their performance, allowing the learner to regulate performance and learning strategies. That is, during a test, the student may utilise this information to decide whether to move to the next task/item or stay on the present one, until they are reasonably confident that their answer is the correct one. Similarly, knowledge about which items generated a small degree of confidence may assist the learner to focus on the material in which they lacked confidence.

Metacognitive knowledge and experiences are essential components of successful self-regulated learning practices since they can inform the choice of learning strategies, provide for their adjustment and, when necessary, adjust the expectations/evaluations of one's own performance (Schraw et al. 2006; Sternberg 1997). This chapter provides a broad literature review based on a series of studies conducted with primary school children in Australia and Sweden and aims to answer three questions: (1) What is the best way to measure self-confidence? (2) What are the key factors at school, after school and at home that predict confidence levels in the cognitive domain? (3) What role do self-beliefs play in predicting confidence levels?

8.2 Self-Confidence as an Aspect of Metacognitive Experiences

The procedure commonly used in research for assessing the self-confidence construct is integrated within the typical test-taking or decision-making activity. Immediately after responding to an item in a test, participants are asked to give a rating, indicating how confident they are that the chosen answer is correct. In other words, participants are instructed to give a confidence (or 'sureness') rating indicating how confident/sure they are that their chosen answer is the correct one. The level of confidence is expressed in terms of percentages and/or verbal statements. The starting point (the lowest confidence) on a rating scale is defined in terms of

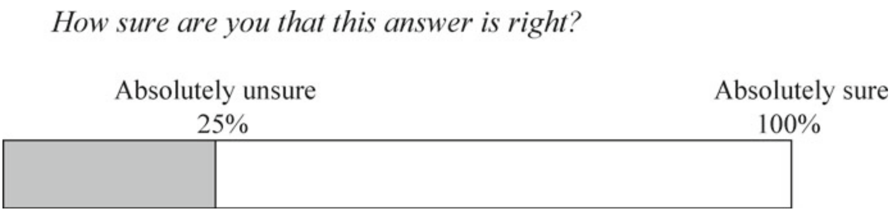


Fig. 8.1 Line confidence rating scale for tests containing multiple-choice questions with four response choices

the number of alternative answers (*k*) given to a question. That is, in multiple-choice questions with five alternative answers, 20% is a starting point because 20% is the probability of answering the question correctly by chance (100/*k*). Younger participants can also be given simple verbal keys to assist their understanding of the confidence scales. Consequently, the confidence scale may include both percentages and labels (e.g. ‘guessing’, ‘fairly sure’, ‘absolutely certain’). Sometimes, confidence scales that contain only verbal descriptors are used, but this practice obviously makes it difficult to evaluate metacognitive performance in a more exact quantitative way. For example, for the four choice multiple-choice question, at the 25% guessing level, the verbal scale may state ‘25% Absolutely unsure (correct 25 times out of 100)’, while for 100% level, the verbal scale may state ‘100% Absolutely sure (correct 100 times out of 100)’.

As measurement of confidence judgments is related to the theory of probability, it is essential that participants understand this concept. Allwood and colleagues (2006) examined four different types of confidence scales with children aged 11 and 12 years. The scales were (a) Numeric, (b) Picture, (c) Line and (d) Verbal. There were no differences between these scales in levels of confidence, suggesting their equivalence in their ability to capture confidence levels and the adequacy of this procedure with children of this age group. The example of the line scale is provided in Fig. 8.1.

The research by Allwood et al. (2006) also provided evidence that the children were not less skilled than adults in using the confidence scale. Here, it is of relevance that Erev et al. (1994) suggested that the within-subject standard deviation of the confidence judgments can be used as an indicator of the presence of noise in the confidence judgments. In general, the presence of noise (non-relevant factors affecting the measurement) is expected to be bigger for more difficult tasks. If the children had been less skilled in using the confidence scale, it is likely that the error variance for their confidence judgments would have been higher than that of the adults. In order to investigate this issue, Allwood et al. (2006), for each of the four confidence scales, compared the error variance within the children’s individual confidence judgments with the corresponding variance in the confidence judgments of a group of adults in a comparable study. The result was that the children did not show higher error variance than the adults for any of the four scales

Subsequently, Allwood and co-workers (2008), using similar methods, demonstrated that children as young as 8–9 years show comprehension of the numerical scale.

In addition, these authors also showed that the 8–9-year-old children were able to give confidence judgments at a level that more or less perfectly mirrored the level of correctness of the specific assertions in their memory recall of an event that they had experienced 1 week earlier. However, this was only the case when they answered an open, free-recall question ('Tell me everything you can remember about the event'). When the children's confidence rated the correctness of their answers to questions posed by another person on specific details of the experienced event, they showed overconfidence bias. That is, their average confidence was higher than the accuracy of their answers. Moreover, the error variance within the children's individual confidence judgments did not differ from the corresponding variance for the adults in this study. This reassures that children as young as eight, as well as adults, understand and utilise well-validated confidence measurement scales. Additional reviews of research on confidence for episodic memory in the calibration tradition are given in Allwood (2010a, b). The studies reviewed in this chapter use two versions of scales (pictorial and line) as evaluated by Allwood et al. (2006).

These confidence ratings immediately follow the cognitive act of providing responses to the typical cognitive test items, rather than relying on a general *perception* of one's own way of acting. As such, these confidence ratings serve as a more accurate measure of self-confidence than the general self-report items such as 'I feel self-assured' and 'I'm self-confident' that rely on Likert scales (Stankov 1999; Kleitman 2008). It is important to note that studies with adult samples indicate limited or no relationships between confidence levels and personality factors which include this type of self-report questionnaire, for example, extroversion (e.g. Dahl et al. 2010). The only exception to this is the openness to experience dimension which shares a positive correlation of low to moderate size (rarely above .30) with these on-task, on-line measures of confidence (see Kleitman 2008, for a review).

The role of confidence judgments in academic work and in everyday memory use has become more comprehensible by the memory model presented by Koriat and Goldsmith (1996). In this model, confidence judgments are an integral part of ordinary memory retrieval and reporting. Three phases are assumed in the model: *retrieval*, that is, activation of information in memory; *automatic monitoring*, that is, evaluation of the correctness of the retrieved information; and finally, *control*, that is, a decision with respect to whether the retrieved information should be reported or not. The control phase is especially relevant in the present context, where it is assumed that the rememberer uses the spontaneously generated confidence judgments to regulate which retrieved memories to report. Thus, when a person can choose what information to report, they can regulate whether the information should be reported or not. This would depend on how confident they are about the memory and on the basis of how important they think it is that they are correct in their current social context. For example, when speaking to a friend, the child may use a lower criterion for what to report than when speaking to a teacher or when giving a testimony in court.

In a study testing this memory model, Koriat et al. (2001) found that when they were given the possibility to choose which questions to answer, 7–12-year-old children were also able to improve the accuracy level of their answers to questions on the content of a slideshow that they had seen earlier.

8.2.1 Self-Confidence Trait in Adults

Confidence judgments have high reliability—both test–retest (Jonsson and Allwood 2003) and internal consistency (e.g. Kleitman 2008; Stankov 1999; Stankov and Lee 2008). There is much empirical evidence attesting to individual differences in confidence ratings in adult populations (see Kleitman 2008; Stankov and Lee 2008). That is, the correlations between accuracy and confidence scores from the *same* test are significant (average between .40 and .50). However, correlations between confidence ratings from a broad battery of diverse cognitive tests have been consistently high enough to define a strong, broad self-confidence factor. This reflects the *habitual* way in which adults assess the accuracy of their cognitive decisions across a diverse variety of cognitive stimuli. That is, adults who are more confident on one task (e.g. general knowledge tests), relative to their peers, also tend to be more confident across other tasks (e.g. math achievement, tests of reasoning or different perceptual tasks). In other words, regardless of the nature of cognitive stimuli, the relative ranking of self-assessment of accuracy of one's own performance remains stable. Thus, the confidence levels converge to define a psychological trait which marks important metacognitive experiences (Kleitman and Stankov 2001, 2007; Stankov 1999; Stankov and Lee 2008).

8.2.2 Self-Confidence Trait in Children

Kleitman and colleagues conducted several studies to examine the generality of confidence levels in Australian children aged 9–13, using a variety of cognitive and achievement tests (see Kleitman et al. 2011, for a review). Again, the results in all our studies show high internal consistency reliability estimates for confidence ratings, ranging between .84 and .96.

Kleitman and colleagues (e.g. Kleitman and Moscrop 2010) employed factor analysis to examine the consistency of confidence judgments in children. Their results demonstrated that a self-confidence factor, similar to the one found among adults, exists in children as well. In other words, confidence judgments in children across different cognitive domains tend to define a single factor. Just as with adults, this factor belongs to the metacognitive realm (Kleitman and Moscrop 2010).

8.3 Importance of Self-Confidence

One might ask, why is this self-confidence factor of any importance? The answer was provided by several studies. Kleitman and Moscrop (2010) demonstrated that in primary school children (age range between 9 and 13 years), higher levels of confidence predicted higher grades after controlling for age, gender, intelligence, school fees and parent–child family dynamics. That is, teachers tend to assign higher grades to children who assessed their own performance more favourably compared to children who were less confident in their performance. This was true irrespective of child's age, gender, intelligence and other key factors. This attests that students with higher levels of confidence appear to be getting better reports from school, which most likely would positively influence their level of confidence. This cycle may continue, influencing children's and then adults' subsequent confidence, aspirations and performance.

8.4 Factors That Influence Self-Confidence

Since metacognitive experiences of self-confidence hold promise for improving learning outcomes, it is important to identify those factors that affect confidence levels. In our studies, we typically employ a variety of cognitive tests which capture different areas of learning (reading, writing and mathematics) and cognition (crystallised and fluid intelligence). Depending on the research design, these tests also assess confidence levels. Confidence ratings for all attempted test items are averaged to give an overall confidence score, which is used in statistical analyses and reports. Throughout this chapter, we use the term self-confidence to refer to the broad psychological trait which emerges from confidence scores on different tests when used together within the study. Before we can start exploring the factors that affect the self-confidence trait, it is necessary to point to an important distinction between *internal* (person-driven) and *external* (ecological) factors that influence metacognition. We shall first consider internal influences.

8.4.1 The Most Important Internal Factors: Self-Beliefs

Self-confidence denotes a psychological trait, thus there are stable *person-driven* factors in confidence ratings (see Kleitman 2008; Stankov 1999; Stankov and Lee 2008, for reviews). However, self-confidence, as we have mentioned above, is *not* strongly related to personality constructs (see Stankov 1999). What then underlies such stability?

Before attempting to answer this question, it is necessary to locate self-confidence within a broader domain of what is sometimes referred to as self-beliefs. Brief definitions and examples of measures of each self-belief construct are as follows:

(a) Metacognitive beliefs are a part of the knowledge aspect of metacognition. In this chapter, we refer to a specific subset of these beliefs, the students' perception of competency of their fundamental cognitive abilities, memory and reasoning. Example: 'I can remember more material than the average student' (memory competence) and 'To solve a problem, I rely on my good reasoning abilities' (reasoning competence).

(b) Academic self-concept refers to multidimensional and hierarchical self-beliefs that students hold with respect to their performance on traditional school curriculum subjects—reading and mathematics—and general school performance (Marsh 1988). Examples: 'I get good marks in reading' (reading self-concept), 'Work in mathematics is easy for me' (math self-concept), and 'I am good at all school subjects' (general school self-concept).

(c) Academic self-efficacy refers to the belief indicated that if one is engaged in a particular learning act/behaviour, one will achieve a positive and desired outcome within a specific learning task/domain (Bandura 1993). Example: 'Even if the work in school is hard, I can learn it'.

In adults, academic self-concepts (see Efklides and Tsiora 2002; Kröner and Biermann 2007) and metacognitive self-beliefs regarding competencies of one's own reasoning abilities positively predicted confidence levels after controlling for accuracy of performance (Kleitman 2008; Kleitman and Stankov 2007; Stankov and Lee 2008). Our findings indicate similar results with children. In particular, metacognitive beliefs in the competency of one's reasoning and memory abilities, together with academic self-concept and self-efficacy judgments, positively predict levels of confidence that children hold in their cognitive performance (Kleitman and Gibson 2011).

Our results also suggest that self-beliefs converge together to define the self-beliefs factor. That is, in primary school children, we found moderate to strong positive correlations between memory and reasoning and self-concept and academic self-efficacy (ranging between .46 and .68, $p < .01$) (Kleitman and Gibson 2011). Moreover, a factor defined by measures of self-beliefs explained about 70% of the total variance in these measures. This suggests that children who hold higher metacognitive beliefs about the competence of their cognitive faculties also hold strong beliefs about their academic self-efficacy.

This factor serves as both an important predictor of self-confidence and a key mediator of the predictions that the other variables have on self-confidence. In particular, this self-belief factor predicted confidence levels regardless of a child's intelligence, gender, school fees and some key school factors that we overview below (Kleitman and Gibson 2011).

Importantly, our findings also show that there is a negative relationship between these self-beliefs and avoidance behaviours known as self-handicapping tendencies (Kleitman and Gibson 2011). Such behaviours include procrastination, generating excuses or staying up late before an exam. Self-handicapping strategies are often used deliberately, and they are detrimental to learning as they are linked to poorer exam performance, the use of surface-learning strategies and lower tendencies to self-regulate (Thomas and Gadbois 2007). Therefore, the fact that students with stronger self-beliefs were less likely to utilise self-handicapping behaviours is important. Thus, strong self-beliefs may be acting as a possible buffer against detrimental self-handicapping behaviours.

Furthermore, the self-belief factor also mediated relationships between certain key classroom environment variables and the self-confidence factor. That is, mastery goal orientation and self-efficacy of the teacher positively predicted metacognitive beliefs, which in turn predicted the self-confidence factor. Thus, in light of the self-concept and self-efficacy theories, metacognitive beliefs serve as both a predictor and a mediator variable of the predictions that the other variables have on self-confidence.

Overall, the results of our studies demonstrate that self-beliefs play a key role in academic settings: they predict positively higher confidence levels and reduce detrimental learning-avoidance behaviours. They also mediate the predictions that other variables have on the self-confidence factor.

8.4.2 Predictive Validity of Self-Confidence Versus Other Self-Beliefs

In the paragraphs above, we reviewed some of the evidence that point to the importance of self-beliefs. Given that there are important links between self-beliefs and self-confidence, it is legitimate to compare the power of each set of beliefs for predicting educational outcomes. Recent work in education indicates a particularly potent role of internal student variables in predicting academic achievement. Three self-belief factors—self-concept, self-efficacy and mathematics anxiety—emerged in the work of Lee (2009), who reported findings from the Programme for International Student Assessment (PISA) 2003 data that is based on 15-year-olds from 41 countries. Examples of measures of each self-belief construct in this work are as follows:

(a) Self-concept. Example: 'In my Mathematics class, I understand even the most difficult work'.
(b) Self-efficacy. Example: 'I am sure I can do difficult work in my English/ Mathematics class'.
(c) Anxiety—one's physio-emotional reactions when she/he thinks about or performs a task. Example: 'I often worry that it will be difficult for me in Mathematics classes'.

Although there is a large body of literature on all three self-belief factors on their own and Lee (2009) reports significant correlations of each one of these with measures of achievement in mathematics, little has been known about the relationship between these constructs and self-confidence. Current data from Singapore and several other Confucian Asian and European countries indicates that (a) self-confidence indeed correlates with all three self-concepts listed above and defines a common factor that has a significant correlation with accuracy of cognitive performance, and (b) confidence is by far the best single predictor of accuracy of cognitive performance. Furthermore, in most data sets available by now, it absorbs the predictive variance of the three other self-constructs listed above when they are considered as separate predictors of accuracy (Stankov et al. in press). This second point is particularly important as it suggests that our procedure of measuring self-confidence absolves the researcher from employing separate scales of self-efficacy, self-concept and anxiety.

An additional, very important property of the self-confidence construct derives from its broadness as a factor. The other three self-beliefs listed above are said to be domain-specific, implying that, for example, a specific subject area—mathematics in the above examples—is not related to the same constructs in other areas such as English. Thus, the self-confidence measure obtained from an English test predicts self-confidence obtained from a mathematics test. Furthermore, given the link between self-confidence and performance mentioned in point (b) above, one can use self-confidence scores from the English test to predict achievement scores in mathematics! Only a few constructs in psychology show similar properties, with IQ measures being the best known example.

8.4.3 Important External Factors: Classroom Environment and Out-of-School Influences

Among the external factors that may influence self-confidence, the three sources we have investigated are (1) classroom environment, (2) after-school environment and (3) parental influences. In this section, we review evidence that points to the role of these influences in the development of self-confidence.

8.4.4 Classroom Environment

There is evidence that positive relationships with teachers influence the learning habits and academic aspirations of children (Burchinal et al. 2002). Although the nature of the relational bond is different between parent–child and teacher–child interactions, the essence of the relationship is similar: caring, closeness, warmth and open communication (e.g. Crosnoe et al. 2004).

Social self-efficacy beliefs with the teacher refer to how competent a student feels about communicating with and relating to their teacher (Schunk 1989). For instance, items that capture this construct are ‘I can explain my point of view to my teacher’ and ‘I can get my teacher to help me when I have problems with other students’. The ability to interact effectively with the teacher is likely to play a facilitative role in fostering self-beliefs (Kleitman and Gibson 2011; Patrick et al. 1997). That is, students who communicate effectively with their teachers may feel more comfortable in asking questions and receiving feedback, helping them to acquire a variety of useful cognitive and metacognitive information (e.g. about strategies, their cognitive strengths and weaknesses, and the ways in which they can approach tasks).

This reasoning is supported by the memory model presented by Koriat and Goldsmith (1996). Given that children with high confidence employ the same control criterion as other children for when to report retrieved information, they are likely to report more of their retrieved information because they will, due to their high self-beliefs, on average, feel more confident about the retrieved information than other children.

Classroom goal orientation—when teachers influence student perceptions of the purposes of achievement behaviour—is another key factor in the school environment (Ames 1992). Mastery-oriented classrooms encourage the attribution that effort leads to success, and emphasise developing new skills, understanding concepts and improving competence, thereby highlighting the intrinsic value of learning (Ames 1992). In contrast, performance-oriented classrooms emphasise student ability, as demonstrated by outperforming others or surpassing normative standards. The results demonstrated that a mastery goal orientation is associated with a stronger self-belief factor (Kleitman and Gibson 2011). However, we found no relationship between performance goal orientation and self-beliefs, and self-confidence and achievement factors (Kleitman and Gibson 2011). This suggests that while explicitly emphasising student ability and achievement appears to do no harm, it seems to have no benefit for metacognitive variables or achievement. In contrast, a mastery (rather than performance) orientation in the classroom appears to foster stronger self-concept, which in turn predicts higher confidence levels.

8.4.5 After-School Environment

Previous studies have found benefits of extracurricular activities on academic performance (Whitley 1999). When we investigated after-school activities of children, we categorised these into several areas: total time for sport, leisure activities and time with adults (Kleitman et al. 2011; Lau 2009; Mak 2009; Young 2009). Although giving many positive predictions for the other variables in our study (e.g. physical competence, physical self-concept), time spent on sport inside and outside of school did not predict the cognitive self-confidence factor or academic achievement (Kleitman et al. 2011).

8.4.6 Parental Role

Previous research demonstrates that children can develop metacognitive skills through early interactions with parents (Neitzel and Stright 2003). The ‘time spent with adults’ variable we utilised included talking, spending time and/or doing activities with mother, father or other adults (combined time in hours). Our results indicate that time spent with adults positively predicted the academic self-concept factor, which in turn positively predicted higher confidence levels and achievement (Kleitman et al. 2011). This finding aligns with the theory that involvement of parents/adults in a child’s time outside of school encourages the child’s motivation, forming positive attitudes towards school and learning.

8.5 Discussion

Metacognition is one of the three fundamentals of self-regulated learning, along with cognition and motivation (Schraw et al. 2006). Efficient test-taking behaviour and test-taking outcomes signify academic success. Moreover, the metacognitive confidence judgments which students assign to their ongoing performance are at the core of this test-taking behaviour. The studies reviewed in this chapter focused on two metacognitive variables—metacognitive beliefs and confidence ratings—indexing metacognitive knowledge and experience, respectively. We outlined the existence of the self-confidence trait in primary school children and the several ways to capture it using the confidence scales validated for such measurement (see Allwood et al. 2006, for a review).

The available evidence suggests that there is a *habitual* response pattern of confidence levels, or a trait, which is stable across different cognitive tasks. The picture of the resulting effect of, for example, having a pattern of high confidence, irrespective of the task, is somewhat complex. The general finding in the research on semantic knowledge that is likely to be tested in a school context is that children and adults show overconfidence (i.e. they are more confident than they are correct), but there is of course individual variation around this average tendency (Griffin and Brenner 2004; McClelland and Bolger 1994). When the knowledge level (i.e. accuracy) is constant, an extra addition of confidence in such situations is likely to result in increased overconfidence. However, for children with a consistently high level of confidence, an effect of this trait is that they, for example, in the school setting, will report more of their retrieved memory information. Additionally, as an effect of this, they will receive more constant feedback from teachers and parents. Such more pervasive feedback is likely to function as a more efficient influence of the student's academic behaviour, when compared with students with a lower level of confidence. These students may be likely to report less of their retrieved information from memory and receive more haphazard feedback from teachers and parents, both of which could lead to poorer academic performance. Students with high levels of confidence and who are provided with more specific and continuous feedback are more likely to demonstrate better academic performance. Research reported in this chapter showed that the self-confidence factor held a predictive power on academic achievement irrespective of a child's intelligence, gender, school fees and parental bonds (Kleitman and Moscrop 2010). Thus, it is important to understand what influences predicted levels of confidence. In this chapter, we focused on metacognitive beliefs as well as key external factors: dynamics at school and after school.

With respect to self-beliefs, the message is clear—they are an important positive predictor of confidence levels, such that children who hold higher self-beliefs also have higher confidence in their answers. Moreover, metacognitive beliefs mediate the predictions that other key educational variables have on confidence judgments. In particular, mastery goal orientation and self-efficacy with the teacher predicted the self-belief factor, which in turn predicted the self-confidence factor.

With respect to the key external factors, our results indicate that spending more time with parents, maintaining positive relationships with teachers and possessing a mastery goal orientation of the classroom positively predicted metacognitive beliefs, which in turn predicted confidence levels. Thus, parents (or other significant adults) and teachers are instrumental in supporting a child's development of strong self-beliefs.

Leisure time is also important for a child to establish a healthy routine in their lives and can pave the way for their future enjoyment of activities outside of school. Our results demonstrated that time spent with adults fosters higher levels of confidence and accuracy in primary school children, emphasising the significance of a holistic approach to a child's life, both inside and outside of school.

8.6 Practical Implications for Teachers and Parents

The studies reviewed in this chapter indicate that investment in the development of students' metacognitive beliefs and skills may advance performance in academic areas. Knowledge that a child as young as 9 years is already habitually assessing their own thinking is a crucial and powerful tool, one which can undoubtedly assist parents, teachers, school counsellors and child psychologists to foster self-regulated learning (see Schraw et al. 2006, for a review). For example, to foster accuracy of performance, many teachers repeatedly tell children to 'check their work'. Hence, the use of confidence ratings as a part of regular in-class practices could improve metacognitive skills in students and provide a more efficient way for teachers and parents to clarify the 'problem' areas. Teachers can also use self-monitoring measures as informal assessment to determine students' level of understanding in specific learning areas—literacy, mathematics and science.

Spending time with adults outside of school seems to have an important relationship with a child's academic development. Teachers and school counsellors often meet with parents to provide them with strategies they can use to assist their children with everyday homework, or for children who may experience difficulties with their learning. In such meetings, parents' understanding of their children's metacognitive skills should be promoted to empower parents' involvement in their children's metacognitive development. That is, knowledge of the child's cognitive strengths could suggest powerful directions for parents on how to assist the learning process. In addition, encouragements from parents like 'that was a good reasoning strategy you just used' or 'that was a nice memory recall' may foster the child's stronger self-beliefs, which predict better academic outcomes and stronger confidence.

Spending time with adults after school may also assist with the development of time management skills. For example, students in primary school (particularly in years 5–6) are expected to develop the skills of time management and prioritisation. Parents can assist their child's metacognitive development by scaffolding and modelling the creation of timetables, and use of homework diaries and lists, as well as specific metacognitive strategies for problem solving and memory recall in individual homework tasks. Students can learn to plan their homework completion according to difficulty, time and effort required. As a student undertakes such tasks, monitoring and evaluation strategies are inevitably used, so that the student reviews the advantages and possible disadvantages of their approach, for example, the detrimental effects of commencing a difficult task late at night after spending time on easier tasks. These skills are integral to a child's academic success in high school, and eventually, also in everyday adult life.

8.7 Conclusion

In summary, future studies in education should aim to include confidence assessment as both an important outcome and an important predictor of academic and physical achievements, especially if the self-regulated model of learning is adopted. These studies will require well-validated methods of assessment of confidence in cognitive domains: the methods which allow for both the reliable assessment of the confidence levels and the ability to immediately confirm their veracity.

References

Allwood, C. M. (2010a). The realism in children's metacognitive judgments of their episodic memory performance. In A. Efklides & P. Misailidi (Eds.), *Trends and prospects in metacognition research* (pp. 149–169). New York: Springer.

Allwood, C. M. (2010b). Eyewitness confidence. In P. A. Granhag (Ed.), *Forensic psychology in context* (pp. 281–303). Uffculme, Devon: Willan Publishing.

Allwood, C. M., & Granhag, P. A. (1999). Feelings of confidence and the realism of confidence judgments in everyday life. In P. Juslin & H. Montgomery (Eds.), *Judgment and decision making: Neo-Brunswikian and process-tracing approaches* (pp. 123–146). Mahwah: Lawrence Erlbaum Associates, Inc., Publishers.

Allwood, C. M., Granhag, P. A., & Jonsson, A. C. (2006). Child witnesses' metamemory realism. *Scandinavian Journal of Psychology, 47*, 461–470.

Allwood, C. M., Innes-Ker, Å., Holmgren, J., & Fredin, G. (2008). Children's and adults' realism in their event-recall confidence in response to free recall and focused questions. *Psychology, Crime & Law, 14*, 529–547.

Ames, C. (1992). Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*, 117–148.

Burchinal, M. R., Peisner-Feinberg, E., Pianta, R., & Howes, C. (2002). Development of academic skills from preschool through second grade: Family and classroom predictors of developmental trajectories. *Journal of School Psychology, 40*(5), 415–436.

Crosnoe, R., Johnson, M. K., & Elder, G. H., Jr. (2004). Intergenerational bonding in school: The behavioral and contextual correlates of student–teacher relationships. *Sociology of Education, 77*, 60–81.

Dahl, M., Allwood, C. M., Rennemark, M., & Hagberg, B. (2010). The relation between personality and the realism in confidence judgments in older adults. *European Journal of Ageing, 7*(4), 283–291.

Efklides, A. (2001). Metacognitive experiences in problem solving: Metacognition, motivation, and self-regulation. In A. Efklides, J. Kuhl, & R. M. Sorrentino (Eds.), *Trends and prospects in motivation research* (pp. 297–323). Dordrecht: Kluwer.

Efklides, A. (2006). Metacognitive experiences: The missing link in the self-regulated learning process. *Educational Psychology Review, 18*, 287–291.

Efklides, A., & Tsiora, A. (2002). Metacognitive experiences, self-concept, and self-regulation. *Psychologia: An International Journal of Psychology in the Orient, 45*, 222–236.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*, 519–527.

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 177–199). Malden: Blackwell Publishing.

Jonsson, A., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences, 34*, 559–574.

Kleitman, S. (2008). *Metacognition in the rationality debate. Self-confidence and its calibration*. Saarbrucken, Germany: VDM Verlag Dr. Mueller Aktiengesellschaft & Co. KG.

Kleitman, S., & Gibson, J. (2011). Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students. *Learning and Individual Differences, 21*, 728–735.

Kleitman, S., & Moscrop, T. (2010). Self-confidence and academic achievements in primary-school children: Their relationships and links to parental bonds, intelligence, age, and gender. In A. Efklides & P. Misailidi (Eds.), *Trends and prospects in metacognition research* (pp. 293–326). New York: Springer.

Kleitman, S., & Stankov, L. (2001). Ecological and person-driven aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15*, 321–341.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*(2), 161–173.

Kleitman, S., Mak, K., Young, S., Lau, P., & Livesey, D. (2011). Something about metacognition: Self-confidence factor(s) in school-aged children (pp. 103–115). In S. Boag & N. Tiliopoulos (Eds.), *Personality and individual differences: Theory, assessment, and application*. New York: Nova.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

Koriat, A., Goldsmith, M., Schneider, W., & Nakash-Dura, M. (2001). The credibility of children's testimony: Can children control the accuracy of their memory reports? *Journal of Experimental Child Psychology, 79*, 405–437.

Kröner, S., & Biermann, A. (2007). The relationship between confidence and self-concept – Towards a model of response confidence. *Intelligence, 35*(6), 580–590.

Lau, P. (2009). *Predicting achievement and self-confidence: Interpersonal and intrapersonal predictors in school-aged children.* Unpublished Hon thesis, The University of Sydney, Sydney, Australia.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences, 19*, 355–365.

Mak, K. (2009). *Metacognitive regulation in the physical domain: An investigation of school aged children's movement confidence and its relationship with self-concept and big five personality traits.* Unpublished Hon. thesis, The University of Sydney, Sydney, Australia.

Marsh, H. W. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology, 80*, 366–380.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980–1994. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). New York: Wiley.

Metcalfe, J., & Shimamura, A. P. (Eds.). (1994). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press.

Neitzel, C., & Stright, A. D. (2003). Mothers' scaffolding of children's problem solving: Establishing a foundation of academic self-regulatory competence. *Journal of Family Psychology, 17*(1), 147–159.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: The MIT Press.

Patrick, H., Hicks, L., & Ryan, A. M. (1997). Relations of perceived social efficacy and social goal pursuit to self-efficacy for academic work. *Journal of Early Adolescence, 17*, 109–128.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*, 460–475.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*, 351–371.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*, 111–139.

Schunk, D. H. (1989). Self-efficacy and achievement behaviours. *Educational Psychology Review, 1*(3), 173–208.

Stankov, L. (1999). Mining on the "no man's land" between intelligence and personality. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait and content determinants* (pp. 315–367). Washington: DC: American Psychological Association.

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology, 100*, 961–976.

Stankov, L., Lee, J., Morony, S., Luo, W. S., & Hogan, D. J. (in press). *Confidence: challenging the role of self-efficacy and self-concepts in education*.

Sternberg, R. (1997). *Thinking styles*. Cambridge, MA: University Press.

Thomas, C. R., & Gadbois, S. A. (2007). Academic self-handicapping: The role of self-concept clarity and students' learning strategies. *British Journal of Educational Psychology, 77*, 101–119.

Whitley, R. L. (1999). Those "dumb jocks" are at it again: A comparison of the educational performances of athletes and nonathletes in North Carolina high schools from 1993 through 1996. *The High School Journal, 82*, 223–233.

Young, S. (2009). *Examining the relationships between children's environments, self-concept, achievement and self-confidence.* Unpublished Hon. thesis, The University of Sydney, Sydney, Australia.

Part II
Tools for Implementing Self-Directed Learning Oriented Assessment

Chapter 9
Using Item Response Theory as a Tool in Educational Measurement

Margaret Wu

9.1 Introduction

The main objective of this chapter is to explain about the use of item response theory (IRT) in the area of educational measurement. Item response theory is also referred to as modern test theory (Crocker and Algina 1986), in contrast to classical test theory (CTT) which was developed before IRT. While IRT can be a stand-alone topic to study, a good understanding of IRT will include a clear view of what IRT adds to CTT. A comparison between IRT and CTT can highlight the advantages of IRT as well as the limitations of IRT. Further, IRT and CTT should be applied in a complementary way, rather than with the exclusion of one from the other. This chapter examines how both CTT and IRT can be used as a tool to build quality assessments, with an emphasis on the interpretation of IRT statistics in contrast to CTT statistics. Further, this chapter demonstrates how the software program, ConQuest (Wu et al. 2007), can be used to analyse item response data. Before we demonstrate the applications of CTT and IRT in item analysis, the following is a brief introduction to CTT and IRT.

9.1.1 Classical Test Theory

The fundamental concept of classical test theory (CTT), also known as true score theory, is that a test score consists of two components: a true score and an error component (e.g. Lord and Novick 1968). For example, Amy takes a test and obtains a score of 34 out of 50. One asks the question that if similar tests were taken by

M. Wu (✉)
Work-based Education Research Centre, Victoria University, Melbourne, Australia
e-mail: wu@edmeasurement.com.au

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_9, © Springer Science+Business Media Dordrecht 2013

Amy, what would be the variability in Amy's test scores, and what would be Amy's average test score. The average score is known as the "true score," and the variability in test scores on similar tests is attributed to an error component when each test is administered. Consequently, the variability in test scores provides an indication of whether the test scores are trustworthy or *reproducible,* or *reliable*. Therefore, a key concept of CTT is the computation of the *reliability* of a test. When the error component is small, the variability of test scores around the average score (true score) will be small (e.g. Amy's test scores on similar tests range between 33 and 36, which is a small range), and the test will have high reliability. In contrast, if Amy's test scores on similar tests range between 25 and 40, a somewhat large range, then the test has low reliability. More generally, CTT deals with test scores and statistics derived from test scores. There is no assumption that the test scores necessarily reflect some underlying latent ability. For example, under CTT, we don't assume that a student's score on a single mathematics test necessarily reflects a student's underlying mathematics ability over and beyond the set of questions included in the particular test. Nevertheless, when people use test scores, they often want to make inferences about a student's "ability" or proficiency in a more widely defined discipline for which the test is but a sample of items. There are a few problems in using test scores to infer some underlying (latent) ability. First, because test scores are bounded by 0 and the maximum score on a test, test scores, in theory, cannot be truly reflecting abilities which are supposed to be unbounded (i.e. there is no lower bound and upper bound for an underlying ability). Second, differences in test scores may not reflect the magnitude of differences in an underlying ability. For example, consider three students, Ann, Bev and Cath, obtaining 20, 30 and 40, respectively, out of a maximum of 50 on a test. While the differences in scores between Ann and Bev is the same as the difference between Bev and Cath, we can't make the assumption that a 10 score difference on the test has the same meaning for the difference in the underlying abilities. To overcome these difficulties with the interpretation of test scores, the development of item response theory (IRT) has taken on some momentum in the past six decades.

9.1.2 Item Response Theory

Item response theory, also called latent trait theory, first appeared in the work of Frederic Lord (1952) and Georg Rasch (1960). While classical test theory does not make any assumption about a postulated person attribute that determines performance on a test, IRT theorises a single proficiency variable, θ, often known as latent ability that underlies a person's performance on a test. That is, there is a notion of a distinctive trait or ability, although not directly observable, which can be used to predict how well a person will perform on a test designed to measure that ability. The more a person possesses this trait, the higher will be the person's expected score on this test. Furthermore, IRT posits a mathematical probabilistic model to make predictions of item responses for a person, where, in the case of the Rasch model, the probability of obtaining a correct answer is expressed as a function of the

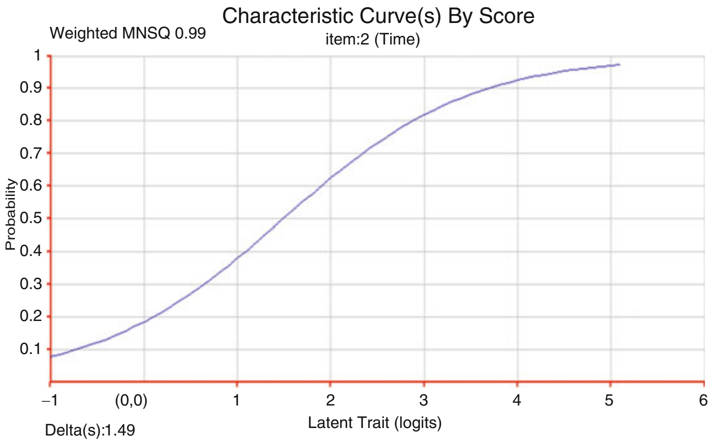


Fig. 9.1 Probability function of an item as a function of ability

difficulty of the item (*δ*) and the ability of the person (*θ*). Therefore, under IRT, the item responses are the units of analysis, while under CTT, the overall test scores are the units of analysis. As an example, Fig. 9.1 depicts the probability of success on an item as a function of a person's ability, using the Rasch model.

The horizontal axis in Fig. 9.1 is the ability scale, with low-ability persons located on the left and high-ability persons located on the right. The unit on this scale is known as logit (abbreviation for "log of odds unit"). The unit does not have any substantive meaning other than that it is a continuous numerical scale that shows measures from low to high. The item difficulty measure as defined in IRT is the ability at which a person has 50% chance of being successful on the item. In this example, we see that persons with an ability measure of 1.49 logits on the ability scale have a probability of 0.5 of obtaining the correct answer for this item. Therefore, the IRT item difficulty measure for this item is 1.49. The fact that item difficulty is defined on the ability scale is the key to many uses of IRT results. For example, if an item has a difficulty of 0.2 and a person has an ability of 0.8, then we can conclude that the person has more than 50% chance of being successful on this item because the person's ability is higher than the item difficulty. In fact, because there is a mathematical function for the probability of success, we can compute the probability that this person will obtain the correct answer on this item, as shown below:

$$p = P(X=1) = \frac{\exp(\theta - \delta)}{1+\exp(\theta - \delta)} = \frac{\exp(0.8-0.2)}{1+\exp(0.8-0.2)} = 0.65, \tag{9.1}$$

where θ is the ability of the student on the logit scale and δ is the difficulty of the item defined on this ability scale.

Consequently, once we know a person's ability and an item's difficulty, we can make statements about the chance that the person will be successful on the item. Such statements are not easily made under CTT, where test scores and item scores are the main statistics computed. For example, when we know a person's test score

Fig. 9.2 Data file: Test371.dat

```
202010414140410
322111424130233
212102424131013
224111424133323
202101424132233
212113424132233
212113424132233
212102424140030
212113424132233
204100414523030
203111424113292
..............................
```

(e.g. 70 out of 100) and an item's difficulty (e.g. 70% of the students obtained the correct answer), we cannot easily work out the person's chance of being successful on this item, since it depends on what other items are on the test and who else took the test.

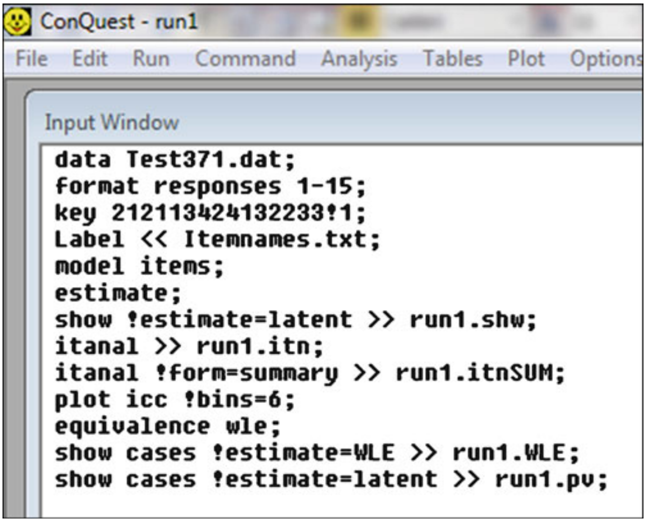
9.2 IRT Analysis of a Mathematics Test

In this section, we will use an example to illustrate the application of CTT and IRT to analyse item response data. The example data set was collected from an online administration of a Grade 5 mathematics test. There are 15 questions in the test. The test questions are in Appendix A.1. The IRT software, ConQuest (Wu et al. 2007), was used to analyse the item responses. The following shows a two-step process for carrying out the item analysis.

9.2.1 Step 1: Recoding Student Answers for Open-Ended Questions

Since item response modelling procedures can only deal with a limited number of response categories for each item, student answers to open-ended questions need to be recoded. For example, for Question 2, students could write down answers ranging between 0 and 60 min. After inspecting the frequencies of various answers to this question, a coding scheme was designed as follows: a response of 25 was coded 1, 30 was coded 2, 35 was coded 3, 40 was coded 4, and all other responses were coded 0. Missing responses were coded 9. A full list of the recodes for all 15 questions is shown in Appendix A.2. The recoding was carried out using the statistical package SPSS. After recoding the item responses, a data file was prepared. An extract of the data file, *Test371.dat*, is shown in Fig. 9.2.

The data file, *Test371.dat*, is a text file (or ASCII file). Each line contains 15 item responses of one student. For example, the first line contains the item responses of



```
data Test371.dat;
format responses 1-15;
key 212113424132233!1;
Label << Itemnames.txt;
model items;
estimate;
show !estimate=latent >> run1.shw;
itanal >> run1.itn;
itanal !form=summary >> run1.itnSUM;
plot icc !bins=6;
equivalence wle;
show cases !estimate=WLE >> run1.WLE;
show cases !estimate=latent >> run1.pv;
```

Fig. 9.3 ConQuest control file for the IRT analysis

Student 1, where the recoded response is 2 for Question 1, 0 for Question 2, 2 for Question 3, etc. In this data set, there are 3,930 students. So there are 3,930 lines in the data file.

9.2.2 Step 2: Run IRT Software ConQuest

To analyse the data, a "control file" is required to tell ConQuest where the data file is, how the data file is organised and how to score the item responses. An example ConQuest control file is shown in Fig. 9.3.

The ConQuest control file is typed in the ConQuest input window. A line-by-line explanation of the 13 lines of the control file is given in Table 9.1. More detailed descriptions of the output files from ConQuest are given in later sections.

9.2.3 Examine Item Difficulty Statistics

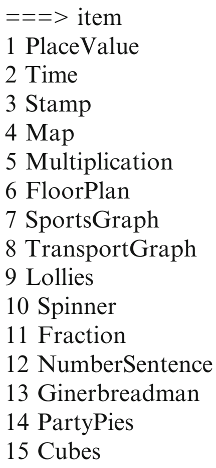
Once ConQuest is run, output files are produced. The output files, *run1.shw* and *run1.itn*, are text files so they can be opened with a text editor such as Notepad.

Under classical test theory, the facility of an item (i.e. percentage correct) is used as a measure of item difficulty. Table 9.2 shows a summary of item level statistics, including item facilities and IRT item difficulty parameter estimates, taken from the output file *run1.itnSUM,* sorted according to item facility.

Table 9.1 Explanations of ConQuest commands

data Test371.dat;	The data command specifies the location and the name of the data file
format responses 1–15;	The format command specifies the column range of the item responses
key 212113424132233!1;	The key command specifies the correct answer for each question so ConQuest can score the responses. For example, for Question 1, response "2" is the correct answer and should be scored 1. Other responses should be scored 0. For Question 2, response "1" is the correct answer and should be scored 1, etc.
Label << Itemnames.txt;	The label command specifies the name of the file containing item labels. The item labels file is shown in Fig. 9.4 Item labels file *Itemnames.txt*. This command is not necessary for the program to run. It is for the readability of the output where item names will be shown, in addition to item numbers
model items;	The model command specifies the IRT model to be used. The items argument refers to the Rasch model for dichotomous items where the score for each item is 0 or 1
estimate;	The estimate command tells ConQuest to begin the estimation process
show !estimate = latent >> run1.shw;	The show command requests output of results to be written to a file called *run1.shw*. This "show" file contains a summary, item difficulty parameters and a person-item map
itanal >> run1.itn;	The itanal command requests an output of CTT results including percentages correct and discrimination indices
itanal !form = summary >> run1.itnSUM;	The itanal command requests a summary of item level results to be written to a file called *run1.itnSUM*
plot icc !bins = 6;	The plot command specifies that item characteristic curves (ICC) are to be plotted with ability measures grouped into six ability groups
equivalence wle;	The equivalence command requests a table showing the correspondence between test score and ability estimate (weighted likelihood estimate)
show cases !estimate=WLE >> run1.WLE;	The show cases command requests student ability estimates to be written to a file call *run1.WLE*. Weighted likelihood estimate (WLE) is requested
show cases !estimate=latent >> run1.pv;	The show cases command requests student ability estimates to be written to a file call *run1.pv*. Plausible values are requested

Fig. 9.4 Item labels file *Itemnames.txt*



```
===> item
1 PlaceValue
2 Time
3 Stamp
4 Map
5 Multiplication
6 FloorPlan
7 SportsGraph
8 TransportGraph
9 Lollies
10 Spinner
11 Fraction
12 NumberSentence
13 Ginerbreadman
14 PartyPies
15 Cubes
```

Table 9.2 A summary of item statistics

Item no.	Item label	Facility (%)	Discrimination (CTT)	Fit wt mean sq	Difficulty IRT (logit)
Item:7	(SportsGraph)	96.59	0.24	0.98	−2.45
Item:1	(PlaceValue)	95.24	0.22	1.02	−2.06
Item:10	(Spinner)	93.97	0.29	0.97	−1.78
Item:4	(Map)	93.21	0.28	0.99	−1.63
Item:8	(TransportGraph)	93.10	0.31	0.97	−1.61
Item:9	(Lollies)	91.78	0.27	1.03	−1.39
Item:3	(Stamp)	75.80	0.41	1.04	0.20
Item:12	(NumberSentence)	70.94	0.49	0.95	0.53
Item:11	(Fraction)	66.67	0.37	1.11	0.80
Item:5	(Multiplication)	64.63	0.42	1.05	0.93
Item:14	(PartyPies)	62.52	0.46	1.00	1.05
Item:15	(Cubes)	61.30	0.43	1.04	1.13
Item:2	(Time)	54.91	0.48	0.99	1.49
Item:13	(Ginerbreadman)	53.08	0.50	0.95	1.59
Item:6	(FloorPlan)	26.67	0.44	0.93	3.18

Only one item (Item 6) has a facility below 50%. This shows that the test is easy for the students. The IRT item difficulties also range from low to high, with an average difficulty of zero (This is the default setting of ConQuest, where the zero on the scale is set to the average item difficulty). From the set of IRT item difficulties, it is not possible to tell whether the test is easy or difficult for the students, since the IRT scale is set so that the zero on the scale is the average of item difficulties. However, when we compare student ability estimates with item difficulties, we will be able to assess whether the test is easy or difficult, as we explain in the section on the examination of student abilities. Figure 9.5 shows a plot of CTT item facilities against IRT item difficulties.

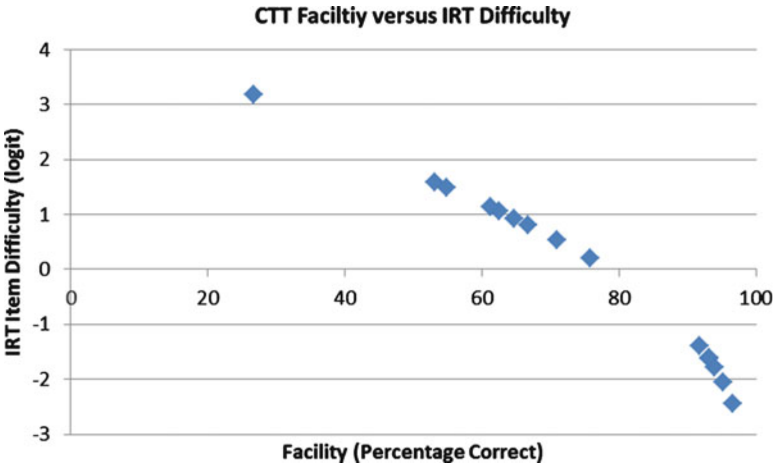


Fig. 9.5 CTT item facility versus IRT item difficulty

A number of observations can be made:

1. There is a one-to-one relationship between CTT item facility and IRT item difficulty. That is, the ranking of items in order of difficulty is exactly the same whether we use CTT item facility or IRT item difficulty.
2. The relationship between CTT item facility and IRT item difficulty is not a linear one. At the end of the scale (e.g. items with high facilities, say, above 90%), the IRT item difficulties are more "stretched out" (i.e. further apart) than for items in the middle range of facilities. However, for items in the middle range of facilities (say, between 20% and 80%), the relationship between CTT item facilities and IRT item difficulties is close to a straight line.

9.2.4 Examine CTT Item Discrimination Statistics

CTT item discrimination index is computed as the correlation between students' scores on an item and students' total scores on the test (excluding the item for which the discrimination is computed). For example, Table 9.3 shows an excerpt of students' scores on Item 2 and their total scores on the test (excluding Item 2), arranged in order of students' total scores. To calculate item discrimination for Item 2, we compute Pearson's correlation between the total score and item score (columns 2 and 3 in Table 9.3). This correlation is the discrimination index for Item 2.

If an item has high discriminating power in separating high- and low-ability students, we would expect that students with high total scores on the test to be more likely to have a score of 1 on the item, and students with low total scores on the test to have a score of 0 on the item. In contrast, if an item has no discriminating power, we would expect a lack of positive relationship between students' total test scores and their scores on the item. For example, if all students randomly guessed the answer

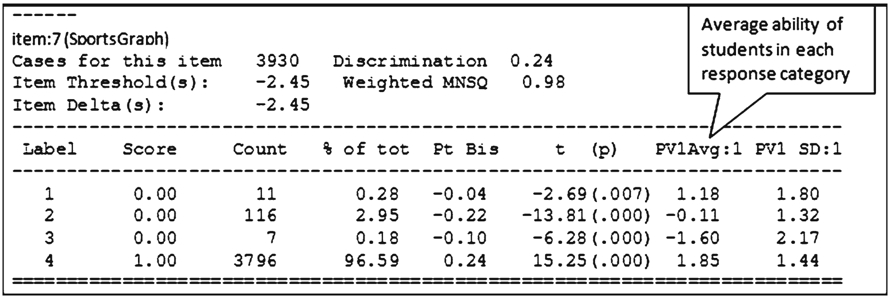
Table 9.3 CTT item discrimination computation

Student (arranged according to test score)	Total test score (excluding item 2)	Score on item 2
1	0	0
2	1	0
3	1	0
4	1	0
5	1	0
…	…	…
2,317	11	0
2,318	11	1
2,319	11	1
2,320	11	0
2,321	11	1
2,322	11	0
…	…	…
3,926	14	1
3,927	14	1
3,928	14	1
3,929	14	1
3,930	14	1
	Correlation between these two columns	

on an item, then the correlation between students' test scores and their scores on the item will be close to zero. In Table 9.3, we can see that, as the total score increases, there are more students obtaining a score of 1 on Item 2. The CTT discrimination index for Item 2 is 0.48, which is regarded as good discrimination power for a dichotomously scored item.

Item discrimination is an important index for assessing the quality of an item, particularly if the main purpose of the test is to separate students by ability levels. An item with a low discrimination index indicates that the item is testing something that is unrelated to what is being tested by the other items. Consequently, checking item discrimination should be one of the first priorities in examining the results of item analysis. In general, items with higher discrimination index are better items than those with lower discrimination index.

The information provided by item discrimination is different from the information provided by item difficulty. Item difficulty tells us about *how many* people obtained the correct answer. But item discrimination tells us *who* obtained the correct answer (e.g. high-ability students or students from a range of ability levels). Item difficulty does not tell us about the quality of an item. If an item is difficult, it may still be a very good item, provided that the few students who obtained the correct answer are the highest-ability students and, in which case, the item discrimination should be reasonably high.



```
------
item:7 (SportsGraph)
Cases for this item     3930   Discrimination   0.24
Item Threshold(s):      -2.45   Weighted MNSQ    0.98
Item Delta(s):          -2.45
-----------------------------------------------------------------------------
 Label    Score     Count    % of tot   Pt Bis      t   (p)    PV1Avg:1 PV1 SD:1
-----------------------------------------------------------------------------
   1       0.00       11       0.28     -0.04    -2.69(.007)   1.18     1.80
   2       0.00      116       2.95     -0.22   -13.81(.000)  -0.11     1.32
   3       0.00        7       0.18     -0.10    -6.28(.000)  -1.60     2.17
   4       1.00     3796      96.59      0.24    15.25(.000)   1.85     1.44
=============================================================================
```

Average ability of students in each response category

Fig. 9.6 Item analysis of Item 7 in the *run1.itn* file

However, it should be noted that, since the discrimination index is a correlation between two sets of scores, it is sensitive to the degree of variability of each set of scores. If an item is very easy or very difficult, there will not be much variation in students' scores on that item (i.e. mostly 0 or mostly 1), and the correlation will tend to be lower. That is, the CTT discrimination index will tend to be lower for very easy and very difficult items. In these cases, a low discrimination index may not reflect a poor quality item, but it is simply an artefact of the difficulty of the item. For example, Item 7 is a very easy item (facilities of 97%, Table 9.2). The discrimination value is also somewhat low (0.24). However, an examination of the item analysis for Item 7 (Fig. 9.6) shows that the students who obtained the correct answer are of higher average ability than students who obtained the incorrect answers. The low discrimination is a result of the lack of variation in students' scores on this item (i.e. mostly 1) and not a result of poor item construction.

Nevertheless, while item difficulty has an impact on the magnitude of CTT item discrimination index, the conceptual differences between item difficulty and item discrimination should be understood as outlined in this section.

9.2.5 Examine Item Discrimination Using IRT

Under IRT, item discrimination can be checked in two ways. First, the item characteristic curves (ICC) show the steepness of the observed ICC. In ConQuest, the command "plot icc !bins =6;" produces item characteristic curves, with abilities grouped into six groups. Figures 9.7 and 9.8 show the ICCs for two items.

The solid line graphs in Figs. 9.7 and 9.8 are theoretical ICCs, computed using the Rasch model probability function after the item difficulties are estimated. The dotted line graphs are observed ICCs based on the data collected for the item (i.e. empirical ICCs). For example, for Item 5, the observed ICC is slightly *flatter* than the theoretical ICC; and, for Item 6, the observed ICC is slightly *steeper* than the theoretical ICC. A flat ICC shows that the item has less discriminating power than expected theoretically. In the extreme case, if an observed ICC is very flat (see an example in Fig. 9.9), students with low abilities have similar chances of getting the

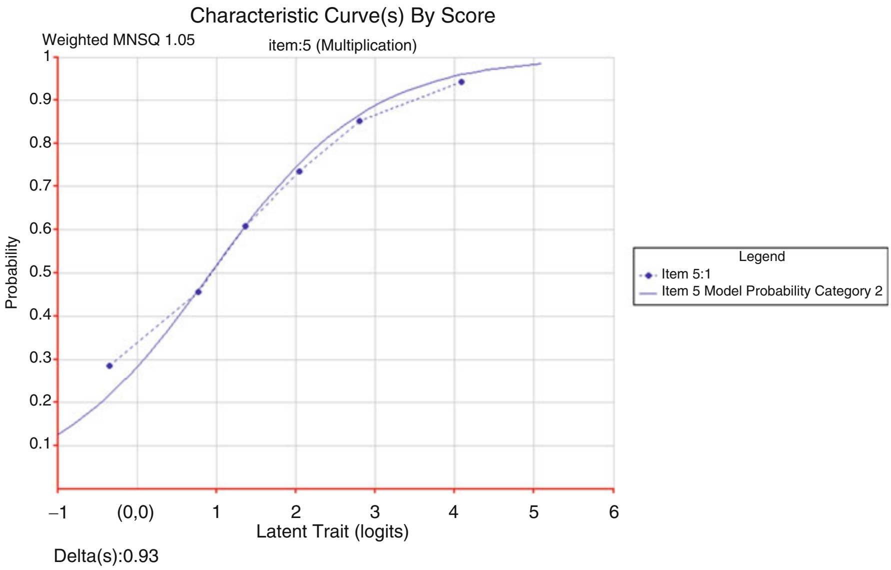


Fig. 9.7 Item characteristic curves for Item 5

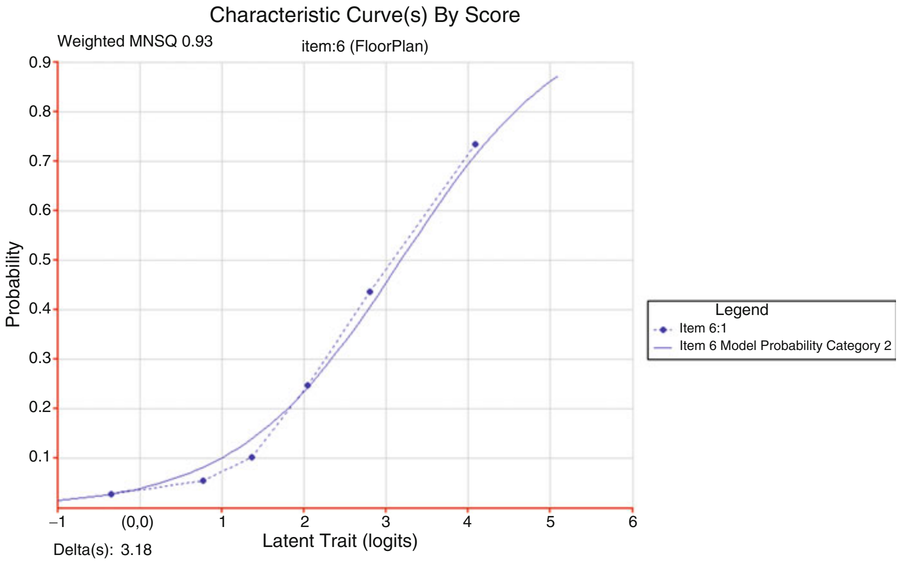


Fig. 9.8 Item characteristic curves for Item 6

correct answer as students with high abilities. Clearly, such an item is not working well in measuring student ability. Consequently, the steeper the ICC, the more power an item has in separating students of different ability levels. It should be noted that the steepness of the ICC is not influenced by the item difficulty, in contrast to CTT where the discrimination index is influenced by item difficulty.

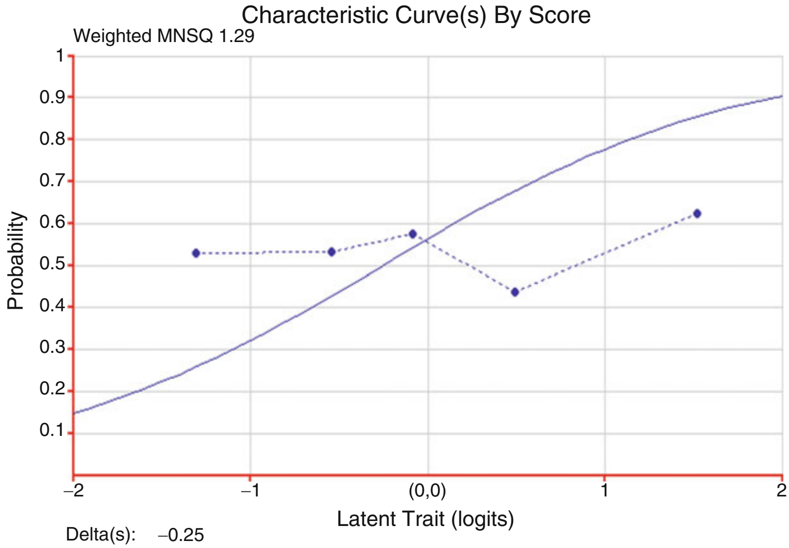
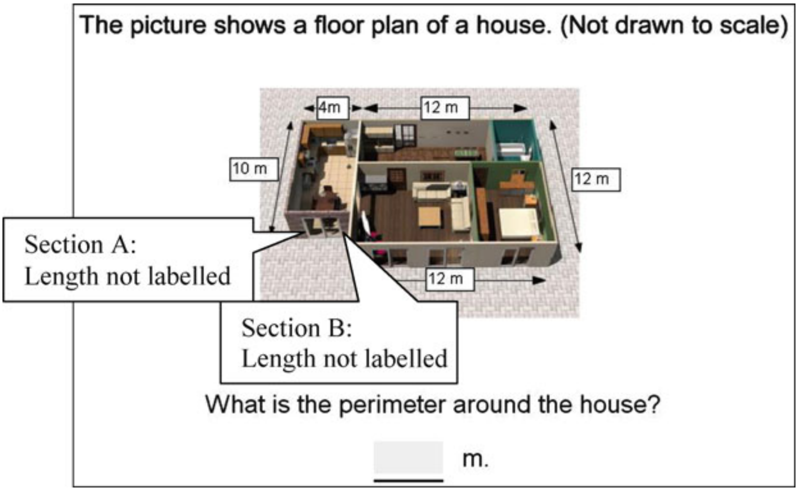


Fig. 9.9 An example of a "very flat" observed ICC

The second way to examine item discrimination in IRT is to use the fit statistics (Wu and Adams 2008). For the Rasch model for dichotomous items, the residual-based fit statistics (fit mean squares) as reported by ConQuest reflect the discrimination of the items. For example, in Fig. 9.7, Item 5 is less discriminating than expected (observed ICC flatter than the theoretical ICC), and the weighted fit mean squares statistic is greater than 1 (1.05). In contrast, in Fig. 9.8, Item 6 is more discriminating than expected (observed ICC steeper than the theoretical ICC), and the weighted fit mean squares statistic is less than 1 (0.93). By checking whether the fit mean squares statistic is more than 1 or less than 1, we can get an indication of whether the item is less discriminating or more discriminating than expected.

9.2.6 Examine Response Categories (Distractor Analysis)

In preparing the data set for analysis, it was decided that we should retain as much *raw* information as possible. That is, we have retained the actual student responses to the items for multiple-choice items, rather than scoring these before carrying out the item analysis. For open-ended items, in most cases, we have categorised student responses into a number of response categories. The scoring of the items is carried out within ConQuest, and the key statement (the third line in the ConQuest control file) is included for this purpose. An advantage of using raw responses instead of scored responses is that we can obtain information on how each response category worked and use this information to revise items if necessary. The *itanal* file (*run1.itn*) shows response category statistics for each item. Figures 9.10 and 9.11 show an example item and the corresponding response category statistics.

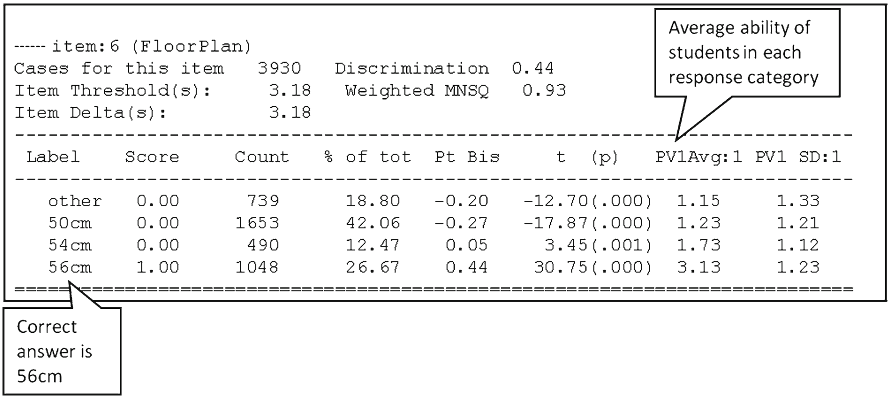


The picture shows a floor plan of a house. (Not drawn to scale)

What is the perimeter around the house?

_____ m.

Fig. 9.10 Item 6 – FloorPlan



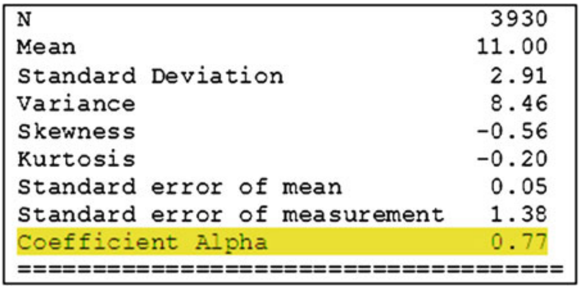
Average ability of students in each response category

```
------ item:6 (FloorPlan)
Cases for this item   3930   Discrimination  0.44
Item Threshold(s):      3.18   Weighted MNSQ   0.93
Item Delta(s):          3.18
--------------------------------------------------------------------------
 Label     Score      Count   % of tot  Pt Bis      t   (p)   PV1Avg:1 PV1 SD:1
--------------------------------------------------------------------------
   other    0.00       739      18.80    -0.20   -12.70(.000)  1.15      1.33
   50cm     0.00      1653      42.06    -0.27   -17.87(.000)  1.23      1.21
   54cm     0.00       490      12.47     0.05     3.45(.001)  1.73      1.12
   56cm     1.00      1048      26.67     0.44    30.75(.000)  3.13      1.23
==========================================================================
```

Correct answer is 56cm

Fig. 9.11 An example of response categories statistics

Figure 9.11 shows that students who obtained the correct answer (56 cm) have the highest average ability (3.13, under the column headed “PV1Avg:1”). Students who answered 54 cm have higher average ability than students who gave other incorrect answers. To obtain 54 cm as the answer, students have spotted one part of the house (Sect. A, horizontal to our view) with an unlabeled length, but they have failed to spot the other part (Sect. B, vertical to our view) of unlabeled length. Further, to work out the length of Sect. B, a subtraction needs to be carried out (12 – 10 cm). To obtain 50 cm as the answer, students have simply added up the labelled sections of the house. They have failed to spot the two unlabelled sections. Overall, the average abilities for the response categories, and the



```
N                                    3930
Mean                                11.00
Standard Deviation                   2.91
Variance                             8.46
Skewness                            -0.56
Kurtosis                            -0.20
Standard error of mean               0.05
Standard error of measurement        1.38
Coefficient Alpha                    0.77
=========================================
```

Fig. 9.12 CTT statistics in the item analysis output (run1.itn)

point biserial correlations for the categories, are as we would expect for this item. The response category statistics tell us that the majority of the students understand the notion of *perimeter* (i.e. the length around the house), as most students carried out additions of the lengths given. But students differ in their degree of observation of the information given. More importantly, students who gave *better* answers are more able students.

9.2.7 Examine Student Test Scores with CTT Statistics

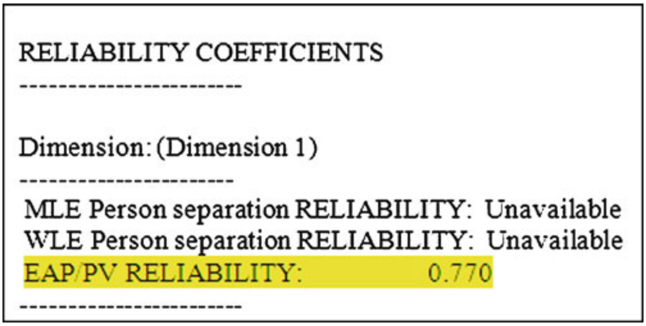
Figure 9.12 shows some CTT statistics as reported at the end of the *itanal* file, *run1.itn*, regarding student test scores.

As Fig. 9.12 shows, classical test theory focuses on statistics based on raw test scores. The average score of 3,930 students on the test is 11, out of a maximum of 15 on the test. The test is relatively easy for most students. A standard deviation of 2.91 indicates that the range of student scores is mostly between 5 and 15 on the 15-item test (assuming that 95% of the scores are in the range "mean ± 2 × standard deviation"). The standard error of measurement is 1.38, indicating that, should similar tests be administered, a student's score could vary by ±2.7, a range of around 5 scores. For example, a student's score could vary between 8 and 13, should similar tests be given. A range of 5 score points shows that there could be considerable variability in a student's test scores and that the current test does not provide a very accurate measure of student performance. This is as expected because a test of 15 questions is a rather short test. The test reliability, coefficient alpha, is 0.77, indicating that the test is not exceedingly reliable.

9.2.8 Examine IRT Ability Estimates

Under IRT, a statistic called person separation reliability is reported by ConQuest in the *show* file (*run1.shw*). Figure 9.13 shows the results.

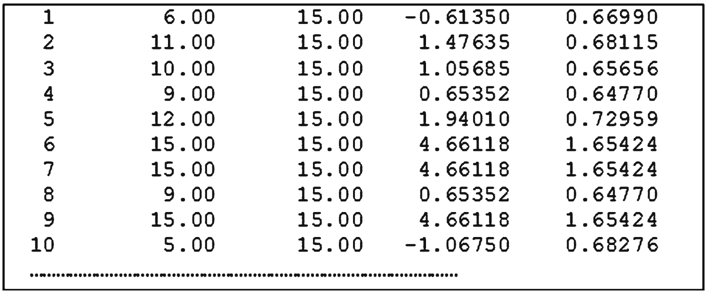
The IRT person separation reliability index is computed based on the measurement accuracy (known as measurement error) of individual ability measures and the variation of ability measures across the group of students taking the test. While the



```
RELIABILITY COEFFICIENTS
------------------------

Dimension: (Dimension 1)
-----------------------
 MLE Person separation RELIABILITY:  Unavailable
 WLE Person separation RELIABILITY:  Unavailable
 EAP/PV RELIABILITY:              0.770
------------------------
```

Fig. 9.13 IRT person separation reliability



```
  1        6.00       15.00    -0.61350     0.66990
  2       11.00       15.00     1.47635     0.68115
  3       10.00       15.00     1.05685     0.65656
  4        9.00       15.00     0.65352     0.64770
  5       12.00       15.00     1.94010     0.72959
  6       15.00       15.00     4.66118     1.65424
  7       15.00       15.00     4.66118     1.65424
  8        9.00       15.00     0.65352     0.64770
  9       15.00       15.00     4.66118     1.65424
 10        5.00       15.00    -1.06750     0.68276
.........................................................
```

Fig. 9.14 An excerpt of the student ability estimates file run1.WLE

formula is not the same as the CTT coefficient alpha, the interpretation is the same. In our example, the IRT person separation reliability is essentially the same as the CTT coefficient alpha.

An IRT measure that is equivalent to the CTT standard error of measurement (see Fig. 9.12) is the standard error of the estimated ability. In the output file *run1.WLE*, the standard error of each WLE ability estimate is given. Figure 9.14 shows an excerpt of the file.

There are five columns of numbers in Fig. 9.14. The first column shows the student number. The second column shows the test score of the student on the test. The third column shows the possible maximum score on the test. The fourth column is the weighted likelihood (WLE) ability estimate of θ, for each student. The last column is the standard error of the estimated ability estimate. A number of observations can be made:

1. One should note that the magnitude of the standard errors is rather large, showing that each ability measure is not very precisely estimated. For example, for Student 1, the 95% confidence interval of the student's ability is between −1.9 and 0.7 ($-0.61350 \pm 2 \times 0.66990$), a range of around 2.6 logits wide. This degree of inaccuracy is expected from a 15-question test.
2. There is a one-to-one correspondence between test score and IRT ability estimate (WLE). This means that all students with the same test score will have the same

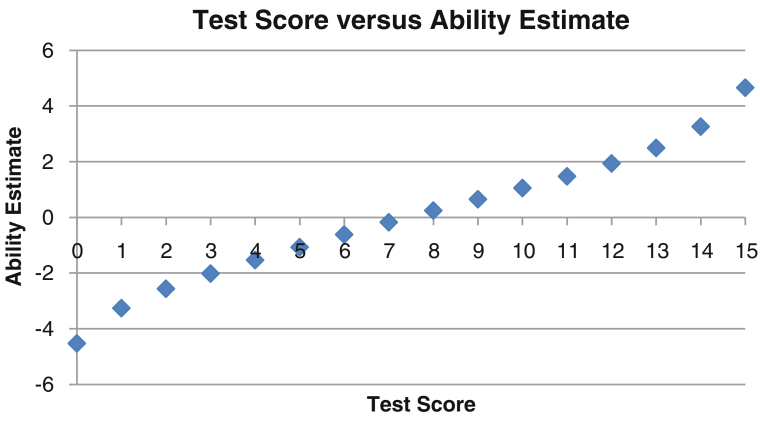


Fig. 9.15 Plot of test score versus ability estimate (WLE)

ability estimate. That is, the rank ordering of students by test score (CTT) or by ability estimate (IRT) will be exactly the same.

3. Equal test score difference does not correspond to equal ability difference. For example, the difference in ability estimates for students with test scores of 9 and 10 is 0.40. The difference in ability estimates for students with test scores of 11 and 12 is 0.46. In general, the difference in ability estimates are more stretched out at the lower and higher ends of the scale, as can be seen in a scatter plot shown in Fig. 9.15. Figure 9.15 shows that the relationship between test score and ability estimate is not a straight line, with ability estimates slightly stretched out at the ends of the scale (lowest and highest scores). However, the relationship is close to a linear one for the middle range of test scores.

9.2.9 *Further Examination of IRT WLE Ability Estimates*

The fact that test scores are bounded (by zero and the maximum score) is a limitation of the use of test scores to make inference on ability. The IRT ability estimates do seem to stretch out the scale a little and, in theory, θ can be a better representation of students' underlying ability. However, in practice, as each test score corresponds to one ability estimate, the range of ability estimates is limited to the number of possible scores in a test (in this case, it is 16 (0–15)). Consequently, when we form a distribution of IRT abilities (e.g. WLE), the distribution is discrete (i.e. not continuous) and bounded, as can be seen in Fig. 9.16.

For a comparison, Fig. 9.17 shows the test score distribution. By and large, there is not a great deal of difference between the use of test scores and IRT WLE ability estimates to construct the ability distribution. The only slight difference is that, at the tail ends of the horizontal scale, the IRT distribution is more "stretched out" (i.e. a larger gap between adjacent scores).

The fact that the ability distribution constructed using the IRT ability estimates (WLE) is somewhat skewed is because the test is easy for the group of students, and most

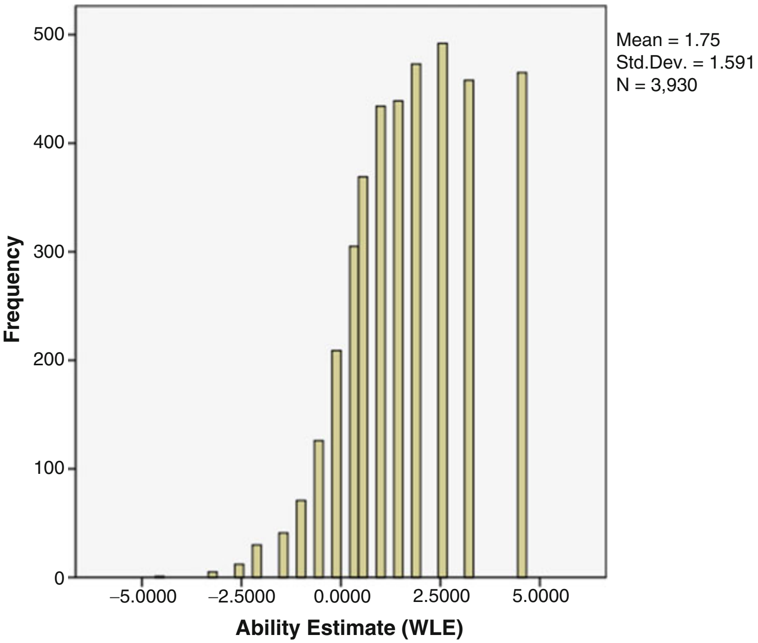


Fig. 9.16 Frequency distribution of ability estimates (weighted likelihood estimates)

students obtained high test scores. While Fig. 9.16 may be an accurate representation of test scores obtained by the group of students on this test, it does not seem a good representation of the underlying students' mathematics abilities, as we would expect a more symmetrical distribution. The ceiling effect of an easy test has skewed the shape of the ability distribution. If we administer a test that is better targeted at the average ability of the group of students, we would expect an ability distribution (and a score distribution) that is more symmetrical, with more students in the middle of the scale (a bell-shaped curve, like the normal distribution). That is, with the IRT weighted likelihood ability estimates, the problem with the boundedness of test scores has not really been solved.

9.2.10 Alternative IRT Ability Estimates: Plausible Values

The software program, ConQuest, differs from a number of Rasch modelling programs in the estimation method used. ConQuest uses marginal maximum likelihood (MML) estimation method rather than joint maximum likelihood or conditional maximum likelihood method. While a discussion of estimation methods is beyond the scope of this chapter, we will briefly explain the idea of MML estimation and the consequences of using this estimation method.

Equation 9.1 shows the item response model. For most Rasch model software programs, this is the only assumption of the model. However, for marginal maximum likelihood (MML) estimation method, there is an additional assumption about

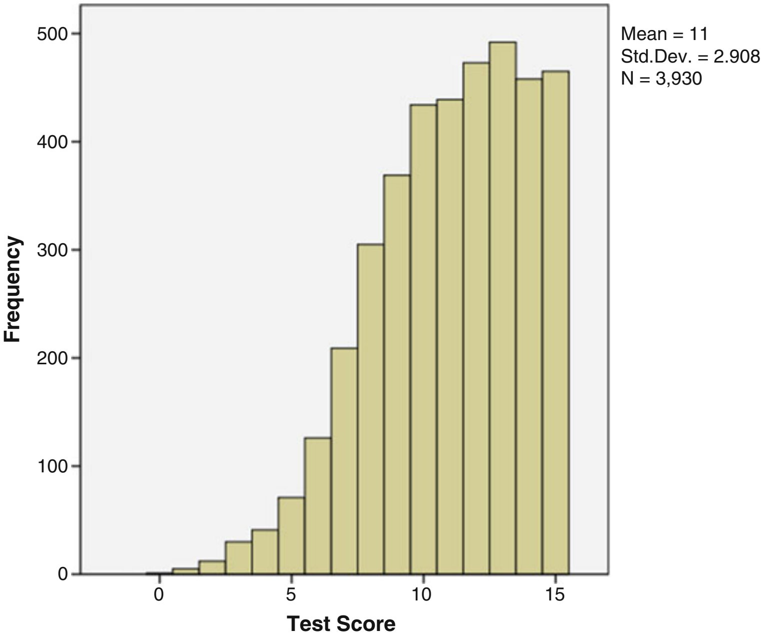


Fig. 9.17 Frequency distribution of test scores

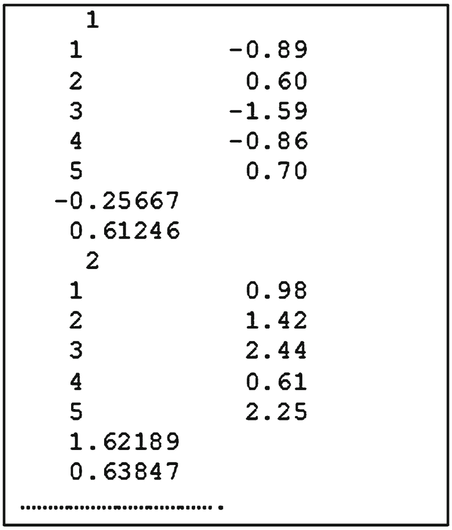
the shape of the population ability distribution. In most cases, an assumption is made that the population ability distribution is normally distributed (i.e. bell shaped) with mean μ and standard deviation σ, where (μ, σ) are estimated using the item response data. The advantage of making an assumption of the shape of the population ability distribution is that the mathematics involved in deriving the ability estimates can make use of this assumption so that a resulting estimated population distribution will have the shape of a normal distribution, despite whether the test is well targeted to students' abilities or not. The disadvantage of making a population assumption is that if the assumption is incorrect, then the results could be invalid. One needs to weigh the benefit against the cost, as in any mathematical modelling.

One type of ability estimate under the marginal maximum likelihood estimation method is called plausible values (Wu 2005). The idea of plausible values is that, given the item responses of a student and the overall shape of the ability distribution, we can work out the propensity of the ability range of a particular student. For example, we are able to make likelihood statements such as "there is one in ten chance that the student's ability is at 0.8, and two in ten chance that the student's ability is at 1.3," etc. Each plausible value computed for a student is a probable ability location of the student, according to the likelihood statements.

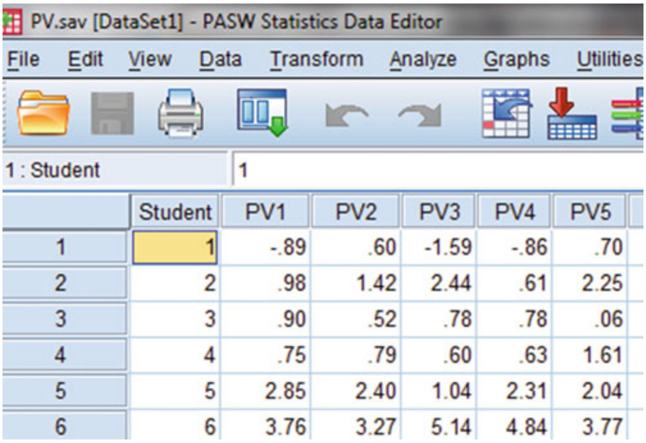
The ConQuest command (see Table 9.1, last command) for producing plausible values is straightforward. The word "latent" in this ConQuest command specifies "plausible values." The resulting output file is shown in Fig. 9.18.

The plausible values file, *run1.pv*, contains eight lines per student. Five plausible values are produced for each student. For example, for Student 1, the five plausible

Fig. 9.18 An excerpt of the plausible values file *run1.pv*



```
   1
  1          -0.89
  2           0.60
  3          -1.59
  4          -0.86
  5           0.70
 -0.25667
  0.61246
   2
  1           0.98
  2           1.42
  3           2.44
  4           0.61
  5           2.25
  1.62189
  0.63847
.................................
```



	Student	PV1	PV2	PV3	PV4	PV5
1	1	-.89	.60	-1.59	-.86	.70
2	2	.98	1.42	2.44	.61	2.25
3	3	.90	.52	.78	.78	.06
4	4	.75	.79	.60	.63	1.61
5	5	2.85	2.40	1.04	2.31	2.04
6	6	3.76	3.27	5.14	4.84	3.77

Fig. 9.19 A snippet of the SPSS file of plausible values

values are –0.89, 0.60, –1.59, –0.86 and 0.70. These are five probable ability estimates for Student 1, generated using the likelihood statements derived from the MML model. Importing the plausible values file into SPSS, one can examine the distribution of the plausible values (Fig.9.19).

Using the first plausible value, PV1, to construct a frequency distribution, (Fig. 9.20), it can be seen that a more symmetrical ability distribution is formed, in comparison to the skewed distribution shown in Fig. 9.16. Thus, the use of plausible values as ability estimates produces better estimates of population characteristics than the use of test scores or the weighted likelihood estimates (WLE). The shape of the estimated population distribution is less dependent on the test difficulty.

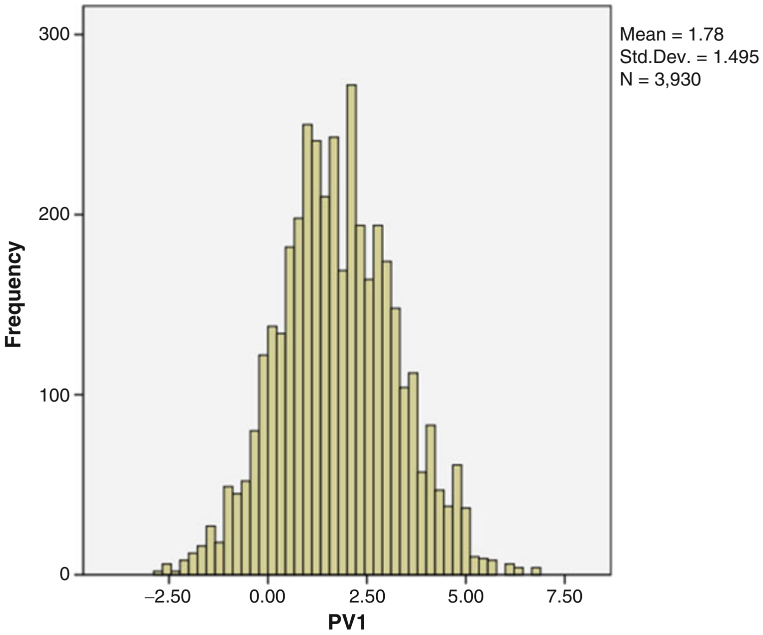


Fig. 9.20 Frequency distribution of plausible values

However, it should be noted that if individual student results are reported, then plausible values are not suitable, since we would not want to provide a set of probable scores to a student. Instead, we want to report a single ability estimate. In that case, the weighted likelihood estimate (WLE) is still the best one to use.

9.2.11 Mapping Student Abilities to Item Difficulties

So far, in examining item properties and student abilities, CTT and IRT provide similar and complementary information. However, one analysis of IRT that is not readily obtainable from CTT is the mapping between student abilities and item difficulties. Figure 9.21 shows a person-item map (see output file *run1.shw* for this map).

In the left-hand panel of Fig. 9.21, the distribution of student abilities is shown. Each "*x*" represents 5.6 students in this case. Note that this distribution is built with plausible values (PV) ability estimates and not with the weighted likelihood (WLE) ability estimates. The ConQuest command "show !estimate = latent >> run1.shw;" requests that the ability distribution be built with plausible values by specifying the option "!estimate = latent." Had we requested WLE ability estimates, the ability distribution would be a skewed and discrete distribution, as shown in Fig. 9.16.

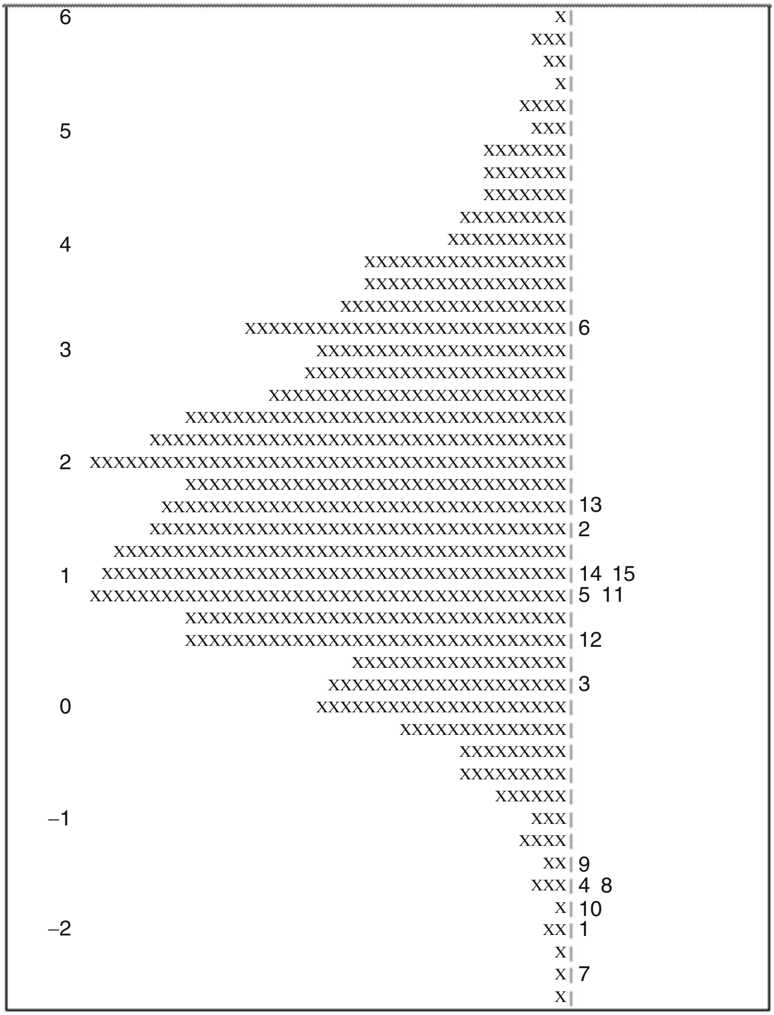


Fig. 9.21 Person-item map

In the right-hand panel of Fig. 9.21, the items are located according to their difficulty values. For example, we see that Item 7 is located at the bottom of the map, as it is the easiest item in the test (facility of 97% and item difficulty estimate of–2.45). At the top end of the scale, Item 6 is the most difficult item (facility of 27% and item difficulty estimate of 3.18).

It is possible to place items on the same scale as students because, under IRT, item difficulties are defined on the ability scale. This important IRT property enables us to make statements about the likelihood of the tasks students can do in relation to the items. For example, students located at around 1 logit will have 50% chance of obtaining the correct answers to Questions 5, 11, 14 and 15. Further, they will be able to answer Questions 1, 4, 7, 8, 9 and 10 relatively easily, but they will have difficulties in answering Question 6. Such probabilistic statements provide descriptions of the skill sets of students, in addition to the numerical values of ability we assign them. It should be noted, however, that a single test of 15 questions does not provide very accurate ability estimates, as shown in Fig. 9.14. So any inferences made about individual students need to take into account of the margin of error surrounding individual student ability estimates.

9.2.12 Potential for Constructing a Described Proficiency Scale

While a test of 15 questions cannot provide a detailed description of skills along the ability scale, it is possible to combine a number of tests and many items together and calibrate the items along the same scale. The example test used in this chapter is one of a set of mathematics tests constructed for Grade 5 students. These tests can be calibrated together so that all items can be placed on the same scale. Such a process is called *equating*. While equating is outside the scope of this chapter, Table 9.4 shows an example of a described proficiency scale.

Table 9.4 is useful in providing substantive descriptions of skills attached to numerical ability measures. Such descriptions can assist the teachers and curriculum designers to link student test results to a set of proficiency statements.

9.3 Summary and Conclusions

In this chapter, a data set of item responses to a mathematics test is used to illustrate how the data set can be analysed using the software program, ConQuest. Further, the results of the analysis are discussed with respect to the interpretations of the CTT and IRT statistics. In examining item statistics, the notions of item difficulty and item discrimination are discussed. In examining student ability estimates, the concepts of test reliability, standard error of measurement and different ability estimates are discussed.

This chapter demonstrates that many CTT and IRT statistics provide similar results. If one is simply interested in the results of one single test, there is not much advantage in using IRT over CTT, although IRT does provide some niceties in presenting results graphically. The main advantage of IRT over CTT is that IRT enables the placement of items and students on the same scale, leading to the possibility of the construction of described proficiency statements along the scale. Under IRT, we can provide not only a numerical ability measure but also a substantive description of skills underlying each ability measure. Further,

Table 9.4 An example of a described proficiency scale

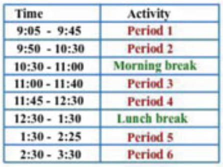
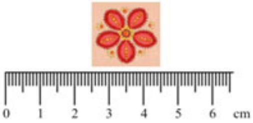
Level	Logit range	Description
3	2–4	Typically, students working at this level can perform two-digit additions and subtractions with carrying, as well as simple three-digit additions and subtractions. They can work out the total cost given the unit price, e.g. the cost for three people at \$3.50 each. They can work out simple number sentences such as 12 + ? = 7 + 8. Students at this level typically understand the notion of chance and can compare the likelihood of events. They can read a simple bar chart with axis labels and identify the frequencies of occurrences of events. They typically know names of polygons and understand the notions of sides, lines and simple spatial orientation. They can read a simple map and work out distances between locations using addition
2	0–2	Typically, students working at this level can carry out simple two-digit additions and subtractions by setting it out formally. They can perform simple multiplication operations without resorting to counting and adding. For example, they can use the multiplication operation to work out the total cost for eight people at \$2 each. They can count in multiples and recognise simple number patterns. They can work out clock time and elapsed time. Students at this level typically understand the relationship between hours and minutes, and they use formal units such as centimetres, metres, kilograms and litres
1	−2–0	Typically, students working at this level understand number representations up to four digit numbers. They can add and subtract two-digit numbers using counting. They understand the notion of multiplication and division operations as repeated addition and repeated sharing. They can use a ruler to carry out simple measurements of length and read a clock face to tell the time. They understand the notion of the likelihood of events in tossing a coin or spin a spinner. They can read simple bar charts and observe the most frequent and the least frequent events

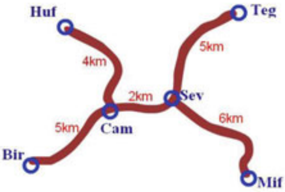
different tests can be combined together through an equating process under the IRT framework, giving rise to an even wider scope of building a coherent set of assessments.

It should be noted that there is a set of assumptions under IRT. In particular, there is a mathematical probability function that relates the item responses to item difficulty and student ability. The valid use of IRT is contingent on the fact that the observed item responses fit the IRT mathematics model. There is no guarantee at all why a set of item responses will fit a particular IRT model. To be able to claim the benefits of IRT, we must establish that the data fit the IRT model sufficiently well for the purposes of the assessment.

To conclude, both CTT and IRT provide useful tools for building quality assessments. But to utilise these methodologies to the fullest, there must be a clear understanding of the uses, as well as the limitations, of the various statistics provided.

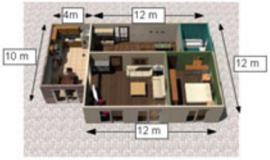
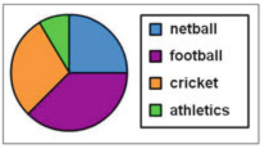
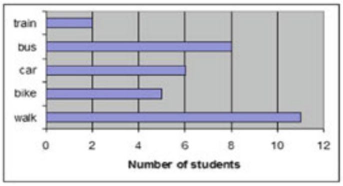
9.4.1 Appendix A.1 Mathematics Test

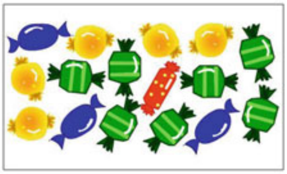
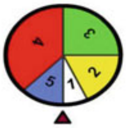
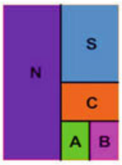
<table>
<tr><td>1</td><td>PlaceValue2</td><td>What is the value of 6 in 1762?
○ 6 ones
○ 6 tens
○ 6 hundreds
○ 6 thousands</td></tr>
<tr><td>2</td><td>Time6</td><td>The following shows a school timetable:

<table>
<tr><th>Time</th><th>Activity</th></tr>
<tr><td>9:05 - 9:45</td><td>Period 1</td></tr>
<tr><td>9:50 - 10:30</td><td>Period 2</td></tr>
<tr><td>10:30 - 11:00</td><td>Morning break</td></tr>
<tr><td>11:00 - 11:40</td><td>Period 3</td></tr>
<tr><td>11:45 - 12:30</td><td>Period 4</td></tr>
<tr><td>12:30 - 1:30</td><td>Lunch break</td></tr>
<tr><td>1:30 - 2:25</td><td>Period 5</td></tr>
<tr><td>2:30 - 3:30</td><td>Period 6</td></tr>
</table>
The school day from Period 1 to the end of Period 6 is 6 hours and ____ minutes</td></tr>
<tr><td>3</td><td>stamp1</td><td>What is the width of this stamp?


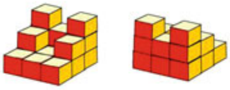
○ 1 cm
○ 2 cm
○ 3 cm
○ 4 cm</td></tr>
<tr><td>4</td><td>Map2</td><td>The map shows 6 towns and the distances between the towns by road.


Huf
Teg
5km
4km
2km
Sev
6km
5km
Cam
Bir
Mif

To go from Mif to Huf by road, you need to travel ____ km.</td></tr>
</table>

5	Multiplication2	**215** × **8** = ? The number in the box with the question mark is .
6	FloorPlan1	The picture shows a floor plan of a house. (Not drawn to scale)  What is the perimeter around the house? m.
7	SportsGraph2	In Daniel's class, there are 24 students. Each student has to choose one physical activity at school. The Pie chart shows the proportion of students taking each activity.  Which activity is the LEAST popular? ○ netball ○ football ○ cricket ○ athletics
8	TransportGraph1	The graph shows how students in Amy's class go to school.  How many students go to school by bike? students.

9	Lollies2	A bag contains these lollies.  Sam picks one from the bag without looking. Which one is Sam LEAST likely to pick?
10	Spinner5_2	This spinner is used for a game.  The spinner will be LEAST likely to stop on Number ____.
11	ShapeFraction4	The webpage of a City Council is divided into 5 sections as shown.  **N:** main news **S:** sports news **C:** community announcements **A:** advertisements **B:** contact information Estimate the fraction of the page that is devoted to the Sports News. $\frac{1}{2}$ $\frac{1}{3}$ $\frac{1}{4}$ $\frac{1}{5}$
12	NumberSentence1	$12 \times \boxed{?} = 18 \times 2$ The number in the box with the question mark is ____.

13	Gingerbreadman3	Each gingerbread man costs $2.50. Tim bought 5 gingerbread men and gave the cashier $20. How much change did Tim get back? $ ______
14	PartyPie1	 Party pies are sold in boxes of 6. 25 children are at a birthday party. If each child gets one party pie, how many boxes need to be bought? ____ boxes
15	Cubes2	The following two pictures show a FRONT view and a BACK view of a shape made of cubes.  How many cubes are used to build this shape? ____ cubes

9.4.2 Appendix A.2 Recoding of the Item Responses

Question	New code = student response
1. Place value	1 = 6 ones
	2 = 6 tens
	3 = 6 hundreds
	4 = 6 thousands
	9 = missing response
2. Time	1 = 25 (minutes)
	2 = 30 (minutes)
	3 = 35 (minutes)
	4 = 40 (minutes)
	0 = all other responses
	9 = missing response
3. Stamp	1 = 1 cm
	2 = 2 cm
	3 = 3 cm
	4 = 4 cm
	9 = missing response

Question	New code = student response
4. Map	1 = 12 (km)
	0 = all other responses
	9 = missing response
5. Multiplication	1 = 1720
	0 = all other responses
	9 = missing response
6. Floor plan	1 = 50 (m)
	2 = 54 (m)
	3 = 56 (m)
	0 = all other responses
	9 = missing response
7. Sports graph	1 = netball
	2 = football
	3 = cricket
	4 = athletics
8. Transport graph	1 = 4 (students)
	2 = 5 (students)
	0 = all other responses
	9 = missing response
9. Lollies	1,2,3,4 according to the order of the 4 response options.
	9 = missing response
10. Spinner	1 = 1
	2 = 2
	3 = 3
	4 = 4
	5 = 5
	9 = missing response
11. Shape fraction	1 = 1/2
	2 = 1/3
	3 = 1/4
	4 = 1/5
	9 = missing response
12. Number sentence	1 = 2
	2 = 3
	3 = 36
	0 = all other responses
	9 = missing response
13. Gingerbread man	1 = ($) 5
	2 = ($) 7.5
	3 = ($) 8.5
	4 = ($) 10
	0 = all other responses
	9 = missing response

Question	New code = student response
14. Party pies	1 = 3 (boxes)
	2 = 4 (boxes)
	3 = 5 (boxes)
	0 = all other responses
	9 = missing response
15. Cubes	1 = 14
	2 = 15
	3 = 16
	4 = 17
	0 = all other responses
	9 = missing response

References

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: CBS College Publishing.

Lord, F. M. (1952). A theory of test scores. *Psychometrika Monograph, No. 7, 17* (4, Pt. 2).

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Wu, M. L. (2005). The role of plausible values in large-scale surveys. In Postlethwaite (Ed.), *Special issue of studies in educational evaluation (SEE) in memory of R M Wolf. 31 (2005)* (pp. 114–128).

Wu, M. L., & Adams, R. J. (2008). *Properties of Rasch residual fit statistics*. Unpublished paper.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACERConQuest version 2: Generalised item response modelling software*. Camberwell: Australian Council for Educational Research.

Chapter 10
A Concurrent-Separate Approach to Vertical Scaling

Zi Yan, Doris Ching Heung Lau, and Magdalena Mo Ching Mok

10.1 Introduction

In the current educational climate, tracking students' academic growth in subjects (i.e. mathematics, reading, etc.) over time is of great interest to educators, as well as to the public. An implicit requirement of tracking is that performance and test items across grades can be compared using an established framework. It is obvious that the scores across grades obtained in achievement tests routinely used by schools or large-scale assessment programs cannot be compared directly because the difficulty of such tests and programs differs between grades. Suppose students A and B got the same score, for example, 80 points, in Primary 1 mathematics test and Primary 4 test, respectively. Although they have the same score, student B certainly has higher mathematical ability than student A because the test for Primary 4 level is more difficult than the test for the Primary 1 level. Such scores obtained from the different tests must be placed on a common scale before they can be compared and interpreted under the same framework.

Z. Yan (✉)
Department of Curriculum and Instruction, The Hong Kong Institute of Education,
Tai Po, Hong Kong
e-mail: zyan@ied.edu.hk

D.C.H. Lau
Formerly Centre for Assessment Research and Development, The Hong Kong Institute
of Education, Tai Po, Hong Kong
e-mail: chlaudoris@gmail.com

M. Mo Ching Mok
Department of Psychological Studies, and Assessment Research Centre,
The Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po N. T.,, Hong Kong
e-mail: mmcmok@ied.edu.hk

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_10, © Springer Science+Business Media Dordrecht 2013

Vertical scaling places the scores obtained from tests with different difficulty levels and measures the same construct on a common scale. The scale developed through vertical scaling is called a vertical scale (also referred to as a developmental scale) (Briggs and Weeks 2009; Harris 2007; Tong and Kolen 2007).

Vertical scaling is usually derived from a set of tests that are developed to assess the same domain across a range of grades. These tests are linked through common items (or linking items) that are shared by adjacent grades. A statistical procedure, usually using unidimensional item response theory (IRT), is then applied to the set of tests, and all of the items in those tests are calibrated on the same latent scale. The resulting vertical scale consists of an item pool, with each item having a fixed difficulty estimate.

10.2 Importance of Vertical Scaling

Vertical scales facilitate monitoring of students' academic growth over time. This has proved challenging for traditional grade-by-grade assessment approaches due to the incomparability of scores obtained on different tests, which are comprised of different items with various difficulty levels. Vertical scales overcame this problem by calibrating all of the items in different tests on a common scale. It provides a stable framework for comparison and interpretation of students' abilities estimated from different tests. Once an item is calibrated on the vertical scale, it has a unique estimate of difficulty on the scale, and this estimate remains invariant for all students and all test situations. Teachers, parents or anyone who wishes to measure students' achievement levels can formulate a test by drawing items from the item pool provided by the vertical scales according to different criteria or different situations. It is just like selecting different "rulers" with different minimum and maximum values, while using the same unit of length from a ruler pool to measure the length of objects with different sizes. A "0–200 cm" ruler can be used to measure adults' heights, and a "0–100 cm" ruler can be used to measure babies' heights. In a similar way, a test can be formulated for grade 4 students by selecting items with a particular range of difficulty levels; a specific item with a lower difficulty level can also be utilised to assess the ability levels of grade 2 students. What is more important and exciting is that students' ability levels measured by the different tests—they are calibrated on the same vertical scale—can be interpreted in the same framework and compared along the same scale. Another important feature of vertical scales developed using unidimensional IRT models is that the scores obtained in the vertical scales are linear and equal-interval measures. The same scores reflect the same amount of the construct measured, irrespective of the source test; moreover, adding one more unit results in equal-size increments. For example, a score of 60 in the grade 2 test represents the same level of ability as that of a 60 in the grade 4 test as long as both tests are vertically scaled. The growth from 60 to 70 (10 points) is the same as the growth from 70 to 80 (10 points) on a vertical scale. Therefore, as long as the item pool covers a corresponding range of difficulty, vertical scales make it feasible to track students' academic growth across a range of grades.

10.3 Challenges in Vertical Scaling

Although vertical scaling is a promising approach for monitoring students' development over time, there are concerns about the utility of these scales in a practical educational context. The most important concern probably relates to doubts about the validity of the unidimensionality assumption of the construct being measured across several grades (e.g. Camilli 1999; Lissitz and Huynh 2003; Yon 2006). As vertical scales are usually developed using unidimensional IRT models, the tests across grades are assumed to measure the same trait, just at different difficulty levels. Violation of the unidimensionality assumption would influence the vertical scaling results. If the assessments are designed to measure several distinct dimensions of the content that explains performance differences, then a vertical scale is not expected to produce usable data (Yen 2009). Therefore, test developers need to ensure that the items in the tests across different grades measure the same dimension of the construct to satisfy the unidimensionality assumption for vertical scaling. However, in practice, this assumption may not hold in many situations. As pointed out by Yen (2007), educational achievement tests are usually multidimensional, although they tend to have a strong principal domain. Not all of the links between different grade tests are strong enough to maintain a robust connection between those grades.

Furthermore, vertical scaling is a complex procedure. Previous research (e.g. Camilli et al. 1993; Petersen et al. 1983; Custer et al. 2006; Hanson and Béguin 2002; Hendrickson et al. 2006; Ito et al. 2008; Kim and Cohen 1998; Pomplun et al. 2004; Tong and Kolen 2007; Wingersky et al. 1987) has shown that vertical scaling results depend on many factors, such as the linking method and the IRT model used, the ability/difficulty estimation method employed and the design of the data collection used in the construction of the scale. A number of important decisions need to be made during the construction of the scale, and the combinations of these decisions probably result in somewhat different vertical scales.

Ito et al. (2008) used real data from a national standardisation assessment study and compared two vertical scaling approaches—concurrent and separate grade-groups linking—for grades kindergarten through 9 for reading and mathematics. They found that reading is more likely than mathematics to have a single prevalent trait across grades because similar results were generated at more grades in reading than in mathematics. The two approaches produced similar results in terms of item difficulties, discriminations and ability estimates. However, the separate grade-groups scaling had better control in terms of scale expansion than did concurrent scaling. Thus, an increase in the score variance at the highest and lowest grades is more salient for concurrent scaling than for separate grade-groups scaling. Kim and Cohen (1998) also found that similar results were generated by concurrent and separate methods except that the separate method provided more accurate estimates when the number of common items was small. In contrast, some research (e.g. Petersen et al. 1983; Wingersky et al. 1987) found that concurrent estimation was better than separate estimation. Hanson and Béguin (2002) also found that concurrent estimation outperformed separate estimation by generating a lower error in most conditions.

Pomplun et al. (2004) compared scaling results from WINSTEPS (Linacre 2011) and BILOG-MG (Zimowski et al. 1996) with both real and simulated data. WINSTEPS and BILOG-MG differ in two respects: WINSTEPS uses joint maximum likelihood estimation (JMLE) as the estimation method, whereas BILOG-MG uses marginal maximum likelihood estimation (MMLE). BILOG-MG also has a group option during estimation, whereas WINSTEPS has not. The findings of concurrent calibration showed that WINSTEPS generated more accurate individual and mean estimates, whereas BILOG-MG produced more accurate standard deviations. In another similar study, Custer et al. (2006) further compared results generated with WINSTEPS and BILOG-MG. Based on simulated vocabulary tests, they conducted vertical scaling with the Rasch model for grades kindergarten through 10. They used a common item block design and concurrent calibrations for scaling. Their results suggested that the convergence setting in the program was an important factor that influenced the parameter estimation. BILOG-MG generated more accurate individual and mean estimates than did WINSTEPS under default convergence settings. Tightened convergence settings enabled both programs to produce more accurate estimates than did default convergence settings. Furthermore, under tightened convergence settings, WINSTEPS and BILOG-MG produced similar scaling results. They recommended using MMLE with the direct group option of BILOG-MG to estimate group parameters in concurrent vertical scaling.

Tong and Kolen (2007) employed two data collection designs: the scaling test (SC) design and the common-item (CI) design. Under the SC design, the scaling test was calibrated concurrently while the tests for different levels were separately calibrated, and then these calibrations for the different levels were placed on the common scale. In the CI design, grade 3 was chosen as the base grade, and the other grades were separately calibrated to the grade scale. The results, in line with Hendrickson et al.'s (2006) research, found that the base grade chosen for vertical scaling under the common-item design had no substantial impact on the scaling results. In other words, choosing the lowest grade or the highest grade or the middle grade had little impact on the final scale results. However, Tong and Kolen (2007) noted that using as few links as possible might reduce the extent of scale shrinkage, which is common in vertical scaling with IRT models. Therefore, using a middle grade instead of the lowest or highest level as the base grade might be a better choice. The results also showed that the choice of scaling design has an important impact on the scaling result. Estimated student growth under the CI design was greater than that under the SC design. The parameter estimates generated by the SC design were more accurate. The multiple linking involved in the CI design possibly introduced more linking errors. The results also indicated that the real data were sensitive to the scaling procedure because many assumptions imposed by scaling methods were not met in the real data. The different scaling methods generated different scaling results for real data. However, the simulated data showed great tolerance to variation in the scaling methods. The different scaling methods produced very similar scaling results for simulated data.

In sum, vertical scaling is a complex procedure, which is influenced by many factors. Researchers usually determine the vertical scaling procedure according to their own situations and purposes. There is no agreement in the literature with regard to which approach generates the "best" vertical scales. Scale developers should make their own decisions based on their conception of estimated student growth and the nature of the scale to be developed.

10.4 Mathematics Competency Vertical Scale

In spite of the complexity of scale construction and the lack of consensus on the optimal approach, vertical scales are still attractive to researchers and test publishers. The Mathematics Competency Vertical Scale (MCVS) was created to measure the development of competency of Hong Kong students in mathematics; the scale utilises real data from 9,531 students between Primary 2 (P2 or grade 2) and Secondary 3 (S3 or grade 10). The MCVS was built using a new approach, the concurrent-separate approach, under the Rasch model. Both concurrent and separate calibrations were used at different stages of the vertical scaling procedure.

The MCVS covers a wide range of mathematical developmental competencies from P2 to S3. Two assessment booklets were designed for each grade to measure the mathematical competencies of students who had just completed their first semester (e.g. P2_1, P3_1.) and the competencies of those who had completed the second semester (e.g. P2_2, P3_2). The MCVS comprises 16 measurement booklets, with each pair of adjacent booklets (e.g. P2_1 and P2_2, P2_2 and P3_1, P3_1 and P3_2) having several common items through which all of the papers are interlinked. Figure 10.1 depicts the assessment design for the scale.

The number of items in each measurement booklets ranges from 29 to 42. As indicated by the overlap between the blocks in Fig. 10.1, there is a set of common items in the adjacent booklets. The number of common items for each booklet ranges from 4 to 14. All of the items in the booklets were developed according to the Mathematics Curriculum Guide (P1–P6) (Hong Kong Education Bureau 2000) and the Syllabuses for Secondary Schools–Mathematics (Secondary 1–5) (Hong Kong Education Bureau 1999). There are three types of items: multiple-choice questions, short questions requiring a brief answer and open-ended questions requiring steps and reasons for the answer. All items in the booklets for the primary students are grouped into five content strata: numbers, measures, shapes and spaces, data handling and algebra. All of the items in the booklets for the secondary students are grouped into three content strata: number and algebra, measure, and shape and space.

In the common-item design, the quality of the common items is important, and they should be considered carefully from both content and statistics perspectives. Lack of examination of the quality of the common items probably leads to unsatisfactory scaling results. In the design of MCVS, all of the common items were designed according to the suggestions provided by previous research (e.g. Kolen and Brennan 2004; Patz and Yao 2007). They argued that the common items should

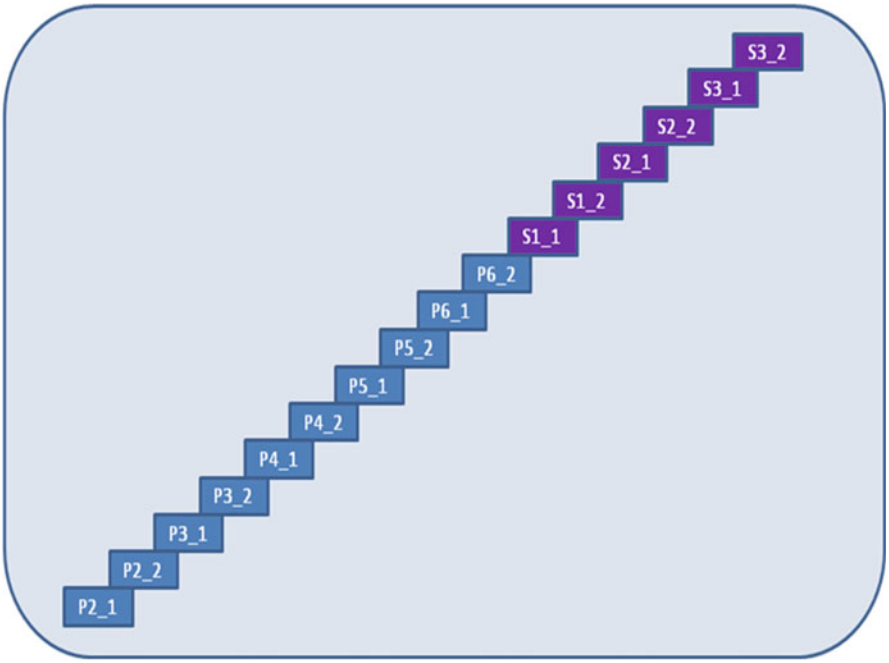


Fig. 10.1 Assessment design for the scale

(1) be appropriate in difficulty for the adjacent grades linked through the common items; (2) be representative of the whole test in terms of the representation of standards, the range of difficulty and the item's format; and (3) be in a similar position with the same appearance across test papers.

All of the data were collected in the academic year 2006–2007. The tests for the first semester (i.e. P2_1, P3_1, etc.) were administered in December 2006 or January 2007, and the tests for the second semester (i.e. P2_2, P3_2, etc.) were administered in May or June 2007. The study sample comprised 5,755 primary students enrolled in grades P2 through P6 from 24 schools and 3,776 secondary students enrolled in grades S1 through S3 from 11 schools in Hong Kong. The sample size for each booklet varied with a range from 177 to 1,405. According to Kolen and Brennan (2004), most of the booklets have a sufficient number of examinees (more than 400) for vertical scaling with the Rasch model. The number of items for each booklet and the number of participants who completed each booklet are presented in Table 10.1.

As discussed earlier in this chapter, both concurrent and separate linking methods have advantages and disadvantages. The separate method calibrates the parameters for items and individuals grade by grade and, thus, suffers from measurement error. Since for the calibration at each grade, there is estimation error and the error might be cumulative across the calibrations for different grades, more rounds of

Table 10.1 The item and participant distribution for booklets

Booklet	Number of items	Number of participants
P2_1	47	659
P2_2	42	650
P3_1	31	515
P3_2	35	514
P4_1	36	380
P4_2	36	382
P5_1	36	862
P5_2	36	756
P6_1	35	495
P6_2	36	542
S1_1	29	382
S1_2	35	227
S2_1	31	1,405
S2_2	34	1,393
S3_1	31	192
S3_2	32	177
Total	562	9,531

calibrations might imply greater cumulated error. This may explain why some research (e.g. Ito et al. 2008) has reported that as the grade deviates from the base grade, the best-fit linear line through the pairs of item discriminations start to rotate away from the identity line. In contrast, the concurrent method calibrates all of the parameters simultaneously in one analysis and, therefore, minimises the errors associated with calibrations. However, Hanson and Béguin (2002) noted that concurrent calibration imposes more constraints on item parameter estimates than the separate method, especially when calibrating many forms of tests at the same time, and that this might contaminate the resulting scale. Kolen and Brennan (2004) further pointed out that although, in theory, concurrent calibration that makes full use of all available information might be preferable, additional considerations, including violation of the unidimensionality assumption, might favour separate calibration.

Considering the inherent defects of using the single method, either concurrent or separate, to create a vertical scale, we adopted a combination of the two approaches, i.e. concurrent-separate. The concurrent and separate methods were carried out at different stages. This approach was partially inspired by that proposed by Wright (1996) and elaborated on by Wolfe and Chiu (1999) who measured the changes in person or item estimates across different times. To disentangle changes in persons (or items) from changes in items (or persons) in the measurement context, Wolfe and Chiu (1999) stacked the data collected from different time occasions together and obtained a set of category threshold calibrations of a rating scale that were shared by all time occasions. These threshold calibrations provided a unique and stable framework in which person and item estimates for each time occasion were calibrated. In addition, in the same framework, all person and item estimates could be compared and the development in individual abilities or changes in item difficulty could be interpreted.

The procedure for constructing MCVS consists of three steps which are illustrated in the following section.

10.4.1 Step 1: Identify Qualified Linking Items

The main purpose of this step was to identify quality linking items that are invariant in item difficulty across adjacent grades. For each grade, two rounds of analyses were undertaken. The first round of analysis was to identify the underfit persons whose OUTFIT or INFIT MNSQ were larger than 2.0 because they have a negative impact on the construction of the scale (Linacre 2011). The second round of the analysis was conducted by excluding all underfit persons identified in the first round of the analysis. Each linking item has two estimates of difficulty, one for each of the two adjacent grades. Two criteria were used to examine the quality of the lining items: the goodness of fit to the Rasch model and the invariance across adjacent grades. The linking items were disqualified and treated as different items in subsequent steps if any of the criteria below was satisfied.

(1) The item's OUTFIT or INFIT MNSQ was less than 0.5 or larger than 1.5; and
(2) The standardised difference of the item difficulties for adjacent grades was larger than 2.0, and the actual difference of the item difficulties was larger than 0.5 logits.

Any overfit (OUTFIT or INFIT MNSQ was less than 0.5) or underfit (OUTFIT or INFIT MNSQ was larger than 1.5) items were disqualified as linking items because of their misfit to the Rasch model. The items identified by the second criterion were also disqualified as linking items because they are not invariant in terms of item difficulty across grades.

As a result, 37 linking items were identified as quality linking items and used in the following steps.

10.4.2 Step 2: (Concurrent Analysis) Obtain the Item Measures for the Quality Linking Items

The main purpose of this step was to obtain the difficulty estimates for the quality linking items identified in step 1. All of the data from different grades were stacked together. The data for the quality linking items were placed in the same column, and the disqualified linking items were treated as different items. Rasch analysis of the stacked data was conducted. Similar to step 1, two rounds of analyses were undertaken. The first round of the analysis identified the underfit persons whose OUTFIT or INFIT MNSQ were larger than 2.0, and the second round of the analysis without the underfit persons identified in the first round of the analysis calibrated the difficulty estimates for all the quality linking items. As the quality linking items were calibrated based on the whole data set, the results yielded a shared framework for the following separate calibrations.

10.4.3 Step 3: (Separate Analysis) Obtain the Item Measures for All Items and Construct the Scale

In this step, separate analyses for each grade were conducted with the quality linking items anchored at the value that had been calibrated in step 2 to generate item measures for all of the items. Similar to the previous steps, the first round of the analysis was undertaken to identify the underfit persons whose OUTFIT or INFIT MNSQ were larger than 2.0, and the second round of the analysis without the underfit persons identified in the first round of the analysis was used to calibrate the difficulty estimates for all of the items. Any items showing misfit to the Rasch model, i.e. the OUTFIT or INFIT MNSQ was larger than 2.0, were removed from the scale. Eight items were identified by this criterion and removed. Furthermore, any items with extremely high or low difficulty were investigated by experts specialised in mathematics to determine whether they were appropriate for inclusion in the scale. Consequently, four items were removed because their difficulties were not appropriate for the corresponding grades. The remaining items comprised the MCVS.

The final version of the MCVS consists of 510 unique items. The details of each final booklet and the whole scale are presented in Table 10.2.

It can be seen that the mean item measures for each booklet ranged from 27.5 (P2_1) to 68.4 (S3_2). These values for the item measures (the second column in Table 10.2) are neither students' raw scores on assessment booklets nor the Rasch calibration in logits: they are *units* in the Rasch analysis, and the meaning of the units depends on the settings in the Rasch analysis. In this case, the mean of item difficulty across all items was set to 50, and one logit was divided into 10 units in the concurrent analysis conducted in step 2. Therefore, one unit of item measured in this method stands for 0.1 logit. Consequently, the mean test difficulty for the booklet ranged from 2.75 logits (27.5/10) for P2_1 to 6.84 logits (68.4/10) for S3_2. In other words, the whole scale covered a difficulty range of 4.09 logits for 7.5 schooling years of development (from the first semester of P2 to the second semester of S3), resulting in 0.55 logits per year. This amount of advancement in difficulty level of items from year to year is consistent with children's development because many studies of their development have shown that it is typical for a child to gain 0.5 logits growth within 1 year.

It can also be seen from Table 10.2 that each booklet had quite good Rasch reliability, ranging from 0.97 to 1.00. The separation index of the booklets ranged from 5.88 to 18.43. The statistical data provide strong confidence in the practical application of the MCVS scale. Figure 10.1 presents the item distribution by grades for the MCVS.

Table 10.2 The details of the MCVS

Booklet	Mean item measure	S.D. of item measure	Max. item measure	Min. item measure	Separation index	Rasch reliability	No. of items	No. of quality linking items
P2_1	27.54	15.05	69.33	−1.58	12.77	0.99	47	1
P2_2	30.73	13.31	52.2	3.03	12.47	0.99	42	2
P3_1	40.43	13.78	67.52	22.23	10.41	0.99	30	5
P3_2	42.66	13.09	74.92	13.93	10.10	0.99	35	7
P4_1	45.38	10.78	73.15	28.65	7.94	0.98	36	8
P4_2	49.00	14.12	75.52	23.21	9.35	0.99	35	7
P5_1	53.38	12.49	73.13	19.81	13.02	0.99	35	4
P5_2	59.85	10.08	76.03	39.88	10.42	0.99	36	3
P6_1	60.34	7.95	77.28	42.94	6.83	0.98	35	8
P6_2	61.47	8.10	81.41	45.13	7.26	0.98	35	5
S1_1	61.61	15.64	97.65	32.04	10.63	0.99	27	3
S1_2	64.21	22.60	101.84	27.95	5.88	0.97	32	6
S2_1	63.28	13.23	86.88	33.67	18.43	1.00	31	6
S2_2	66.84	12.62	108.69	41.11	16.04	1.00	34	4
S3_1	67.80	16.20	101.68	39.75	5.95	0.97	28	5
S3_2	68.36	13.44	97.65	48.51	5.90	0.97	32	3

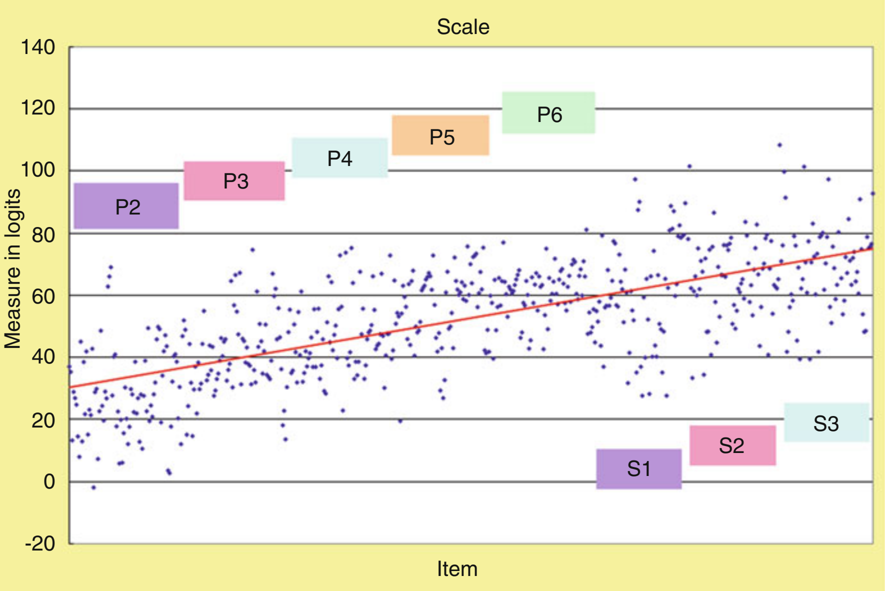


Fig. 10.2 Item distribution of the MCVS

Each dot in Fig. 10.2 stands for a single item. The items are grouped by their grades and placed along the *x* axis from the left to the right. The *y* axis represents item difficulty. It can be seen that, in general, the item difficulty advanced gradually from the lower grades to higher grades. The red solid line is a regression line that indicates that the item difficulty could be predicted, to some extent, by the grade where the item is placed. The *R* square was equal to 0.456, which is far from perfect prediction, but still substantial.

As the item difficulties are on the same scale as person ability, teachers, parents or anyone who wishes to measure students' achievement levels in mathematics could use items from the scale according to the students' mathematics abilities or their grades to form a test, administer the test to the students and analyse the test results under the Rasch model with the items anchored at the values provided by the scale. Thus, the students' mathematics competencies can be calibrated along the scale. More importantly, the competency estimates of the students from different grades could be compared directly, even though they were assessed by totally different sets of items because the items had been calibrated along the same scale, which provides a stable framework for the comparison. Consequently, students' growth in mathematics competencies could be tracked from P2 to S3 with the MCVS.

As noted earlier, all of the items in the MCVS are grouped into five content strata for the primary levels and three content strata for the secondary levels (all of the strata belong to the same dimension, i.e. overall mathematics competency). It can be seen from Figs. 10.2 and 10.3, which illustrate the item distribution by strata

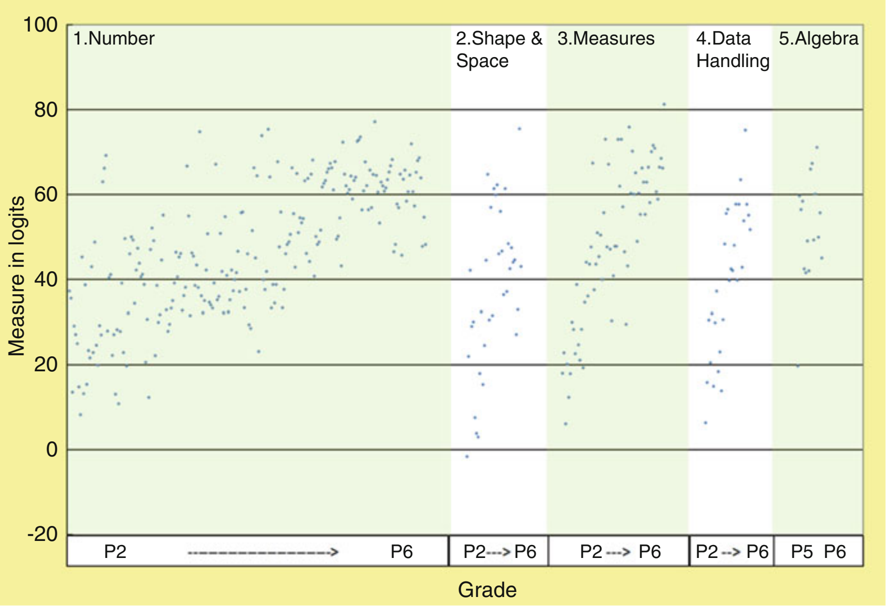


Fig. 10.3 Items by strata of the MCVS (primary)

for the primary levels and the secondary levels, respectively, that the items in each content strata cover quite a wide range of difficulties. The item difficulty advances gradually with grades for each stratum. Such a trend is especially salient for strata at primary levels.

The results presented in Figs. 10.3 and 10.4 indicate that the MCVS could be divided into sub-scales according to the content strata. The items belonging to the same strata could be selected and used to measure students' competencies in a particular mathematical domain, i.e. numbers, measures, shapes and spaces, data handling and algebra for primary levels and number and algebra, measure, and shape and space for the secondary levels. Thus, tracking the students' development in mathematics could be done in a more detailed way.

In sum, the MCVS was built under the Rasch model with a concurrent-separate approach, which incorporates the strength of both concurrent and separate methods. First of all, a separate analysis was conducted to investigate the quality of all of the linking items and identify those items that could be fitted to the Rasch model and invariant in terms of difficulty. The concurrent analysis was then utilised to calibrate the difficulty estimates of the quality linking items and to provide a stable and unambiguous framework for the construction of the scale. With those quality linking items anchored at the values obtained in the concurrent calibration, a separate analysis was undertaken for each booklet to calibrate the difficulty estimates of all of the items and, thus, form the whole scale. Furthermore, the impact of underfit persons was taken into account during the scale construction, and all persons with too large INFIT

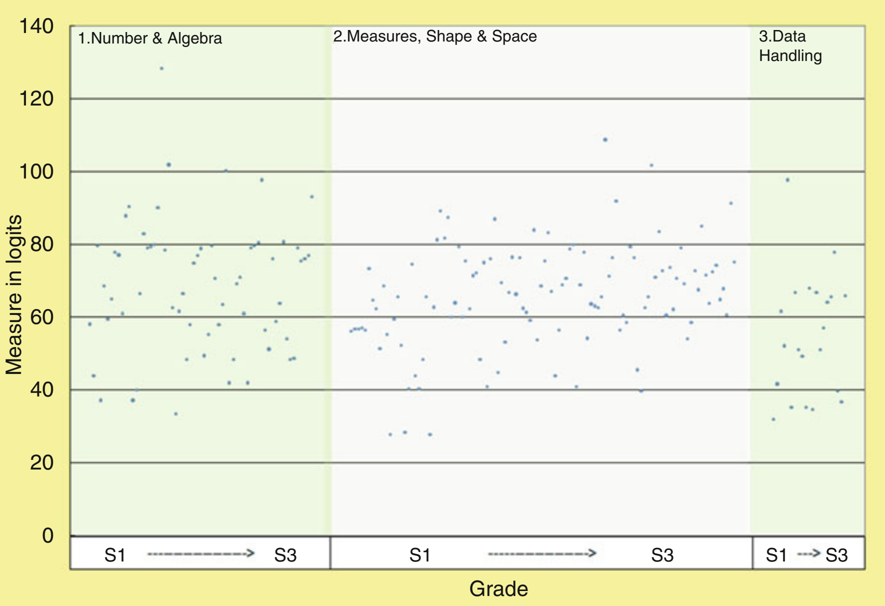


Fig. 10.4 Items by strata of the MCVS (secondary)

or OUTFIT MNSQ were excluded from each round of the analysis, and the "best sample" was used to construct the scale. The resulting scale comprised 16 booklets with a total of 510 items, encompassing P2 to S3 grades. The mean test difficulty for the booklet ranged from 2.75 logits for the first semester of P2 to 6.84 logits for the second semester of S3. Each booklet showed quite good Rasch reliability (ranging from 0.97 to 1.00) and separation index (ranging from 5.88 to 18.43). The properties of the MCVS make it a suitable vertical scale for tracking Hong Kong students' development in mathematics, or in particular domains of mathematics, over time.

Of course, this scale has some limitations in common with all other vertical scales. Previous research (e.g. Harris 2007; Kolen and Brennan 2004; Patz and Yao 2007) emphasised that the common items determine the quality of the constructed scale because all item parameters are estimated based on common items. The pilot study of the current research also showed that a minor change in the linking items (e.g. adding/deleting/changing even only one linking item) has quite a large impact on the calibration of the other items, especially when the number of linking item is small. Thus, this research examined the quality of the linking items from both content and statistics perspectives. Only those linking items that met several prior requirements, such as sufficient goodness of fit to the Rasch model, invariant in terms of item difficulty across grades and appropriate in terms of content were retained. As a result, there were too few qualified linking items for some grades, especially for P2_1 and P2_2. Most of the linking items had to be disqualified because they were not invariant across adjacent grades in terms of difficulty. This research highlights the fact that the linking items should be trait-related but not curriculum-related. Thus, students' performance on linking items should be determined by the trait measured but not by whether they have learned the content in the classroom. If the linking items are overly linked with the curriculum, the linking items will be easy for students who studied with a curriculum that includes knowledge required to solve the items and difficult for students who studied with another curriculum that does not include such knowledge. The difference in curriculum coverage will in turn lead to a large standardised difference in item difficulty. Further studies are needed on the characteristics of quality linking items to shed light on how researchers should select linking items in the construction of vertical scales.

References

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice Winter, 28*(4), 3–14.

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement, 36*, 73–78.

Camilli, G., Yamamoto, K., & Wang, M. M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379–388.

Custer, M., Omar, M. H., & Pomplun, M. (2006). Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG. *Applied Measurement in Education, 19*(2), 133–149.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24.

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233–251). New York: Springer.

Hendrickson, A., Cao, Y., Chae, S. E., & Li, D. (2006, April). *Effect of base year on IRT vertical scaling from the common-item design*. Paper presented at the National Council on Measurement in Education, San Francisco, CA.

Hong Kong Education Bureau. (1999). *Syllabuses for secondary schools: Mathematics (Secondary 1–5).* Retrieved on August 12, 2010, from http://www.edb.gov.hk/index.aspx?nodeID=4905&langno=1

Hong Kong Education Bureau. (2000). *Mathematics curriculum guide: P1–P6.* Retrieved on August 12, 2010, from http://www.edb.gov.hk/index.aspx?nodeID=4907&langno=1

Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education, 21*(3), 187–206.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*, 131–143.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Linacre, J. M. (2011). *Winsteps* (Version 3.72.3) [Computer software]. Chicago: Winsteps.com.

Lissitz, R. W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation, 8*(10). Retrieved on August 11, 2010, from http://PAREonline.net/getvn.asp?v=8&n=10

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253–272). New York: Springer.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137–156.

Pomplun, M., Omar, M. H., & Custer, M. (2004). A comparison of WINSTEPS and BILOG–MG for vertical scaling with the Rasch model. *Educational and Psychological Measurement, 64*, 600–616.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227–253.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration* (ETS Research Rep. 87–24). Princeton: Educational Testing Service.

Wolfe, E. W., & Chiu, C. W. (1999). Measuring change across multiple occasions using the Rasch rating scale model. *Journal of Outcome Measurement, 3*(4), 360–381.

Wright, B. D. (1996). Time 1 to time 2 comparison: Racking and stacking. *Rasch Measurement Transactions, 10*(1), 478–479.

Yen, W. M. (2007). Vertical scaling and no child left behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–282). New York: Springer.

Yen, W. M. (2009). *Growth models approved for the NCLB growth model pilot*. Unpublished manuscript.

Yon, H. (2006). *Multidimensional item response theory (MIRT) approaches to vertical scaling*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Chicago: Scientific Software International.

Chapter 11
Student-Problem Chart: An Essential Tool for SLOA

Magdalena Mo Ching Mok, Sze Ming Lam, Ming-Yan Ngan, Jing Jing Yao, Michael Ying Wah Wong, Jacob Kun Xu, and Stephen Yin Chuen Ting

11.1 Introduction

Recent research has consistently highlighted the importance of quality feedback for learning and academic achievement (Black and Wiliam 1998, 2009; Hattie and Timperley 2007; Shute 2008). Quality feedback enables the teacher to understand students' learning progress, diagnose their strengths and weaknesses in learning, and gauge the effectiveness of teaching strategy. From this feedback, teachers can then adjust their instruction so that it aligns better with the learning state of their students. Feedback is also helpful to the student: It can act like a mirror for the student to understand more about themselves as learners, supports

M. Mo Ching Mok (✉)
Department of Psychological Studies, and Assessment Research Centre,
The Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, N.T., Hong Kong
e-mail: mmcmok@ied.edu.hk

S.M. Lam • M.Y.W. Wong • J.K. Xu
Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong
e-mail: lamsm@ied.edu.hk; mywwong@ied.edu.hk; Jacobxu@ied.edu.hk

M.-Y. Ngan
Department of Curriculum and Instruction, The Hong Kong Institute of Education,
Tai Po, Hong Kong
e-mail: myngan@ied.edu.hk

J.J. Yao
Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong

Department of Psychology, Zhejiang Normal University, Jinhua, China
e-mail: jingjing@ied.edu.hk

S.Y.C. Ting
Formerly Assessment Research Centre, The Hong Kong Institute of Education,
Tai Po, Hong Kong
e-mail: stephentyc@hotmail.com

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_11, © Springer Science+Business Media Dordrecht 2013

student metacognition, and increases students' learning motivation. Specific feedback on the quality of assessment items can strengthen teachers' skills in item setting. This is usually done using item analysis methods. However, traditional item analysis only focuses on item difficulty, item discriminability, test reliability, and test validity – there is no information relating student responses to item quality in traditional item analysis. Unfortunately, the mathematics involved in modern item response theory (Wu 2012, this volume) might be daunting for some teachers and deter them from using assessment feedback to enhance assessment items (Ho et al. 2012, this volume). The purpose of this chapter is to present the student-problem chart (SP chart) as a user-friendly and efficient alternative for teachers to use to obtain invaluable feedback regarding student performance as well as item quality.

11.2 The Rationale Behind an SP Chart

Originally created by Takahiro Sato in the 1970s, the SP chart as an assessment *for* learning tool (Sato 1980, 1984, 1985) has benefited from the attention of a number of researchers in developing its theory and application. Most notable is Harnisch, who introduced the method to students in the USA and Hong Kong in the early 1980s (e.g. Connell and Harnisch 2004; Harnisch 1981, 1983; Harnisch and Linn 1981; Harnisch and Romy 1985; Linn and Harnisch 1981). Further applications of the SP chart can be found in the work by Chacko (1998), Dai et al. (2005), Dinero and Blixt (1988), Ngan (2011), Yu (2002), and others. The SP chart uses a number of indices including the disparity index, homogeneity index, item modified caution index, and student modified caution index to diagnose if a student's responses to a test are unusual (Harnisch and Linn 1981). At the same time, the SP chart carries diagnostic information about the extent to which each assessment item attracts normal or aberrant response patterns from the candidates.

11.2.1 Student Curve (S-Curve)

The construction of the SP chart (Chacko 1998; Harnisch 1983; Sato 1980, 1984, 1985) is actually very simple. First, students' responses to a set of assessment items are recorded in a raw student–item response matrix, with each row representing individual students' responses and each column representing responses to individual items. A score of 0 is given to an incorrect response, and a score of 1 is given to a correct response. From the raw student–item response matrix, the total number of items scored correctly by each student can be computed (row total), and the total number of students answering an item correctly can be computed for each item (column total). Based on the row totals, the raw student–item response matrix can be rearranged such that the students are arranged in descending order according to

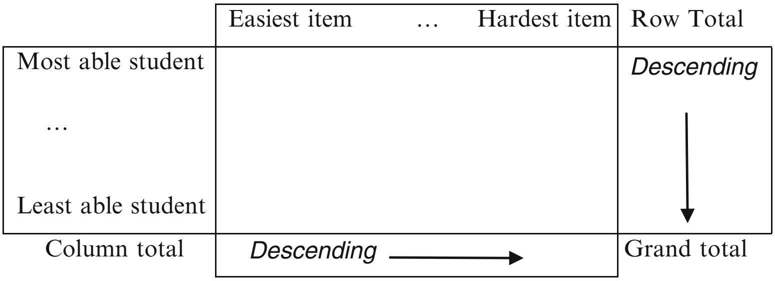


Fig. 11.1 Student–item response matrix

Item / Student	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Row total
A	1	1	1	0	1	1	▌1	6
B	1	1	1	1	1	▌0	0	5
C	1	1	1	1	▌0	0	0	4
D	1	0	1	1	▌1	0	0	4
E	0	▌1	0	0	0	0	0	1
F	1	▌0	0	0	0	0	0	1
Column total	5	4	4	3	3	1	1	21

Fig. 11.2 Rearranged student–item matrix of six students and seven items

their total score. Students with higher marks are placed at the top of the matrix, and students with lower marks at the bottom. Similarly, the columns are rearranged with items in ascending order from left to right. The resulting arrangement places more able students at top end of the student–item matrix (as far as the current assessment is concerned) and less able students at the lower end of the matrix. Similarly, items to the left of the matrix are easier than those to the right. An easier item is one which has more students answering it correctly. This is illustrated in Fig. 11.1.

Figure 11.2 displays the rearranged student–item matrix of a hypothetical situation involving six students attempting seven items. A student curve (S-curve) can be constructed based on the assumption that a student should be able to answer an easier item first before she/he can answer a more difficult item. For example, student A in Fig. 11.2 scored six out of seven items correctly; thus, one could count the six easier items on the left side of the rearranged student–item matrix and draw a vertical thick line to indicate that student A is expected to answer items Q1, Q2, Q3, Q4, Q5, and Q6 correctly and get the more difficult item Q7 wrongly.

Similarly, a vertical thick line can be drawn between Q5 and Q6 for student B to indicate the expectation that student B should get items Q1 to Q5 correctly and both Q6 and Q7 wrongly. The same operation can be applied to all students (Fig. 11.2). These student vertical expectation lines for the students can then be joined together to form the S-curve in Fig. 11.3.

Item Student	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Row total
A	1	1	1	0	1	1	1	6
B	1	1	1	1	1	0	0	5
C	1	1	1	1	0	0	0	4
D	1	0	1	1	1	0	0	4
E	0	1	0	0	0	0	0	1
F	1	0	0	0	0	0	0	1
Column total	5	4	4	3	3	1	1	21

Fig. 11.3 S-curve for six students and seven items

11.2.2 Student Modified Caution Index

It should be noted that in this example (Fig. 11.3), although student A is expected to get items Q1 to Q6 correctly and Q7 wrongly, student A does not perform entirely according to expectation. Student A got Q4 wrongly (score 0) but Q7 correctly (score 1). This may happen for a number of reasons, for example, the student was careless when responding to Q4, had good luck in answering Q7, made a good attempt to Q7 based on partial knowledge, was dishonest in answering Q7, etc. Regardless of the reason, the teacher should be concerned if the observed pattern of student responses deviates too much from expectation. A modified caution index (MCI) (Harnisch and Linn 1981; Tatsuoka 1984) for the *i*th student can be computed as follows:

$$\mathrm{MCI}_i = \frac{\sum_{j=1}^{n_{i.}} (1-u_{ij}) n_{.j} - \sum_{j=n_{i.}+1}^{J} u_{ij} n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - \sum_{j=J+1-n_{i.}}^{J} n_{.j}}$$

where

i is the i^{th} person,
j is the j^{th} item,
J is the total number of items in the assessment,
u_{ij} is the answer of the i^{th} student to the j^{th} item (correct answer = 1; wrong answer = 0),
$n_{i.}$ is the total score of the i^{th} student, and
$n_{.j}$ is the total score of the j^{th} item.

In theory, all student MCI values can range from 0 to 1. The higher the MCI value, the more caution should be used when interpreting the student's response patterns. Computation of the MCI can be easily done using the following simplified method illustrated using responses from student D: $\mathrm{MCI}_i = \frac{W-X}{Y-Z}$

$\text{MCI for student D} = \text{MCI}_D = \frac{W - X}{Y - Z} = \frac{4 - 3}{16 - 8} = 0.125$	
W = For all the items to the left of the S-Curve which are answered incorrectly by student D, add up the number of students who have correctly answered these items.	X = For all the items to the right of the S- curve which are answered correctly by student D, add up the number of students who also have correctly answered these items.
Student D answers Q2 incorrectly, which is on the left of the S-Curve. However, Q2 is answered correctly by four students , so W = 4.	Student D answers Q5 correctly, which is on the right of the S-curve. However, Q5 is answered correctly by three students, so X = 3.
Y = For all the items to the left of the S-Curve, add up the number of students who have correctly answered these items.	Z = Sum the number of students who answer the items to the right of the S-curve correctly, starting from the m^th^ item, where the m^th^ item = (total number of items in the test + 1 – student score).
There are four items (Q1, Q2, Q3, Q4) to the left of the S-curve. The numbers of students who answer these items correctly are 5, 4, 4, and 3, respectively, so Y= 5 + 4 + 4 + 3 = 16.	Student D score = 4; there are seven items in the test, so m^th^ item means (7 + 1 – 4 = 4) the 4th item. Add up the number of students who answer Q4, Q5, Q6, and Q7, correctly, so Z =3 + 3+ 1 + 1 = 8.

Fig. 11.4 Computation of MCI for student D

Students with same scores may have very different MCIs, depending on their response pattern in relation to the expected pattern. For example, student D gets an easier item (Q2) wrong while gets a harder item (Q5) correct, and so the MCI of student D is 0.125, using the method illustrated in Fig. 11.4. Student C, however, got all the easier items correct and failed to score on the harder items – i.e. their response pattern matches the expected pattern – and so their MCI is 0.00. The next question then is as follows: How large can the MCI be before the teacher needs to be concerned? The literature has different views on this question. In general, MCI values between 0.0 and 0.3 are considered normal, an MCI value between 0.3 and 0.5 indicates some aberrant responses of the student, and MCI values greater than 0.5 suggest rather abnormal response patterns and the teacher should take a closer look into the responses and perhaps follow up with the students concerned.

11.2.3 Student Types

The student's MCI can be used in combination with his/her total score to support the teacher in providing evidence-based feedback. Students can be broadly classified into four types:

Type A: A type A student is one who is performing well (getting 50% or more items correct) and whose MCI is low (less than 0.3). Their response pattern indicates that this student has a satisfactory and steady performance. To a certain extent, teaching and learning at this stage are effective.
Type B: A type B student is one who is performing well (getting 50% or more items correct) and whose MCI is high (equals 0.3 or higher). Their response pattern shows that although this student can answer relatively difficult items, she/he is not able to answer relatively easy items. This suggests that their learning is good (perhaps) but not stable. The teacher should check to see if the student really does have high ability but merely has been careless in this assessment or whether the student is actually of lower ability but has been lucky and dishonest, has learned more difficult topics elsewhere (e.g. at a tutorial school), or gets right answers for the more difficult items for some other reasons than ability.
Type C: A type C student is one who has performed poorly (less than 50%) and whose MCI is high (equals 0.3 or higher). The response pattern of a type C student is both unsatisfactory and unstable. The teacher should keep an eye on the learning progress of type C students. Their low performance together with high MCI suggests that their performance might be due to carelessness, insufficient academic preparation, poor understanding of subject content, or a lack of examination skills.
Type D: A type D student is one who has performed poorly (less than 50%) and whose MCI is low (less than 0.3). The learning of type D students is unsatisfactory but stable. They have not attained the required knowledge level, and their learning is incomplete. Teachers should help students to understand their learning weaknesses and support them to manage their learning problems in order to raise their learning capacity.

Figure 11.5 presents distributions of student types in three hypothetical schools, across two classes. These examples are inspired by some real cases discussed in Mok (2010). In Fig. 11.5, most of the students in class 1 at school P are type A. They scored 50% or higher in the assessment, and their MCI values were less than 0.3. Only two students scored below 50% and three others with MCI values greater than 0.3. Most students in this class are performing satisfactorily and with stability. Class 2, from the same school, is very different: Most students in class 2 at school P are type D – they scored below 50%, and their MCI values were less than 0.3. Teachers of class 2 may consider providing remedial support as these students are consistently not performing to the expected level. It is possible that the students in class 2 have not yet grasped the essence of the topic being taught and that the class is not ready to progress to the next topic at this stage. School administrators can use this information to consider differential resource allocation for the two classes; obviously class 2 needs more support.

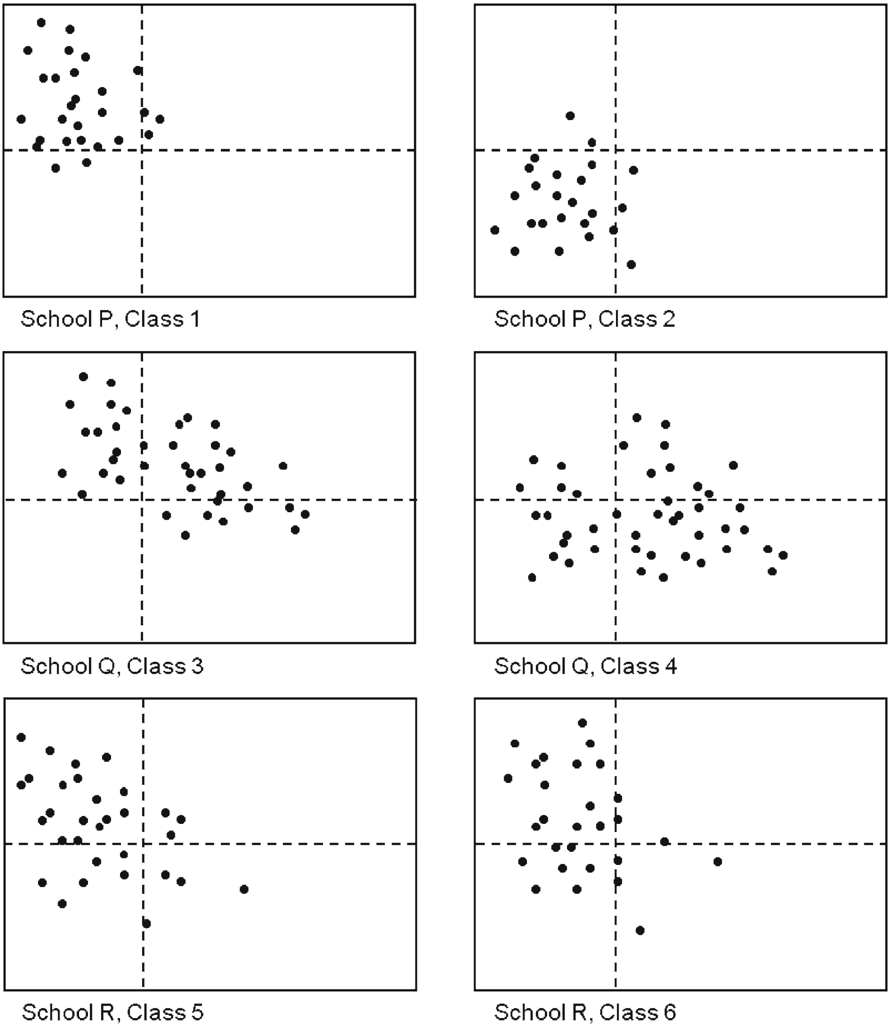


Fig. 11.5 Distribution of student types for three hypothetical schools (Note: In each of the scatter plots, the *x-axis* represents the MCI and the *y-axis* the percentage of correct answers)

Class 3, from school Q, comprises students who are performing satisfactorily. If the teacher only refers to the total score, however, she/he may misinterpret them to have mastered the topic being assessed. Analysing the performance x MCI scatter plot (Fig. 11.5) provides deeper insight revealing that many of the students have very high MCI values, categorizing them as type B students. This means that while they appear high performing, their performance is unstable. In reality, this type of student may not have totally mastered the topic, and there is the chance that she/he will fail items if she/he is under stress (e.g. in high-stake examination conditions) or if more difficult items are assigned in the test.

Item Student	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Row total
A	1	1	1	0	1	1	1	6
B	1	1	1	1	1	0	0	5
C	1	1	1	1	0	0	0	4
D	1	0	1	1	1	0	0	4
E	0	1	0	0	0	0	0	1
F	1	0	0	0	0	0	0	1
Column total	5	4	4	3	3	1	1	21

Fig. 11.6 Item curve for six students and seven items

Class 4, from the same school, has a slightly lower performance level than class 3, although there are some students who are performing well. However, of particularly note is that there are many students whose performances are unstable – in other words, there are both type B and type C students in this class. If both classes from the same school show unstable performances, then questions have to be asked. Maybe the school needs to be teaching better study habits. For example, do the students check their answers before handing in their papers? Or maybe an alternative assessment method will give more reliable results. For example, if only multiple-choice items were used in this assessment, the teacher might want to consider setting some items requiring constructed responses or include some two-tiered items (Tam et al. 2012, this volume) in order to solicit diagnostic insights from the assessment data about the students' state of knowledge.

Classes 5 and 6, from school R, have a very similar distribution of student types (Fig. 11.5). The two classes have similar averaged performances and MCI values. Most of the students in these two classes are of type A, although there are also a couple of types B, C, and D in each class. Many non-streaming schools would have similar class distributions to school R. Most of the students in the two classes have a steady and satisfactory performance and are ready to proceed to the next stage of learning. The majority of students' current approach to learning appears to be effective; for the few type D students, teachers can provide remedial support accordingly and follow up with more diagnostic assessment for the type C and type D students, where appropriate (see Tzuriel 2012, this volume).

11.2.4 Item Curve (P-Curve)

An item curve (or P-curve) can be constructed from the rearranged student–item matrix using similar logic as that used for the S-curve. That is, a P-curve can be constructed based on the assumption that for any item, more able students should have a higher probability of answering it correctly than less able students.

In the example presented in Fig. 11.6, five students answer item Q1 correctly, so a horizontal line (the wave line in the figure) can be drawn between the fifth and sixth students (students E and F, respectively) counting from the top. Using the logic that in the rearranged student–item matrix, students from the top are more able than those at the bottom of the column, it is expected that the five students above the horizontal line will answer the Q1 correctly, but the student below it will not. Similarly, a horizontal line can be drawn between the fourth and fifth students counting from the top of Q2 and of Q3 because these two items both had four students getting the right answer. The process can be repeated for all the items in the assessment. The P-curve is formed by joining these horizontal lines, as shown in Fig. 11.6. It can also be seen from Fig. 11.6 that some items have response patterns that conform more to the expected patterns than others. For instance, although Q2 and Q3 both have four students answer the item correctly, only Q3 conforms to the expected student response pattern. A closer analysis shows that while the top four students answered Q3 correctly, for Q2 it is the top three students plus the fifth student who answered correctly. If the response pattern conforms entirely to the expected pattern, then all students above the P-curve would have scores of 1 and all students below the P-curve would have scores of 0 – any score of 0 above the P-curve or any score of 1 below it represents deviation from expectation. Furthermore, the further a score of 0 is located above the P-curve or a score of 1 below the P-curve, the more serious is the deviation from expectation. Using this logic, Harnisch and Linn (1981) recommended a modified caution index (MCI) for items, which is given by the following mathematical expression (see also Fig. 11.7):

$$\mathrm{MCI}_{j} = \frac{\sum_{i=1}^{n_{j.}} (1-u_{ij}) n_{.i} - \sum_{i=n_{j.}+1}^{I} u_{ij} n_{.i}}{\sum_{i=1}^{n_{j.}} n_{.i} - \sum_{i=I+1-n_{j.}}^{I} n_{.i}}$$

where

i is the i^{th} person,

j is the j^{th} item,

I is the total number of students in the assessment,

u_{ij} is the answer of the i^{th} student to the j^{th} item (correct answer = 1; wrong answer = 0),

$n_{j.}$ is the total score of the j^{th} item (i.e. the total number of students answering this item correctly), and

$n_{.i}$ is the total score of the i^{th} student.

11.2.5 Item Types

Values of an item MCI can range from 0 to 1, with values of between 0 and 0.3 being considered acceptable. Like student MCIs, item MCI values between 0.3 and 0.5 indicate aberrant pattern, and the teacher should take care when interpreting results involving that item. If an item MCI is greater than 0.5, then the response

$$\text{MCI of item Q2} = \text{MCI}_{Q2} = \frac{W - X}{Y - Z} = \frac{4 - 1}{19 - 10} = 0.33$$

W = For all the students above the item-Curve who answered this item wrongly, add together the total scores of these students.	X = For all the students below the item-curve who answered the item correctly, add together the total scores of these students.
Students A, B, C and D are above the P-curve, and only student D answered Q2 incorrectly. Student D's total score is 4 and so W = 4.	Only student E is below the P-curve and answered Q2 correctly. Student E's total score is 1 and so X = 1.
Y = For all the students above the P-curve, add together their total scores.	Z = Sum the total scores of the students from the mth student to the last student, where the mth student = (total number of students in the test + 1 – item total score).
Students A, B, C and D are above the P-curve. Their total scores are 6, 5, 4 and 4, respectively, and so Y = 6 + 5 + 4 + 4=19.	Item Q2 total score = 4; there are six students in the test, so mth student means (6 + 1 – 4 = 3) the 3rd student. Summing up the number of total scores from the 3rd to the last student, inclusive, means that Z = 4 + 4 + 1 + 1 = 10.

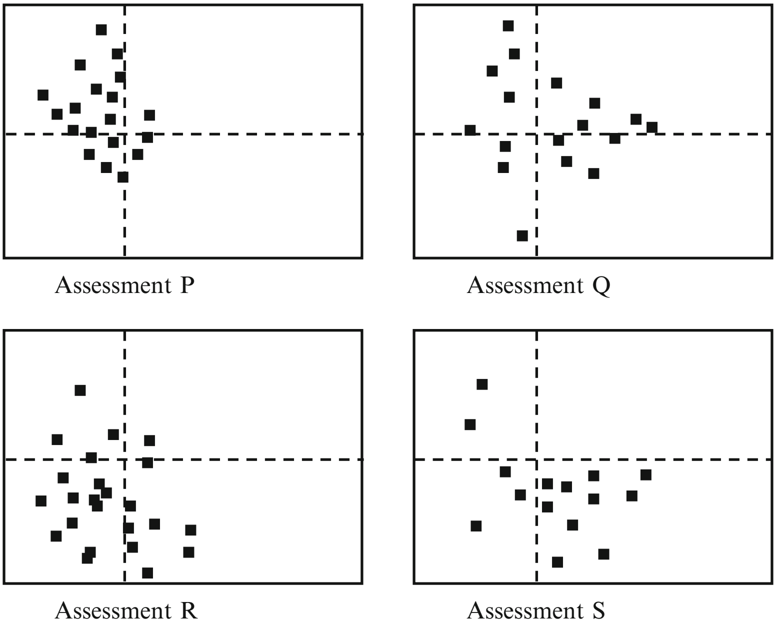
Fig. 11.7 Computation of MCI for item Q2

pattern for that item is very unusual and far from the expected pattern. When this happens, the teacher should be cautious about the item, inspect the item carefully, revise it if possible, and even delete it from the assessment if it is found to have severe defects or does not align well with the students' current state of knowledge.

Items can be categorized into four types according to their MCI and p values, where the p value is the percentage of students who answer a particular item correctly. The four categories are:

Type A: A type A item is one with a p value of at least 50% and a low MCI (less than 0.3.) Type A items have difficulty levels that align well with the abilities of students. These items can effectively assess the knowledge level and learning progress of students and can be used to distinguish students in terms of their ability levels. If a teacher sets items that align with his/her teaching progress and according to item-setting principles, most of the items in the assessment should belong to this type.

Type B: A type B item is one with a p value of at least 50% and a high MCI (0.3 or higher). Type B items attract wrong answers from able students, but less able students tend to get the right answers. There are many possible reasons why an item is type B. The most common reason is because the item is not well aligned



Assessment P

Assessment Q

Assessment R

Assessment S

Fig. 11.8 Item p value against item MCI of four hypothetical assessments (Note: the *x-axis* represents item MCI and *y-axis*, item *p*-value)

with teaching and learning or it may require other traits in addition to the trait being measured. For example, in a mathematics test, those items requiring high language ability might deter some students who are capable in math but weak in language.

Type C: A type C item is one with a p value less than 50% and a high MCI (0.3 or higher). Type C items are difficult items, and the difficulties tend to be due to ambiguous item expression and error(s) in the item or because the item is poorly aligned with the teaching and learning. Teachers need to revise type C items or even delete them from the assessment.

Type D: A type D item is one with a p value less than 50% and a low MCI (less than 0.3). These items are too difficult for most students, and because only students of high ability are able to answer type D items, these items cannot differentiate between students of middle and low abilities. Nevertheless, type D items are not necessarily of poor quality – they are just too difficult for that learning programme. Teachers should avoid giving too many type D items in one assessment because difficult items like these can discourage students.

Presented in Fig. 11.8 are scatter plots of item p values against item MCI values of four hypothetical assessments. Assessment P is good quality because most items are type A, and the items range from quite difficult to easy, with most of the items having difficulty levels aligning well with ability of students. Only three items have MCIs slightly greater than 0.3.

Hypothetical assessment Q (Fig. 11.8) is problematic because although the item p values are spread over a good range, many items have MCI values greater than 0.3. These are type B or type C items, which means response patterns to these items tend to be random. These items need to be refined.

Hypothetical assessment R (Fig. 11.8) is too difficult for the students (with many having p values below 0.5), even though many items in the assessment are of acceptable quality (MCI values less than 0.3). The pattern of responses indicates that the students have not reached the standards expected of the items. The teacher may need to revisit the target topics as well as lowering the difficulty level of the items.

Hypothetical assessment S (Fig. 11.8) is a poorly set assessment. Many of the items have p values less than 0.5 as well as MCI values exceeding 0.3. These are type C items, and heavy revision of the assessment is recommended before it is administered to students.

11.3 SP Xpress

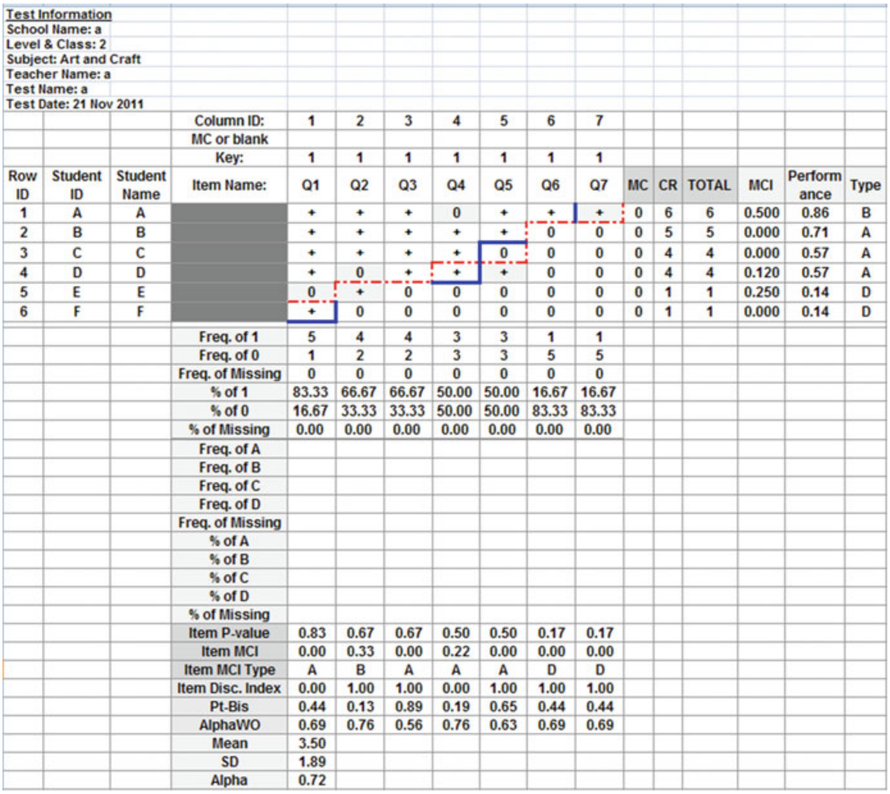
A piece of computer software entitled SP Xpress (version 2.2) (Mok et al. 2011) is now available for SP analysis. The following section will discuss the various outputs from SP Xpress.

11.3.1 S-Curve and P-Curve Output from SP Xpress

Presented in Fig. 11.9 is an example output from the assessment data discussed in Sect. 11.2.1. In the upper left of the output from SP Xpress (Fig. 11.9) is test information including school name, year level and class, subject, teacher name, test name, and test date. Following this information are four rows that give information about the items: the column IDs, an indication on whether each item is multiple choice or otherwise, the key for the items, and the items' names (e.g. Q1 … Q7). Student ID and name information are provided on the left of the table.

The responses of each student are displayed in the next rows of the output (below the school and test information, Fig. 11.9). Constructed-response items correctly answered are represented as '+', and those wrongly answered are shown as '0'. Multiple-choice items correctly answered are also turned to '+', but those wrong options chosen by students (i.e. A, B, C, or D) are listed for easy reference by the teacher.

It can be seen from Fig. 11.9 that the S-curve is printed as a solid thick line and the P-curve as a dotted line. Cells to the left of the S-curve and above the P-curve are expected to be fully filled by '+' since students in this area are of higher ability and items in this area are relatively easy. By contrast, cells to the right of the S-curve and below the P-curve are expected to be fully filled by '0', i.e. wrong multiple-choice



Test Information
School Name: a
Level & Class: 2
Subject: Art and Craft
Teacher Name: a
Test Name: a
Test Date: 21 Nov 2011

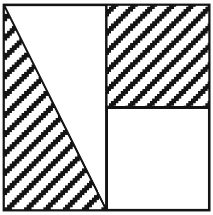
			Column ID:	1	2	3	4	5	6	7						
			MC or blank													
			Key:	1	1	1	1	1	1	1						
Row ID	Student ID	Student Name	Item Name:	Q1	Q2	Q3	Q4	Q5	Q6	Q7	MC	CR	TOTAL	MCI	Performance	Type
1	A	A		+	+	+	0	+	+	+	0	6	6	0.500	0.86	B
2	B	B		+	+	+	+	+	0	0	0	5	5	0.000	0.71	A
3	C	C		+	+	+	+	0	0	0	0	4	4	0.000	0.57	A
4	D	D		+	0	+	+	+	0	0	0	4	4	0.120	0.57	A
5	E	E		0	+	0	0	0	0	0	0	1	1	0.250	0.14	D
6	F	F		+	0	0	0	0	0	0	0	1	1	0.000	0.14	D
			Freq. of 1	5	4	4	3	3	1	1						
			Freq. of 0	1	2	2	3	3	5	5						
			Freq. of Missing	0	0	0	0	0	0	0						
			% of 1	83.33	66.67	66.67	50.00	50.00	16.67	16.67						
			% of 0	16.67	33.33	33.33	50.00	50.00	83.33	83.33						
			% of Missing	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
			Freq. of A													
			Freq. of B													
			Freq. of C													
			Freq. of D													
			Freq. of Missing													
			% of A													
			% of B													
			% of C													
			% of D													
			% of Missing													
			Item P-value	0.83	0.67	0.67	0.50	0.50	0.17	0.17						
			Item MCI	0.00	0.33	0.00	0.22	0.00	0.00	0.00						
			Item MCI Type	A	B	A	A	A	D	D						
			Item Disc. Index	0.00	1.00	1.00	0.00	1.00	1.00	1.00						
			Pt-Bis	0.44	0.13	0.89	0.19	0.65	0.44	0.44						
			AlphaWO	0.69	0.76	0.56	0.76	0.63	0.69	0.69						
			Mean	3.50												
			SD	1.89												
			Alpha	0.72												

Fig. 11.9 Output from SP Xpress (version 2.2)

options are chosen since students in this area are of lower ability and items in this area are relatively difficult. When this happens, the S-curve would overlay the P-curve perfectly, similar to the Guttman's (1950) perfect scale, but such a situation is rarely found in reality. Aberrant cells (i.e. with a 0 mark or wrong multiple-choice options to the left of the S-curve and above the P-curve or right answers (represented by +) to the right of the S-curve and below the P-curve) are shaded in the SP Xpress output for easy reference by the users.

11.3.2 *Information on Students' Academic Performance from SP Xpress*

On the far right of the SP chart output from SP Xpress is information about each student's academic performance which includes the number of multiple-choice items she/he correctly answered (MC), the number of constructed-response items correctly answered (CR), the total number of items correctly answered (TOTAL),



What is the fraction that best describes the shaded parts of the figure?

A. 1
B. 4
C. 1/2
D. 2/4

Fig. 11.10 A multiple-choice item with four options

their modified caution index (student MCI) and performance (Performance), and their student type, as classified according to the student's MCI (Type).

The teacher can inspect the information to get a better understanding of the knowledge of each student. For example, the analysis shown in Fig. 11.9 tells the teacher that although student A's total score is much higher than that of student B, the performance of student A is not stable. She/he has MCI value of 0.50 and belongs to type B. The MCI value of student A reflects the fact that although the student has answered six items out of seven correctly, the student has failed on Q4, which is of intermediate difficulty and easier than Q7, which student A has answered correctly. Based on his/her knowledge of the student, the teacher might want to explore this aberrant response pattern further in order to find out whether the error was due to misconception or carelessness.

11.3.3 Item Information and Item Statistics from SP Xpress

Above the SP chart, there is item information which includes a serial column ID (automatically generated by the software), item type (multiple choice or constructed response), a key for each item, and item name (Fig. 11.9).

Below the SP chart are statistics for each item: the number and percentage of students who have correct (scored as 1) and incorrect (scored as 0) answers for constructed-response items, the frequency and percentage of choice for each option of multiple-choice items, the item p value (i.e. the ratio of students who correctly answered the item), its modified caution index (item MCI), and what type the item is according to its MCI (Fig. 11.9). The teacher can use this information to refine the assessment as well as to get diagnostic information about the current understanding of the class. For example, if the multiple-choice options are designed carefully, each option can reveal a different misconception of the students. Presented in Fig. 11.10 is one such example.

In Fig. 11.10, students who choose option A may have come to this answer using the following logic: The shaded part of the left part of the figure is 1/2, and the shaded part of the right part of the figure is 1/2, so the total shaded part = 1/2 + 1/2 = 1. Those who choose option B either do not know what is meant by a fraction or do not understand what is asked by the question, for example, they may not understand what is meant by 'the shaded parts'. Those who choose option D probably have not grasped the concepts or skills of fraction simplification. If the SP Xpress output

Student proportion for each option within ability group			
Option / Ability	High	Middle	Low
A	0.00	0.00	0.00
B	10.00	10.00	0.00
C (key)	80.00	55.00	10.00
D	10.00	30.00	90.00
M (missing)	0.00	5.00	0.00
Column total	100	100	100

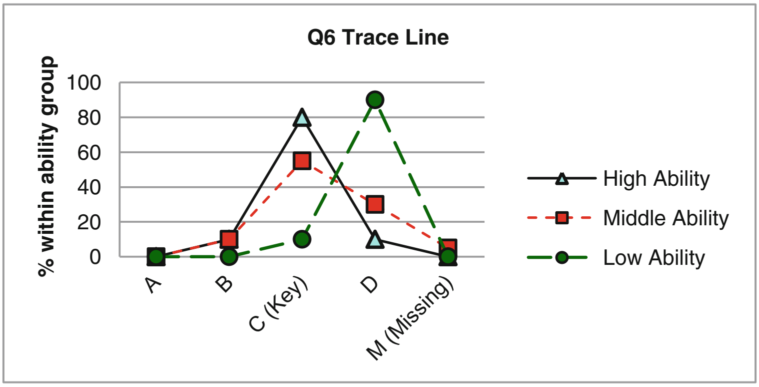


Fig. 11.11 Trace line analysis of Q6

shows that a large proportion of students have chosen option D, then the teacher can revisit the concepts and skills of fraction simplification, but if instead a large proportion of students have chosen option B, then the teacher may need to teach the basic concepts of fractions again.

To assist teachers to optimally extract diagnostic information from options of multiple-choice items, SP Xpress (Mok et al. 2011) also includes in its output trace line analysis for each item in the assessment. Students are first of all divided into three groups according to their total scores: (a) high ability students – the top 25% of students; (b) low ability students – the lowest 25% of students; and (c) middle ability students – the remaining middle 50% of students. Next, the proportion of students in each group who choose each option of a multiple-choice item and who correctly/incorrectly answer a constructed-response item are analysed by the programme. SP Express presents this analysis in both tabular and graphical formats.

A trace line analysis of an example item (Q6) from an assessment involving fictitious data from 40 students and 10 items is presented in Fig. 11.11. It can be seen from the trace line analysis (Fig. 11.11) that Q6 has some good qualities. First, the proportion of students choosing the right option (C) are in descending order of the ability of the students, being highest for the high ability group and lowest for the low ability group. Furthermore, one of the wrong options (D) attracted different proportions of students from different ability groups, being highest for the low ability group and lowest for the high ability group. In addition, another wrong option (B) attracted some students from both the high and middle ability groups. The teacher needs to reflect upon the possible reasons for the response patterns of the three ability groups. Lastly, there is no student choosing the wrong option A. Reflecting on the response patterns of the three ability groups provides insight for the teacher. For example, option A includes some basic knowledge that the teacher would like everyone to grasp before proceeding to more difficult concepts, then the data indicates that this goal has been reached.

Further down in the output of SP Xpress (Fig. 11.9) are item analysis indices and statistics for the entire assessment including the item discrimination index (item disc. index), point-biserial correlation coefficient (pt bis), Cronbach alpha of the assessment without the particular item (Alpha WO), mean of current assessment data (mean), standard deviation of the current data set (SD), and Cronbach alpha: reliability of the current assessment (alpha). The item discrimination index shows the extent to which the item can discriminate students' ability. It is a number with possible values ranging from −1 to 1. When the index is close to 0 from either end, the item has weak discrimination power. When the index is higher than +0.4, the item has high discriminatory power, which means it attracts the right answer from students of higher abilities but incorrect answers from students of lower abilities. When the index assumes a negative value, the corresponding item has negative discrimination power, which in turn means that it attracts incorrect answers from students of higher abilities but correct answer from students of lower abilities. The teacher might want to exclude such items when computing the total score for the assessment.

The point-biserial correlation coefficient shows the correlation between the item and the score of all other items in the assessment. Its value can also range from −1 to 1. A high point-biserial correlation coefficient means that students who correctly answered the item also got a high score in the test, while students who wrongly answered the item also got a low score in the test. Thus, the consistency of the test is high when all point-biserial correlation coefficients in the assessment are higher. An item with a negative point-biserial correlation coefficient means that the trait measured by that particular item might be different from the traits assessed by other items. This could be due to ambiguity in the question, different language requirements for the different items, or a misalignment between the item and the curriculum.

Teachers can gain a holistic view of the class and the alignment between the test and the class from the test mean and SD. The higher the mean value, the lower is the difficulty of the test in general. The higher the SD, the greater is the difference between students' levels. When both mean and SD are low, the items in the test might be too difficult or the teaching outcomes have not been achieved. When the mean value is average and the SD is high, this suggests that there is a wide range of abilities and levels of understanding within the class. The teacher might want to divide the class into different ability groups and address specific issues faced by the groups accordingly. The individual student MCI and performance statistics output from SP Xpress should be of great value under such circumstances.

The Cronbach's alpha shows the internal consistency of the assessment. Its value can range from 0 to 1. The higher the Cronbach's alpha value, the more internally consistent is the assessment. In general, the assessment is considered reliable (internally consistent) when the alpha value is higher than 0.7. Alpha WO shows the alpha value after a particular item is deleted. In general, it increases when items with low item disc. index and pt bis are deleted and vice versa. If after removing the item, the alpha WO value is very much greater than the alpha value, then the item is assessing a trait that is very different from the other items in the assessment.

11.4 Future Development of SP Chart

With the help of the SP chart, teachers can analyse the effectiveness of the assessment tools to discriminate students with different states of knowledge and adjust the assessment tools to meet the needs of different students. Moreover, teachers can provide one-to-one assistance to students according to each student's MCI, helping them to overcome specific difficulties and strengthen their learning skills. Thus, the diagnostic information generated by the SP chart can help teachers improve their assessments and so assist teachers to reach their ultimate goal of helping their students to achieve their learning targets.

The theory and application of the SP chart can be further developed along the following directions:

1. Analyse more deeply the aberrant items to find out why these items tend to attract abnormal response patterns and identify possible distinguishing characteristics of these items using approaches similar to cognitive diagnostic assessment (as discussed in earlier chapters of this volume; see de la Torre 2012, this volume).
2. Undertake further research on the learning habits of students with different types of student MCI. Identify contributing factors in relation to MCI student types including student learning ability, student characteristics, knowledge base, learning strategies, learning habits, and learning difficulties. This knowledge would assist teachers in the formulation of strategies to support student learning.
3. Promote the SP chart to schools for its general application for classrooms use. Our experience shows that the SP Xpress is an easy-to-use tool for the production of an SP chart and associated item and student statistics, and teachers should find SP Xpress very helpful as an assessment analysis tool.
4. Further refine and strengthen the current version (version 2.2) of SP Xpress (e.g. by including a disparity index (Yu 2002)), so that SP charts can be more commonly used to support teaching and learning.

References

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London: GL Assessment.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5–31.

Chacko, I. (1998). S-P chart and instructional decisions in the classroom. *International Journal of Mathematical Education in Science & Technology, 29*(3), 445–450.

Connell, M., & Harnisch, D. (2004). SP charts: Creating a longitudinal view of a technology enabled intervention. In C. Crawford, D. Willis, R. Carlsen, I. Gibson, K. McFerrin, J. Price, & R. Weber (Eds.), *Proceedings of Society for Information Technology and Teacher Education international conference 2004* (pp. 951–954). Chesapeake: AACE.

Dai, C., Cheng, J., & Hsu, Y. (2005). The new meaning of S-P Chart. In P. Kommers & G. Richards (Eds.), *Proceedings of world conference on educational multimedia, hypermedia and telecommunications 2005* (pp. 3074–3079). Chesapeake: AACE.

de la Torre, J. (2012). Application of the DINA model framework to enhance assessment and learning. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. Dordrecht: Springer.

Dinero, T. E., & Blixt, S. L. (1988). Information about tests from Sato's S-P Chart. *College Teaching Journal, 36*(3), 123–128.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer (Ed.), *Measurement and prediction* (The American Soldier, Vol. IV). New York: Wiley.

Harnisch, D. L. (1981). *Analysis of item response patterns: Consistency indices and their application to criterion referenced tests* (ERIC Document Reproduction Service No. ED 209335).

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20*(2), 191–206.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*(3), 133–46.

Harnisch, D. L., & Romy, N. (1985). *SPP: Student problem package on the IBM-PC. User's guide, version 1.0*. Champaign: Office of Educational Testing, Research, and Service/University of Illinois at Urbana-Champaign.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Ho, C. M., Leung, A. W. C., Mok, M. M. C., & Cheung, P. (2012). Informing learning and teaching using feedback from assessment data: Hong Kong Teachers' attitudes towards Rasch measurement. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessments in the Asia-Pacific*. Dordrecht: Springer.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*(2), 109–18.

Mok, M. M. C. (2010). *Self-directed learning oriented assessment: Assessment that informs learning & empowers the learner*. Hong Kong: Pace Publications Ltd.

Mok, M. M. C., Ting, Y. C., Ho, P. H. S., Wong, M. Y. W., Tse, L. C. N., Xu, J. K., & Yao, J.-J. (2011). (In Chinese: 莫慕貞、丁彥銓、何昊璇、黃英華、謝棹南、徐坤、姚靜靜 (2011) 。優化學習導向評估之SP Xpress 2.2。Hong Kong: Pace Publications Ltd.)

Ngan M. Y. (2011). (In Chinese: 顏明仁(2011) 。第八章:試題分析(二)。促進學生學習的當代教育評估理論與實踐。香港:培生教育出版南亞洲有限公司。)

Sato, T. (1980). The S-P chart and caution index. In *NEC Educational Information Bulletin*, 80–1. C&C Systems Research Laboratories, Nippon Eletric Co. Ltd, Takatsu-Ku Kawasaki City, Kanagawa Prefecture 213, Japan.

Sato, T. (1984). *The state of art on S-P analysis activities in Japan*. C & C System Research Labs, Nippon Electric Co. Ltd, Takatsu-Ku Kawasaki City, Kanagawa Prefecture 213, Japan.

Sato, T. (1985). *Introduction to student-problem curve theory analysis and evaluation*. Tokyo: Meiji Tosho.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Tam, H. P., Wu, M., Lau, D. C. H., & Mok, M. M. C. (2012). Using user-defined fit statistic to analyze two tier items in mathematics. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. Dordrecht: Springer.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*(1), 95–110.

Tzuriel, D. (2012). Dynamic assessment of learning potential. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. Dordrecht: Springer.

Wu, M. (2012). Using item response theory as a tool in educational measurement. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. Dordrecht: Springer.

Yu M. N. (2002). (In Chinese: 余民寧(2002) 。第八章:學生問題表分析。《教育測驗與評量:成就測驗與教學評量》。台北市:心理出版社股份有限公司。頁325–369。)

Chapter 12
Using User-Defined Fit Statistic to Analyze Two-Tier Items in Mathematics

Hak Ping Tam, Margaret Wu, Doris Ching Heung Lau, and Magdalena Mo Ching Mok

12.1 Introduction

The two-tier item is a relatively new diagnostic item format for classroom assessment and is gradually gaining popularity in certain areas of educational research. For the past two decades, it has been used to assess at a deeper level students' understanding of the concepts being covered in classes, especially in the area of science education (e.g., Treagust 1988; Treagust and Smith 1989; Tan and Treagust 1999). Its popularity in recent years may be partly illustrated with the following piece of information. In 2007, a whole issue of the *International Journal of Science Education* was devoted to reporting the research design and results of a study entitled National Science Concept Learning Study (NSCLS). This study was conducted in Taiwan in 2003 and involved more than 30,000 students from primary to senior high school

H.P. Tam (✉)
Graduate Institute of Science Education, National Taiwan Normal University,
Taipei City, Taiwan
e-mail: t45003@ntnu.edu.tw

M. Wu
Work-based Education Research Centre, Victoria University, Melbourne, Australia
e-mail: wu@edmeasurement.com.au

D.C.H. Lau
Formerly Centre for Assessment Research and Development,
The Hong Kong Institute of Education,Tai Po, Hong Kong

The University of Hong Kong, Hong Kong
e-mail: chlaudoris@gmail.com

M. Mo Ching Mok
Department of Psychological Studies, and Assessment Research Centre,
The Hong Kong Institute of Education, 10 Lo PingRoad, Tai Po, N.T., Hong Kong
e-mail: mmcmok@ied.edu.hk

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_12, © Springer Science+Business Media Dordrecht 2013

(Guo 2007; Tam and Li 2007). Its main purpose was to assess students' misconceptions of important science concepts from primary to senior high school. What is worth noticing is that all the items adopted in this study were framed in a two-tier format.

A two-tier item can be viewed as a special kind of testlet in that it has a common item stem followed by two subitems, with one of them requiring the respondents to carry out part of the task while the subsequent subitem requiring them to finish the remaining part of the task. In science education, a typical two-tier item is made up of an item stem followed by two portions. Usually, the purpose of the first portion is to assess whether students could identify some factual aspects with respect to a phenomenon stated in the item stem, while the second portion examines if they can supply the correct reason associated with why the phenomenon occurs. Since some students may not be able to identify the correct option associated with the first portion, what they chose as the accompanying reason in the second portion could then reveal valuable information about their knowledge status about the phenomenon being tested. More specifically, the combination of options being chosen across the two tiers has the potential of revealing the misconception being held by students about why some phenomenon happens or does not happen. As a result, this item format has been used as one of the ways to illuminate the kind of misconceptions as well as how widespread they have been among the students taking the test.

Although this format has not quite found its way into research conducted in the area of mathematics education, there are, nevertheless, situations where some mathematics items can be essentially treated as two-tier items. For example, one common way of assessing students' abilities in solving word problems is to ask them to formulate an equation that corresponds to the conditions given in the items. In some tests, partial credits will already be assigned to examinees who have been successful in expressing the correct equation. Afterwards, the examinees are required to solve the equations they have formulated and then provide their final answers. Again in some tests, partial credits may be assigned to those who can provide the correct final answer. Thus accordingly, one can view a word problem as the item stem and the requirement to set up the corresponding equation as the first tier while the compilation of the final answer as the second tier. As a matter of actual practices, this approach has been frequently adopted by mathematics teachers especially in the elementary grades.

Unfortunately, the methodology regarding how the two-tier items should be analyzed is still fairly underdeveloped in the area of science education. For example, many data analysis, as can be currently identified in the literature, is limited to reporting tables of percentages of options being chosen by the examinees across the two tiers for each individual item. This approach is descriptive in nature and is dependent on the sample of students taking the tests. The quality of a two-tier item, moreover, is usually assessed by appealing to the judgment of subject matter experts based on their professional experiences. However, there are many a time when professional judgment cannot be easily made, such as when the two-tier item format appears brand new to the experts. In real practices, it is quite often the other way around with the subject matter experts requesting the data analysts to provide them with supportive statistical information, thereby assisting them in their judgment making regarding whether the two-tier items are in good shape.

One possibility is to use the techniques that have been developed for analyzing testlets or item bundles as reported in the literature. One such alternative is the testlet response theory developed by Wainer et al. (2007). This theory is accompanied by a software program entitled Scoright, which is freely available by way of Educational Testing Service (ETS), thereby making it more attractive to applied researchers. Yet, the technicality behind the testlet response theory is quite involved for most school teachers or even applied researchers to comprehend. In addition, the current version of Scoright is not as user friendly as one would desire. Furthermore, since the theory is based on the Bayesian approach in its estimation of parameters, Scoright can be quite slow in terms of program execution. Though the program allows starting values to be provided by the users so as to speed up the estimation process, many school teachers or applied researchers may find it difficult and need help in deciding on a good set of starting values. There can also be times when the program cannot converge at all in its execution. Thus, it seems that a friendlier approach is much desirable for the common practitioners so that they can handle the analysis of two-tier items in an easy-to-understand manner. Since such information is currently unavailable at large, thus there appears to be a need in developing useful technique for analysis that takes into consideration the relationship between the two tiers within the same item.

12.2 Purpose of Study

A three-step procedure has been proposed in Tam and Wu (2009) as an all-purpose approach to analyze two-tier items. Such practical information as the scoring of the item, the dependence between the two tiers, as well as the functioning of the items can then be provided to the item writers for item evaluation and revision. Since the third step is similar to the item analysis procedure that is commonly seen in a Rasch analysis setting, this chapter aims at illustrating the first two steps that are particularly important for the two-tier item format. More specifically, this chapter will first occupy itself with assessing if there are dependencies between the two tiers for each item on the test as one would expect from the nature of this particular format of item. Afterwards, this chapter will turn its attention to investigate how two-tier items should be scored in the first place. These two steps of data analysis are especially relevant to the data set from a mathematics test with a two-tier structure which will be used to demonstrate the procedure discussed herein. Both the method with its rationale and the data employed for demonstration will be described in more detail in the next two sections. They will then be followed by the results section. Finally, the specific issue about whether all two-tier items should be scored the same way together with other issues of more general interest will be dealt with in the discussion section.

12.3 Method

As a start, one useful yet succinct way of organizing the overall performances of students with respect to a typical two-tier item is to construct a two-by-two cross-tabulation table for the distribution of the students' proportions of right or wrong across the two tiers as illustrated in Table 12.1 below. Among the students who sit for the test, let x be the proportion of those who got both the factual and the reason portions correctly. Similarly, let y be the proportion that got both portions wrongly, z be the proportion that got the factual portion correctly but the reason portion wrongly, and w be the proportion that got the factual portion wrongly but the reason portion correctly.

The original data analysis procedure proposed by Tam and Wu (2009) was comprised of three steps, each tapping into a different kind of information from the two-tier items that appear on the test. The rationale behind this procedure is as follows. Since both tiers, by nature of the item structure, access the same piece of information in the item stem, it is regarded as being safe to assume that students' performances with respect to the two tiers will be related to each other. Hence, the purpose of their first step is to discern systematically if there exists a dependency between the two portions for each two-tier item. If it so happened that the dependency between the two tiers is found to be low for some items, reasonable doubt could then be raised concerning whether these items have functioned according to the intent of the item writers. These items should either be deleted or subjected to revision by referring them back to the item writers. After the relationships between the two tiers have been established, one can then consider how the items should be scored. For example, should the data analyst score the items by using partial credits or should the item be considered correct only when both portions are answered correctly? If the items were inappropriately scored, then any subsequent effort in item analysis and interpretation of results would most likely be led astray. Hence, the second step of the proposed procedure will concentrate on selecting an appropriate item response model that can take into account the dependencies between the two tiers. It is deemed essential to notice that even for those items with justifiable relationship across the two tiers, they have to be properly scored with an appropriate item response model before further item analysis on the items be performed. The main concern of the third step is the provision of item information to subject matter experts that may be useful for revising and rewriting the items. These steps are explained in more detail as follows.

Table 12.1 Proportions of students' performance across the two portions of a two-tier item

Second tier First tier	Right	Wrong	Row total
Right	x	z	$x+z$
Wrong	w	y	$w+y$
Column total	$x+w$	$z+y$	1

Tam and Wu (2009) pointed out that if dependencies exist across the two portions within the two-tier structure, the local independence assumption behind the item response modeling approach will in principle be violated should an item response model be attempted on the data (Embretson and Reise 2000). In order to detect this violation, the user-defined fit statistic as discussed in Adams and Wu (2011) can be applied. The gist of this test statistic will be explained here briefly. Let us first consider an examination that is made up of the usual multiple-choice test items. The user-defined fit statistic allows the data analyst to define several items as a group. The number of items correct within the group is then counted for each participant, which can be regarded as the sum score obtained by each participant. If there is no violation of the local independence assumption within the group of multiple-choice items in the first place, then the sum score should also fit the item response model. However, if there is dependency among the items in a group, then the sum score will tend to be too high or too low than expected, owing to the relationship among the items in the group. Thus, when all the items satisfy the local independence assumption except for those items defined in the group, this group of items can be picked up by the user-defined fit statistic as not fitting the item response model applied. The user-defined fit statistic is implemented in the ConQuest and can be used to compute any groupings of items in a test (Wu et al. 1998). The sum score can be tested against the sum score of yet another group of items also defined by the data analyst. This idea can then be extended to the situation when the multiple-choice test has an extra two-tier item added. For this particular two-tier item, if a respondent scores high on one tier, then it is likely that the same respondent will also score high on the other tier. Thus, when the two portions of the two-tier item are treated as a group, it can be picked up by the user-defined fit statistic. In this chapter, a data set will be used to demonstrate the procedure discussed herewith. This data came from an examination that consisted of both regular items and a number of two-tier items. In our first analysis, an item response model was fit to the two-tier items as if they were all made up of independent items. In other words, the relationships between the two tiers were ignored in this round of analysis. Fit statistics were then computed for the two tiers in each item pair. The magnitude of the fit statistic provides a measure of model violation, thereby revealing how closely the two tiers are related within each item pair.

After the relationship between the two tiers has been established, the second step focuses on selecting an appropriate item response model that can account for the dependencies between the two tiers of the items. For example, each two-tier item can be modeled as one item by scoring it by ways of assigning partial credits. In this step, the data analyst should consider a number of item response models that might reasonably be used to score the two-tier items. These models will then be applied to the data. Model comparisons are made by means of the model fit statistics as well as the test reliability information from each model. The best fitting model could then be used for calibrating and further purposes.

The third step involves the extraction of information at the level of response categories so as to assist item writers in assessing how each item pair functioned. In addition to the frequencies or proportions of respondents in the various response categories, the average ability is, for example, also useful information, as well as the corresponding item characteristic curves by category. As explained earlier, the third step is more familiar to applied researchers, so this chapter will focus on delineating the first two steps. The data set that will be utilized to illustrate the suggested procedure will be described in the next section.

12.4 Data

The aforementioned methodology has been applied to a set of data collected in 2010 by the Assessment Research Centre of The Hong Kong Institute of Education. The test instrument was made up of ten mathematics items cater for students at the fifth-grade level in Hong Kong. The contents being tested included eight items on fractions, one item on rearranging a given set of digits to obtain the smallest number, and one item on finding the greatest common divisor out of a given set of numbers. Of the eight items related to fractions, two of them were purely computational type of items while the other six were word problems. Furthermore, three of the word problems required the respondents to list out their steps before reporting their answers. These three items were regarded as two-tier items in the present study. The other items only required the respondents to write down their answers. A total of 860 fifth-grade students participated in the test. This data set was actually part of a larger study, the purpose of which did not affect in whatever way the methodology proposed hereby in this chapter. All the analysis was performed by using the specialized software program ConQuest (Wu et al. 1998).

12.5 Results

In order to explore if the three two-tier items really did violate the local independence assumption, the step listing portion and the corresponding answer reporting portion were treated as a group for each item. They were labeled as items 7.1, 7.2, 8.1, 8.2, 9.1, and 9.2, respectively. User-defined fit statistics were applied to these three two-tier items. For comparison purpose, individual portion of the three two-tier items were treated as independent items and were randomly paired with the rest of the items on the test form. User-defined fit statistics were also applied to each of these random pairs of items. The results were reported in Table 12.2 below. The user-defined fit statistic in Table 12.2 can be regarded as approximately a *z*-test. When its value is within the range of −2 and 2, the item pair could be regarded as fitting the item response model that is based on the assumption of local independence. When the user-defined fit statistic is greater than 2, the item pair is regarded as having a dependency beyond what the item response model assumes. When it is less than −2, the item pair is regarded as testing different constructs.

Table 12.2 User-defined fit statistics for the two-tier items that were correctly paired and also for items that were randomly paired

Portions of two-tier items correctly paired		Items paired randomly	
Item pair	Fit statistic (weighted)	Item pair	Fit statistic (weighted)
7.1, 7.2	12.162	1, 7.1	–0.672
8.1, 8.2	13.203	2, 7.2	–0.408
9.1, 9.2	13.693	3, 8.1	–1.592
		4, 8.2	–1.830
		5, 9.1	–0.606
		6, 9.2	–2.158
		1, 2	3.161
		3, 4	1.417
		5, 6	1.909
		8.1, 10	–0.087
		8.2, 10	–0.111

It can be seen that when the portions from the two-tier items were correctly paired, the three two-tier items (i.e., items 7–9) all had a user-defined fit statistic much greater than 2, thereby indicating that dependencies existed between the two portions within each two-tier item. In contrast, for items that were randomly paired, as reported in the last two right-hand columns in Table 12.2, their user-defined fit statistics were mostly within the range from −2 to 2. This indicated that the item pairs did not have a dependency over and beyond what the item response model assumed. This was especially the case for both items 7 and 8. When the portions from these items were paired up randomly with other items in the same test, their user-defined fit statistics tended to be much smaller in magnitude. Another interesting observation is that the first portion of each two-tier item tended to have a smaller, at least in a relative sense, user-defined fit statistic when they were randomly paired with the non-two-tier items. This may be attributable to the fact that the computational step of a mathematics item will under most circumstances be unrelated to the answer of another mathematics item. It must be emphasized that the pairing of one portion from a two-tier item with a non-two-tier item is not limited to those listed in Table 12.2. The pairings listed over there are for demonstration purposes. Should any doubt ever arise, another round of random pairings can be pursued and the user-defined fit statistics performed again. However, one should be careful not to capitalize on chances by running too many tests.

It was further noticed that not only did the three two-tier items demonstrate a similar pattern in terms of high user-defined fit statistics, they also shared a similar distribution of respondents' performances in proportions with respect to the two-tier structure. As a typical example, Table 12.3 reported the distribution of participants' proportions across various combinations of right and wrong with respect to the two-tier structure in item 7.

As can be seen from the table, the majority of the respondents (almost 95%) answered either correctly or incorrectly to both tiers of item 7 at the same time. There were a few respondents who had written down the computational portion

Table 12.3 Distribution of respondents' performances in proportions with respect to item 7

First tier \ Second tier	Right (%)	Wrong (%)
Right	66.16	5.47
Wrong	0	28.37

correctly but yet provided the wrong answer. However, there were no respondents who could obtain the correct answer for the second tier yet missed out on the first tier. These findings make empirical sense since the first tier of this item required the respondents to write down the expression that was necessary for computing the answer. Logically speaking, one must first get the computational portion correct before one can obtain the right answer to the item. It is highly uncommon that a wrongly formulated expression would still render the right answer in real life situation. Apparently, the dependence between the two tiers is fairly strong as demonstrated by the distribution of respondents' performances with respect to this item. After taking all the information together, it can be regarded that the two-tier structure of this particular mathematics exam has been substantiated by results from the user-defined fit statistic.

The second step of the suggested procedure in Tam and Wu (2009) involved the selection of an appropriate item response model. Four separate models were fitted to the group of two-tier items in the mathematics exam. The first model attempted was a dichotomous model in which all the portions from the two-tier items were treated as if they were entirely independent items. Accordingly, all the portions were scored either as right or wrong. This model served as the baseline model in this study and was adopted purely for comparison purpose. Since it is deemed improbable by most subject matter experts that a respondent could get the second tier correct and yet missed the first tier, the second model attempted was a partial credit model in which a score of 2 was assigned to the case when both tiers were answered correctly, a score of 1 when only the first tier was correct, and a score of zero for the other combinations. This model was denoted as the 2100 model to facilitate subsequent discussion. As for the third model, another partial credit model similar to the previous one was fitted to the data with slight modification. This time, however, a score of 1 was also assigned to those respondents who obtained the correct answer to the second tier. This model was short-handed as the 2110 model below. Finally, another dichotomous model was attempted as the fourth model in which a respondent was assigned a score of 1 if and only if he/she had answered both tiers correctly. All the other combinations were scored as zero. This model was adopted upon recommendation from some subject matter experts who maintained that both the step and the answer must be correct before mastery of the content being tested could be justifiably assumed. For ease of discussion, this model was short-handed as the 1000 model. It should be noticed that since there were very few respondents who would only get the second tier correctly, hence a partial credit model with the scoring scheme of assigning a score of 3 to both tiers correct, a score of 2 to the first tier correct, a 1 to the second tier correct, and a zero to both tiers incorrect would

Table 12.4 The deviances and the reliabilities for the four-item response models being attempted

Treatment of second tier item	Deviance	Reliability
Individual items	12694.13	0.793
Scored as 2100	10792.24	0.720
Scored as 2110	10798.12	0.720
Scored as 1000	10079.44	0.710

Table 12.5 The percentages and average abilities for respondents manifesting different response patterns with respect to item 7

Response category	Percentages (%)	Average ability (logits)
Both tiers incorrect	28.37	–0.767
Second tier correct but not first	0	N/A
First tier correct but not second	5.47	–0.339
Both tiers correct	66.16	0.841

create calibration problem. For similar reason, more refined scoring schemes were not practically pursued in the present study.

The results are shown in Table 12.4 above. It is found that the fourth model had the lowest deviance statistic among the four models being processed. In addition, the reliability of the fourth model was found to be the smallest even though its value was quite comparable to the other two partial credit models. Meanwhile, the drop in reliability of the 1000 model from the first model in which the two-tier items were treated as independent items was quite prominent. There are two possible explanations for the drop in reliability when we use the 1000 model. First, when items are treated as independent items when they are actually dependent, the reliability will be artificially inflated, as in the 2100 model. Second, the three two-tier items could be more discriminating items, so a maximum score of 2 instead of 1will give more weight to these items, leading to an increase in reliability. In any case, the reliabilities among the 2100, 2110, and 1000 models are very close to each other. As a result, the 1000 model was being adopted as the model to score the two-tier items in this study.

In case further evidences were desired to justify the adoption of the 1000 scoring scheme, then more analysis should be performed at the individual item level. Reported in Table 12.5 above were the proportions of respondents who manifested different response patterns in item 7 together with their average abilities in terms of logits. As can be seen from the table, the average ability for those respondents who were incorrect in both tiers was –0.767, while that for the respondents who were incorrect in the second tier but right in the first tier was –0.339, and 0.841 for those who were correct in both tiers. It is noticed that the average abilities for those who answered both tiers correctly amounted to a positive value that was much larger than the negative average abilities for other two combinations of response categories. These findings seem to reflect that the three groups of respondents were of different abilities, with those respondents answering both tiers correctly attaining the highest average abilities while the other two groups of respondents were of closer average abilities. Thus, this further piece of information warranted strong support regarding the adoption of the 1000 scoring scheme for the two-tier items that appeared on the test, at least with respect to the models attempted.

12.6 Conclusion and Discussion

This study had demonstrated a new and rather comprehensive approach from Tam and Wu (2009) to analyze two-tier items beyond the report of mere proportions of respondents with respect to the various combinations of response categories across the two tiers as a means of data analysis. While such proportions are simple and straightforward to compute, the kind of information that can be gleaned is fairly limited. With mere proportions, incorrect responses to a two-tier item may of course be attributed to some inappropriate mastery on the part of the respondents towards the content being tested. However, it could also be attributed to some underlying deficiency in terms of the conceptualization, design, or even wordings of two-tier items being written. There is not enough information to distinguish between these and other possibilities because they are convoluted with one another. In comparison, the approach suggested herein will be much easier to comprehend by most applied researchers. Under the suggested approach, the results from the first stage of our procedure can reflect whether the two-tier item structures can be substantiated from the empirical data. With respect to the mathematics exam being analyzed, the result from the first stage will reflect whether the computational steps and their respective answers can really substantiate a two-tier structure. If there is no foundation for such claim, careful revision of the two-tier item is advised. On the other hand, the second stage aims at finding a basis concerning how the two-tier items can be scored more appropriately. The decision attained at this stage can subsequently be used to calibrate the items for various parameter estimates as well as generate other useful information.

Furthermore, it was found in this study that the 1000 model had the lowest deviance than the two partial credit models as well as the independent items model being considered. This finding forms the basis for scoring our items in accordance to the 1000 scoring scheme. Thus, consideration of an appropriate scoring procedure should constitute an important step in the analysis of two-tier items. According to our experiences, it appears that the 1000 scoring scheme always performs relatively well with two-tier items. Hence, it is suggested to always include this scoring scheme as one of the options while carrying out the second step of the suggested procedure for two-tier items.

Finally, in order to obtain the best result from the two-tier item format, it is suggested that potential item writers should try every effort to focus on improving the qualities of the items first. Pilot testing on the items is highly recommended. The procedure demonstrated in this study will be quite useful in throwing some light on the quality of the items especially during the pilot testing stage. Rather than jumping to early conclusion with regards to the abilities of the respondents, it is only after careful revision of all items with questionable quality before one should proceed to use the two-tier items for actual assessment purpose.

References

Adams, R. J., & Wu, M. L. (2011). The construction and implementation of user-defined fit tests for use with marginal maximum likelihood estimation and generalized item response models. In N. J. S. Brown, B. Duckor, K. Draney, & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 2). Maple Grove: JAM Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates.

Tam, H. P., & Li, L. A. (2007). Sampling and data collection procedures for the National Science Concept Learning Study. *International Journal of Science Education, 29*(4), 405–420.

Tam, H. P., & Wu, M. (2009). *Analyzing two-tier items with user-defined fit statistics*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, USA.

Tan, K. D., & Treagust, D. (1999). Evaluating students' understanding of chemical bonding. *School Science Review, 81*(294), 75–83.

Treagust, D. (1988). Development and use of diagnostic test to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159–169.

Treagust, D. F., & Smith, C. L. (1989). Secondary students' understanding of gravity and the motion of planets. *School Science and Mathematics, 89*(5), 380–391.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest – Generalised item response modeling software*. Melbourne: Australian Council for Educational Research.

Chapter 13
Dynamic Assessment of Learning Potential

David Tzuriel

13.1 Introduction

The dynamic assessment (DA) of learning potential approach presented in this chapter is based mainly on Vygotsky's (1978) sociocultural theory, specifically the *zone of proximal development* concept, and Feuerstein's *mediated learning experience (MLE)* theory (Feuerstein et al. 1979) and Tzuriel's DA approach developed in the last three decades (Haywood and Tzuriel 1992; Tzuriel 1989, 1997, 2000, 2001, 2002; Tzuriel and Klein 1985). DA refers to an assessment, by an active teaching process, of a child's perception, learning, thinking, and problem solving. The process is aimed at modifying an individual's cognitive functioning and observing subsequent changes in learning and problem-solving patterns within the testing situation (Tzuriel 2001). The term *static* (or *standardized*) test refers to a test where the examiner presents items to the child and records his/her response without any attempt to intervene in order to change, guide, or improve the child's performance.

DA has been motivated by the inadequacy of conventional static tests to provide accurate information about the individual's learning ability, specific deficient functions, change processes, and mediation strategies that are responsible for cognitive modifiability. The need to develop DA tests has emerged because of criticism on static standardized tests and the difference in type of questions asked by DA as compared with standardized testing.

In the following sections of this chapter, I will discuss (a) the main criticism on standardized static tests, (b) the main goals of DA, (c) the major shifts of DA from standardized testing, (d) the major strategies of mediation in DA, (e) the use of DA in educational research, (f) the criticism of DA, and (g) why DA is not applied on a larger scale.

D. Tzuriel (✉)
School of Education, Bar Ilan University, Ramat Gan, Israel
e-mail: David.Tzuriel@biu.ac.il

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_13, © Springer Science+Business Media Dordrecht 2013

13.2 Main Criticism on Standardized Static Tests

The major criticism against standardized testing can be summarized in the following main points:

(a) A frequent argument raised in the literature is that standardized static tests are biased toward minority groups and children with special needs and do not reflect their true ability. Children who come from low socioeconomic status (SES) families do not have adequate learning opportunities or efficient mediation from their parents and therefore fail in academic performance and/or in standardized tests. Their failure, however, does not reflect lack of intellectual abilities but rather lack of learning strategies, deficient cognitive functions (e.g., impulsivity), learning habits, self-efficacy in academic domains, and task-intrinsic motivation (Feuerstein et al. 1979).

(b) Another argument is that standardized tests are characterized many times by selective administration procedures and selective interpretation of results among high-risk children. For example, more lenient procedures (e.g., repeating instruction, showing more sympathy, allowing extra time, and giving hints) are used with children coming from high SES families than with children coming from low SES families. Although the test procedures are standardized, some examiners might use an "under-the-table" strategy of giving little cues for items not answered. This differential response might be on a subconscious or even a conscious level. In DA, on the other hand, mediation is "on-the-table" as the child is given "full-blooming" guidance and help. Another aspect of differential testing is selective interpretation of test results. Some examiners might judge a child's performance, especially a child with special needs, more strictly than a typically developing child and reach more severe conclusions than what actually is the child's level.

(c) A major argument against standardized tests is that motivational, emotional, and personality factors are not well taken. Research literature and teaching experience show that the motivational, emotional, and personality factors are no less important than the "pure" cognitive factors (Haywood 1968, 1971; Haywood and Lidz 2007; Tzuriel et al. 1988). Unfortunately, these factors are not given the proper attention in static tests or even totally neglected.

(d) The strongest argument against static tests is lack of information on learning and metacognitive processes. Those processes are of most importance in explaining the child's learning in the classroom and academic achievements. Teachers are interested in getting information on learning processes no less than on the relative standing of the child as compared with peers. As opposed to standardized tests, DA provides educators with data needed to suggest specific strategies for effective instruction and intervention. The different orientations of DA approach from static test approach derive from the major distinction in the type of questions asked by each approach. While in static testing the focus is on question of *what* is the level of the child's ability relative to same-age peers or what is the child's profile on certain subscales, in DA the questions are focused on *how* the child *processes* the information, *what* are the specific cognitive functions responsible for the child's performance, *how* can we change and improve thinking and learning, and how cognitive, motivational, and emotional changes during testing can be used later to enhance the child's functioning in academic and nonacademic settings.

(e) Very frequently, static tests provide inadequate recommendations on remediation processes, specific interventions strategies, and prescriptive teaching. Many times, there is a "communication gap" between teachers and psychologists regarding translation of test findings into day-to-day teaching activities. It is common to find that teachers do not understand the terminology of static tests; the psychometric information is useless and barely translated to treatment strategies. Very frequently, psychologists do not have much experience with learning processes, and the static test data are not easily translated to specific recommendations.

13.3 Goals of DA

In order to understand deeply how DA is used and how it can help children with learning difficulties, we must understand the goals of DA. These goals may be summarized in the following:

(a) The first goal is to examine the capacity of the child to grasp the principle underlying an initial problem presented to the child and solve it correctly. This goal is very similar to the static test's goal, evaluating the manifested level of performance, or in Vygotsky's terms, the *actual level* of the *zone of proximal development*.

(b) The second goal is to assess the specific deficient cognitive functions as well as the adequate cognitive functions that are responsible for the child's failures and successes, respectively. Cognitive functions were defined as compounds of native abilities, learning habits, attitudes toward learning, motivational orientations, and cognitive strategies (Feuerstein et al. 1979). Adopting an information processing approach, Feuerstein suggested a list of deficient cognitive functions on the *input*, *elaboration*, and *output* phases of the mental act. For example, in the input phase one can identify difficulties in systematic exploratory behavior, simultaneous consideration of two or more sources of information, and spatial orientation. Deficient cognitive functions in the elaboration phase might be expressed by difficulties in planning behavior, comparative behavior, working memory, and episodic grasp of reality. Deficient cognitive functions in the output phase might be expressed by egocentric mode of communication, trial-and-error behavior, and projecting virtual relations. The deficient cognitive functions are considered as key elements for understanding children's performance. The modifiability of cognitive functions and operations (e.g., analogy, seriation) during DA is considered as an indicator for future changes, provided some treatment is given to modify them.

(c) The third goal of DA is to examine the nature and amount of investment required in order to teach the child a given principle or modify a deficient cognitive function. The examiner evaluates *how much* as well as *what types* of mediation are required in order to improve the child's cognitive functioning. This information is crucial in order to recommend later the type of mediation strategies needed as well as the required intensity.

(d) The fourth goal is to examine the extent to which the newly acquired principle is successfully applied in solving problems that become progressively more complex than the initial task. This goal is related to the level of internalization of learning and the amount of transfer the child's show in problem solving.

(e) The fifth goal is to examine the differential preference of the child for one or another modality of presentation of the problem (i.e., pictorial, linguistic, numerical). Understanding of the modality preference may help teachers in the future in designing intervention strategies and techniques.

(f) The sixth goal is to examine the differential effects of different training strategies given to the child to improve his/her functioning. It is important to understand what type of mediation is more effective especially in relation to the type of task that is given. The effects are measured by using the criteria of task level of novelty, level of complexity, language of presentation, and types of operation (i.e., analogy, syllogism, spatial orientation).

13.4 Major Shifts of DA from Standardized Testing

DA can be characterized by four major shifts from static standardized testing:

(a) *Goals of Testing.* The main goal in DA (see above goals of DA) is to assess learning potential and changes in task performance, cognitive functions, and nonintellective factors related to cognitive functioning. These changes are taken as indications for future changes, provided a cognitive intervention will be applied later to actualize the learning potential. In standardized testing, on the other hand, the main goal is to document the existing cognitive repertoire of the individual without any attempt to assess changes or learning processes.

(b) *Change in Nature of the Tasks.* Standardized tests are characterized by an emphasis on psychometric properties of the tasks, graduation of the difficulty levels of items, representation of children's capacities or knowledge, and administration procedures (e.g., test administration is terminated after several failures). Items are generally selected for the test if they coincide with psychometric properties (i.e., normal distribution, interitem reliability). In DA, on the other hand, the tasks are constructed on the basis of their "teaching potential"—the possibility of teaching important cognitive strategies, enhancing cognitive functions, and measuring cognitive changes. The items in DA are also graduated in terms of difficulty level, but the focus is on the teaching of cognitive strategies and operations so that learning of one task prepares the child to perform a more advanced task.

(c) *Change in Test Situation.* Since the objective of static tests is to compare an individual to his or her same-age peers, by definition, the test conditions require standardized stringent conditions for all examinees. Consequently, there is no room for teaching or interactive approach; examiners ask questions and examinees answer. Any guidance or help is perceived a transgression of the standardized conditions. Since the objective in DA is to change the individual's functioning within the test context, an active teaching approach is applied. Thus, there is a major shift in the role of the examiner from passive recording of the child's answers to an active mediation of cognitive strategies, rules, operations, and contents. In other words, while in standardized testing the examiner's roles are limited to administration of test items and later to scoring and interpretations, in DA the examiner intervenes to change the examinee's functioning and interprets future possible changes in view of current changes during assessment.

The DA interactive process is characterized by regulating the child's behavior through inhibition of impulsivity, sequencing and organizing the task dimensions, improving deficient cognitive functions, enriching the child's cognitive operations (e.g., comparative behavior, analogies, seriation) and task-related contents (e.g., labeling of relationships such as "opposite," "up–down"), and creating reflective and metacognitive processes.

The shift in test conditions might be symbolized by the sign frequently seen on the door of standardized testing rooms: "Silence! Testing in Progress." Contrary to the semiexperimental conditions required in standardized testing, in DA parents and teachers are often invited to observe the process. The observation may help later in explaining and reporting to parents the test results and in preparing for future cognitive intervention.

(d) *Change of Focus: From End Products to Process Orientation.* In standardized testing, the focus is on the end product of the mental act: the final answer. In DA, in contrast, the focus is on cognitive processes that bring about changes in specific deficient cognitive functions (e.g., impulsivity) and in nonintellective factors (e.g., need for mastery, resistance to mediation) that affect functioning. In other words, the emphasis is on process components, such as the nature of cognitive behavior, the learning process and strategies, and the specific interventions required to change them. While in standardized testing the emphasis is on the typical level of the child's performance, in DA the emphasis is on unique and qualitative aspects of the child's cognitive behavior. The questions asked in DA are "how" and "why" rather than "what" and "how much."

(e) *Change in Interpretation of Results.* While in standardized testing interpretation of results is based mainly on quantitative aspects, in DA it is based mainly on qualitative aspects of the child's performance, on analysis of the deficient cognitive functions, and on the mediational efforts required to modify them. The child's peak performance (i.e., independent performance after teaching) is taken as indicative of the child's ability rather than an average of all responses. Sometimes, only one bright answer provides a crucial indication of the child's learning potential, an indication that paves the way for deeper exploration of the possible factors that block the child from performing as well in other tasks.

Table 13.1 Major differences between DA and standardized testing

Dimensions	Dynamic assessment	Standardized testing
Goals of testing	Assessment of change Assessment of mediation Assessment of deficient cognitive functions Assessment of nonintellective factors	Evaluation of static performance Comparison with peers Prediction of future success
Orientation	Processes of learning Metacognitive processes Understanding of mistakes	End products (static) Objective scores Profile of scores
Context of testing	Dynamic, open, and interactive Guidance, help, and feedback Feelings of competence Parents and teachers can observe	Standardized Structured Formal Parents and teachers are not allowed to observe
Interpretation of results	Subjective (mainly) Peak performance Cognitive modifiability Deficient cognitive functions Response to mediation	Objective (mainly) Average performance
Nature of tasks	Constructed for learning Graduated for teaching Guarantee for success	Based on psychometric properties Termination after failures

These four major shifts from standardized testing to DA are summarized in Table 13.1.

13.5 Major Strategies of Mediation in DA

(a) *Improvement of (Deficient) Cognitive Functions.* The examiner should know how to identify the cognitive functions required for solution of a problem in the test and the mediation needed to improve the deficient cognitive functions.

(b) *Preparing the Child for Complex Tasks by Establishing Prerequired Thinking Behaviors.* Establishing prerequired thinking behaviors is carried out often by using mediation for transcendence and for self-regulation. Adequate initial investment in preparing the child brings about reduction of mediation efforts in later more abstract and complex problems. It is common to find children who solve difficult advanced problems much easier than the initial easy problems. Mediation of rules and principles (transcendence) has a motivational aspect as the child becomes independent of the examiner's mediation and enhances the child's sense of self-control. Mediation for self-regulation is carried out by focusing on systematic sequencing processes especially in complex problems requiring an analytic approach. The examiner might ask the child to repeat the process of solution in order to crystallize the order of solution and to acquire feelings of mastery and efficiency.

(c) *Self-Regulation by Planning and Organization of the Solution.* One of the most frequent deficiencies among low-functioning children is impulsivity. Inhibition of impulsivity is done many times by decreasing the importance of time for performance. This is carried out by intentional delay of the child's response, longer exposure to the problem, systematic planning of the solution alternatives, verbalization of the problem, representation of the solution before pointing to the correct answer, and metacognitive analysis of the impulsive behavior. An efficient way of coping with impulsivity is by enriching the child's cognitive repertoire with thinking operations, comparative behavior, verbal tools, and hypothesis-testing techniques.

(d) *Enhancement of Reflective, Insightful, and Analytic Processes.* Enhancement of reflective, insightful, and analytic processes is carried out by focusing the child on the relation between his or her own thinking processes and the consequential cognitive performance. The focus is not on the end product but rather on the thinking process in the context of the required operations, type of task, and situation. Creation of insight is important for generalization and transfer of learning. It can be done by a dialogue with the child before solving the problem ("What should we look at before we will start to solve this problem?") or after the solution ("Why did you succeed in solving the problem that was so difficult for you to solve before?"). The most efficient way of enhancing reflective processes is by presenting the child with conflicts, incongruent information, intentional ambiguity, and absurd situations, which will bring about a need to close the cognitive gaps.

(e) *Teaching of Specific Contents that Are Related to the Task-Specific Context.* Teaching of specific contents (concepts, terms, relations) is not for the sake of language enrichment but for further use in problem-solving tasks. For example, the use of the terms up, down, vertical, horizontal, diagonal, similar, opposite, and different is necessary for performing the mental operation. The examiner can deviate for a short time from the task to teach and establish missing concepts and return later to the task to assess the performance efficiency and use of the newly acquired concepts.

(f) *Feedback on Success or Failure in the Learning Process.* The feedback given, which is one of the cornerstones in DA, is mutual—from the child and the examiner sides. It is especially important with low-performing children who are limited in their skills for giving feedback to themselves. This limitation is related to difficulties in self-correction and comparison of findings not only because of lack of knowledge and verbal tools of the children but also because of lack of orientation to make comparisons. Many tests are based on the assumption that trial-and-error behaviors will eventually bring the child to learn the correct answer. This assumption is wrong with regard to low-functioning children who are characterized by episodic grasp of reality. These children do not relate between their behavior and its consequences. A trial-and-error behavior blocks their learning rather than facilitates it. The importance of feedback in DA derives from the examiner's ability to focus the child on the relation between behavior and consequence. The feedback is given not only on wrong answers but also on correct or partially correct answers, in order to teach self-correction. The goal of the feedback is beyond teaching the child a specific response. The aim is to teach insight, lawfulness, and meaning in relation to cognitive and emotional–motivational aspects.

(g) *Development of Basic Communication Skills and Adequate Response Style.* The mediation here is aimed at changing the child's response style so that problem solution will find a proper and efficient external expression. The examiner teaches the child how to communicate efficiently by the use of clear and accepted terms and avoiding egocentric communication. The examiner also teaches the child how to communicate precisely, justify the answer using logical arguments, and use verbal "codes" of expression and abstract high-order concepts rather than body gestures and facial expressions. It should be emphasized that previous communication style is not taken away before establishing new response styles.

13.6 Use of DA in Educational Research

The use of DA in educational research was aimed at (a) establishing the DA measures as more useful and accurate than standardized tests, especially with children showing learning difficulties and other clinical groups (Carlson and Wiedle 1992; Guthke and Stein 1996; Guthke and Wingenfeld 1992; Haywood and Lidz 2007; Hessels 2000; Resing 1997; Sternberg and Grigorenko 2002; Tzuriel 2001; Wiedl 2003), (b) validating theoretical concepts that are at the basis of DA (e.g., *zone of proximal development, structural cognitive modifiability)*, (c) demonstrating the effectiveness of DA in predicting school achievements, and (d) evaluating cognitive education programs. In the following sections, I will focus on two aspects: use of DA with children demonstrating learning difficulties and revealing the effectiveness of cognitive education programs by DA measures. For other aspects readers are referred to the respective literature (Haywood and Lidz 2007; Lidz and Elliott 2000; Sternberg and Grigorenko 2002; Tzuriel 2000, 2001).

13.6.1 DA with Children Demonstrating Learning Difficulties

DA was extensively used in research with children coming from low SES, minority ethnic groups, and different cultural backgrounds (Hessels 2000; Sternberg et al. 2002; Tzuriel and Kaufman 1999), as well as with children with learning and intellectual disability (Hessels-Schlatter 2002; Tzuriel 2000, 2001). In general, previous research has shown that standardized intelligence scores underestimate the cognitive potential of children coming from low SES backgrounds, ethnic minority, and children with special needs, and that DA was proved to be more accurate in revealing their learning potential than static tests do (e.g., Guthke and Wingenfeld 1992; Hamers et al. 1991; Hessels 2000; Lidz and Elliott 2000; Resing 1997; Resing et al. 2009; Sternberg and Grigorenko 2002; Sternberg et al. 2002; Tzuriel 2000, 2001; Wiedl 2003).

DA results have been found to be more sensitive indicators of cognitive potential due to a variety of factors such as sociocultural deprivation, amount and quality of mediation provided at home, specific competencies for taking tests, interruptions in communication between examiner and examinee, test bias, and nonintellective factors such as self-confidence, need for mastery, and intrinsic motivation. By comparing static to DA measures, Guthke and Stein (1996) came to a conclusion that DA does not have a better predictive validity than static tests when used with typically developing students. However, in students with learning difficulties or atypical educational history, DA turned out to be a much better predictor of their future educational performance than static test scores. These findings support the conception of DA as an effective approach for revealing a "hidden" intellectual potential of special needs students. Sternberg et al. (2002) used DA with a group of rural Tanzanian school children ranging in grade levels from 2 to 5. The DA measures were largely based on fluid intellectual abilities such as syllogisms and sorting cards with different geometric figures. Children were assigned to experimental and control groups. The experimental group children received a short intervention phase for each test (well less than an hour per test) in which they were taught cognitive skills and strategies, whereas the control group children received no intervention. The findings showed significant pretest to posttest improvement across different tests in the experimental group as compared with the control group. Furthermore, posttest scores on the dynamic tests (administered in the experimental group only) were better predictors of reference ability and achievement measures than were pretest scores. One of the conclusions of this study, as expected, is that children growing up in difficult circumstances seem to have important intellectual abilities not measured by static tests.

In one of the earlier studies with young children, Tzuriel and Klein (1985) administered the *Children's Analogical Thinking Modifiability* (CATM) test to four groups of children: disadvantaged and advantaged kindergarten children, kindergarten children identified with special needs, and older intellectually disabled (ID) children with mental age equal to kindergarten level. The CATM is composed of three sets of analogies given in preteaching, teaching, and postteaching phases. The operation of analogy has been considered by many authors as a powerful tool for a wide range of cognitive processes and as a principal operation for problem-solving activities (Goswami 1991; Holyoak and Thagard 1997; Gentner and Markman 1997).

The CATM test is composed of 14 items for each phase of administration (preteaching, teaching, and postteaching) and 18 colored blocks that are used to present and solve the analogies. The CATM items, graduated in level of difficulty, require a relatively higher level of abstraction and various cognitive functions. Examples of items from the CATM test are portrayed in Fig. 13.1.

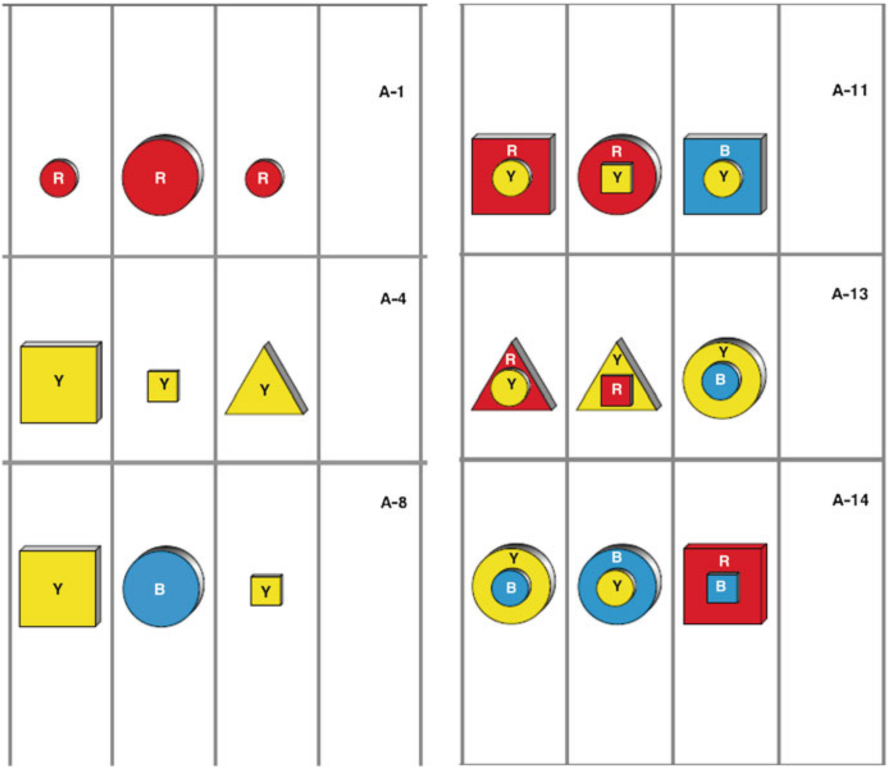


Fig. 13.1 Examples of items from the Children's Analogical Thinking Modifiability (CATM) test (R = Red, B = Blue, Y = Yellow)

In item 13, for example (see Fig. 13.1), the child has to compare the relations of colors in the first pair of the problem, find the rules of the relations, and apply them in the second pair. In the first pair, the relation of colors is opposite: *top*-yellow changes to *bottom*-yellow and *bottom*-red changes to *top*-red. If the rule of opposite is applied in the second pair, then the *top*-blue changes to *bottom*-blue and *bottom*-yellow changes to *top*-yellow. After finding the correct colors, the child can analyze the relations for the other two dimensions of shape and size (of both top and bottom components).

During the teaching phase, the child is mediated to (a) search for relevant dimensions required for the analogical solution, (b) understand transformational rules and analogical principles, (c) search systematically for correct blocks, and (d) improve efficiency of performance.

The CATM may be scored by two methods: "all-or-none" (e.g., a score of 1 is given to full answer) or "partial credit" (e.g., a score of 1 is given for each correct dimension of color, shape, and size). The findings showed that the highest gains from pre- to postteaching phases of the CATM test were found among disadvantaged and advantaged children as compared with children with needs for special education and ID children, who showed small gains. The ID group, however, showed significant improvement when a "partial credit" scoring method was applied. This last finding indicates that the ID group had difficulty in integration of all sources of information and therefore showed modifiability only according to the "partial credit" method. Higher levels of functioning were found for all groups on the CATM than on a static test, the Raven's Colored Progressive Matrices (RCPM Raven 1956). The differences were especially articulated when the analogical items of the RCPM were compared to the analogical problems of the CATM. For example, the advantaged and disadvantaged children scored 69% and 64% on the CATM, respectively, as compared to 39% and 44% on the RCPM, respectively.

In another study on children with special needs, Tzuriel and Caspi (1992) compared deaf children with hearing children on both DA and standardized measures. The kindergarten deaf children were matched to hearing children on variables of age, sex, and a developmental visual-motor test. Both groups were tested on the CATM and RCPM tests. The findings showed that on the CATM-postteaching phase, the hearing and deaf children scored 66% and 54% ("all-or-none" scoring method) and 86% and 81% ("partial credit" scoring method), respectively, as compared to 42% and 39% on the RCPM, respectively. These findings indicate that both groups have a higher level of learning potential than is indicated by static test scores. Comparison of pre- to postteaching tests revealed that the deaf children performed lower than the hearing children on the preteaching test but showed greater improvement after the teaching phase; no significant group differences were found in the postteaching test.

Previous studies with minority and culturally different children have shown that DA provides information different from conventional static tests. Guthke and Al-Zoubi (1987) compared a sample of 200 grade 1 children in Germany to a comparable Syrian sample on both a static measure—the Colored Progressive Matrices (CPM)—and a DA measure. The findings showed that the German children scored significantly higher than did the Syrian children. However, after a training phase, there was only a slight difference between the two groups. These results were interpreted as an indication that both ethnic groups have the same intellectual endowments. Similarly, Hessels and Hamers (1993) reported that although minority children scored significantly lower than Dutch children on learning potential tests, the differences were markedly smaller than with IQ tests. In South Africa, Skuy and Shmukler (1987) and Shochet (1992) used the *Learning Potential Assessment Device* (Feuerstein et al. 1979) and other psychometric tests with groups of children and students of Indian, Black, and "colored" origin. Skuy and Shmukler (1987) reported that although mediation was not generally effective in producing change on transfer measures, it was effective with a subgroup of colored high academic status students. The group that benefited most from mediation was the high academic status colored students. Shochet (1992) investigated the predictability of success in the first year of studies in the university using indexes of cognitive modifiability taken before admission on a disadvantaged student population. The findings showed significant prediction among "less modifiable" students but not among the "more modifiable" students (modifiability was measured by DA prior to start of the studies). It was surmised that they are less susceptible to being modified during the first year, either by direct exposure or by mediated learning experience (MLE).

A unique cross-cultural study was carried out by Tzuriel and Kaufman (1998) on a group of newly arrived Ethiopian children, in grade 1, who immigrated to Israel in the 1990s. They were compared with grade 1 Israeli-born children using static and DA tests. A central question that has been raised recently with new Ethiopian immigrants to Israel is how to assess their learning potential, especially in view of the inadequacy of standard testing procedures to reflect this population's cognitive functioning accurately. The question, however, transcends the specific context of the Ethiopian Jews. Theoretically, it is related to issues such as the influence of cultural changes on the individual's cognitive functioning, internalization of novel symbolic mental tools with transition from one culture to another, and resilience in coping with cultural incongruences. Pragmatically, this question applies to a variety of populations who, for sociohistorical reasons, live as subcultures within a broad culture and whose members might be penalized by inadequate diagnostic procedures.

It should be noted that the Ethiopian immigrants, upon arrival to Israel, had to overcome a gap of civilization and information of many years and had to adapt to the Israeli society. Coming from an illiterate society where their rich culture was transmitted orally, they had to go, upon arrival to Israel, through rapid change and adjust to differences in both material and symbolic tools. All children were administered the *Raven's Colored Progressive Matrices* (CPM Raven 1956), the CATM test, and the *Children's Inferential Thinking Modifiability* test (CITM Tzuriel 1989); the last two are DA measures. The CITM test, which is presented using verbal and pictorial modalities, taps several cognitive functions such as comparative behavior, systematic exploratory behavior, self-regulation of impulsivity, and inferential-hypothetical operations. An example of an item from the CITM is presented in Fig. 13.2.

The CITM test is composed of sets of problems for preteaching, teaching, postteaching, and transfer phases. After presentation of a set of 24 familiar pictures (e.g., clothes, animals, furniture) and naming them, the child is given two example problems and is instructed in the rules and procedures for solving them. Each problem consists of rows of figures, each row presenting partial information about the possible location of objects in houses with different colored roofs. The child is required to compare the information presented in the rows, infer the exact location of the objects, and place them in their right houses. The basic rule is that pictures on the left should be in houses with lines on the right. In Fig. 13.2, for example, the bicycle and cabinet in row 1 should go to the black and red houses, but we do not know which picture goes to which house. The child has to compare rows 1 and 2, identify the common elements, and make the inference (e.g., "the bicycle and the black house appear in both rows therefore the bicycle goes to the black house at the top of the page").

The CITM requires planning behavior, systematic exploratory behavior, a strategic and analytic approach, need for accuracy, and control of impulsivity. Although the tasks were novel to the children in both groups, the mental operations required to solve them are relatively familiar and to some degree are also practiced among the Israeli-born

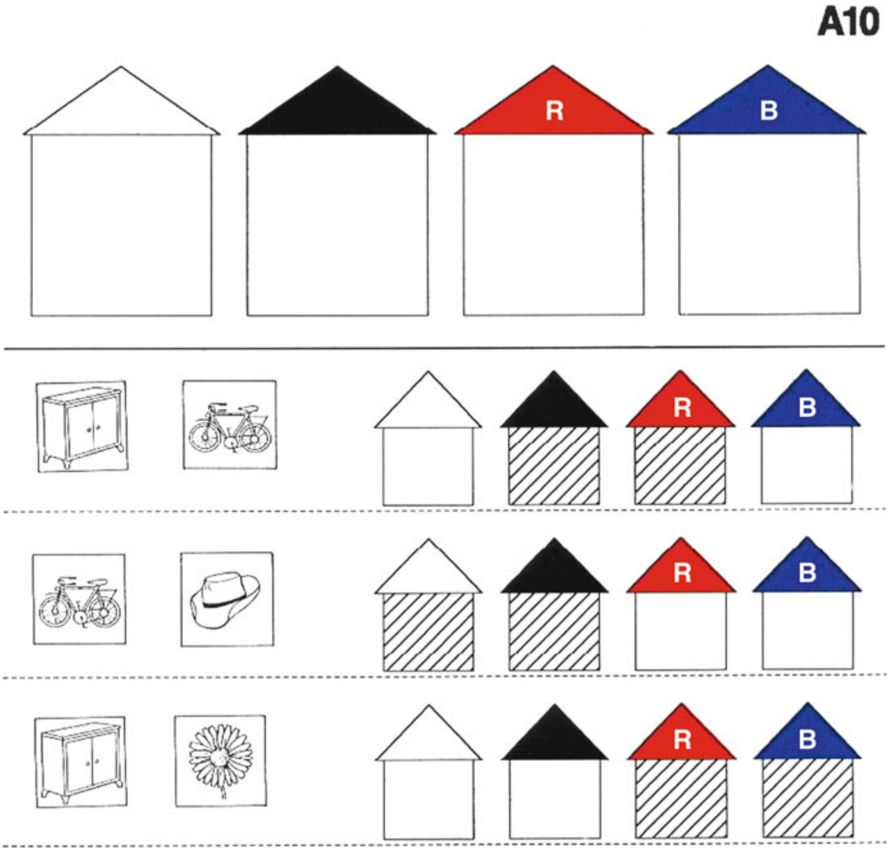


Fig. 13.2 Example of an item from the Children's Inferential Thinking Modifiability (CITM) test (R = Red, B = Blue)

children. For the Ethiopian children, however, these mental activities are new and have no similarity to the type of activities practiced or transmitted in their culture.

The findings showed clearly that the Israeli-born group scored higher than the Ethiopian group on the CPM (static) and the preteaching DA tests. However, the improvement from pre- to postteaching phases of the DA was higher for the Ethiopian than for the Israeli-born group. The findings on The CITM are presented in Fig. 13.3.

As can be seen in Fig. 13.3, the Ethiopian children narrowed the gap on the postteaching phase of the CITM; differences on both postteaching and transfer problems were not significant. The lack of significant differences on the transfer items indicates that the Ethiopian children could benefit from the mediation given to them, internalize the rules, and use them efficiently in the transfer items. The large cognitive change among the Ethiopian children supports both Vygotsky's (1978) ZPD and Feuerstein et al.'s (1979) cognitive modifiability constructs.

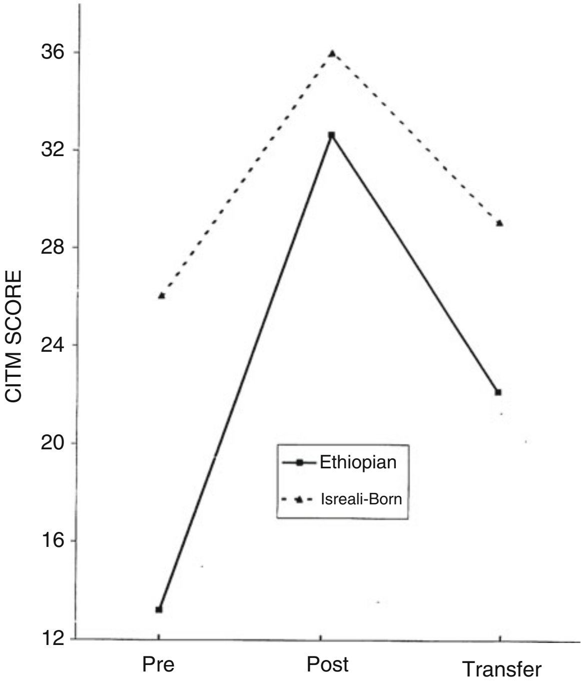


Fig. 13.3 CITM test preteaching, postteaching, and transfer scores of Israeli-born and Ethiopian children (Copied by permission from the *Journal of Cross-Cultural Psychology*)

One of the most intriguing and impressive findings was on the classification phase of the CITM. After finishing the inferential task, children are asked to classify the pictures (cards) presented during the earlier section to categories. There are six categories (e.g., animals, cloths, figures, furniture, means of transportation, plants); each category contains four pictures. Each correctly solved category can get a score of 2 and a maximal score of 12 for all categories. After the first classification phase, all children received a simple mediation phase that lasted between 1 and 2 min in which the principle of classification was explained.

The Ethiopian children achieved a dramatic and significant gain from .70 to 9.00 as compared with a gain from 10.20 to 12.00 among the Israeli-born children who reached a ceiling. It should be noted that the low initial score of the Ethiopian children was not a result of inadequate instruction but of a different understanding of what is expected to perform. For example, a typical classification of objects in the premediation phase among Ethiopian children could be a donkey, a leaf, and a circle. When asked why these three pictures are classified together, the answer was "because the donkey eats the leaf by the well (circle)." After a simple explanation of the meaning of a class (e.g., donkey, dog, cat, and bird; all of them belong to the family of animals), the improvement was drastic. These results coincide with cross-cultural research findings indicating that individuals in many non-Western nations classify items into functional rather than into taxonomic categories (e.g., Greenfield 1997).

In a recent study, ethnic minority children in the Netherlands were compared to indigenous children on a DA test: the Seria-Think Instrument (Tzuriel 2000) using a graduated prompt technique (Resing et al. 2009). The findings showed that children tested by DA changed their strategy behavior into the direction of a more advanced strategy and that this change was the largest for the initial weaker scoring ethnic minority children. More specifically, ethnic minority children initially needed more, but then progressively needed fewer, cognitive hints than did the indigenous children. These findings show that ethnic minority children need support in order to know what to solve and how to do it. Once the situation was clarified, they showed greater progression toward superior strategy use.

13.6.2 Evaluating the Effects of Cognitive Education Programs by DA

DA has been used frequently to assess the effectiveness of cognitive intervention programs. The rationale of using DA is matching the declared objective of the cognitive program (e.g., "learning how to learn") with criterion measures of change and modifiability. DA has been used for evaluating four cognitive intervention programs: *Instrumental Enrichment* (*IE,* Feuerstein et al. 1980), *Bright Start* (Haywood et al. 1986), *Peer Mediation for Young Children (PMYC,* Shamir and Tzuriel 2004; Tzuriel and Shamir 2007, 2010), and the *Analogical Reasoning Program (ARP,* Tzuriel and George 2009). The findings of several studies show clearly that the effectiveness of the program could be revealed only when DA approach was applied. Because of space limitation, I will present here two recent studies, one on *Bright Start* and the second on PMYC program. For a detailed review on revealing the effectiveness of DA in evaluating cognitive education programs, see Tzuriel (2011).

In the first study (Tzuriel et al. 1999), a sample of kindergartners received the Bright Start in their classrooms (*n*=82) and was compared to a group of children (*n*=52) who received a basic skills program. The Bright Start program was applied for 10 months, during which the children in the experimental group received five of the seven small-group units: self-regulation, quantitative relations (number concepts), comparison, classification, and role-taking. The small-group lessons were taught three times a week, each session for a period of 20 min, for a total of 1 h per week and a total number of 32 h for the academic year. The comparison group was given the basic skills program during the academic year, and the teachers were visited periodically to observe their skills-based program. Two DA instruments were administered: the CATM and a young children's version of the *Complex Figure* test (Tzuriel and Eiboshitz 1992). Since the finding of the Complex Figure is very similar to those of the CATM, only the CATM findings are reported here.

After gathering the preintervention data, we realized that the cognitive scores of the experimental group were lower than those of the comparison group. Unfortunately, there was no possibility of random assignment of children in each class to the treatment groups without raising the parents' resentment. It would also have been confusing to the kindergarten teacher who would have had to implement both

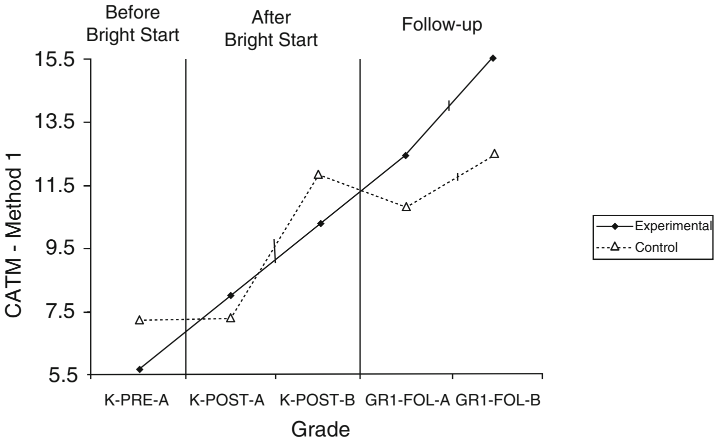


Fig. 13.4 The CATM Scores of the Experimental and Comparison Groups Before and After the Bright Start Program, and in the Follow-up Phase (Copied by permission from *Early Childhood Research Quarterly, ECRQ*) [(K = Kindergarten, GR1 = Grade 1, Pre-A = Pre-Intervention Preteaching Problems (A), Post-A = Post-Intervention Preteaching problems (A), Post-B = Post-Intervention Postteaching Problems (B), Fol-A = Follow-up Preteaching problems (A), Fol-B = Follow-up Postteaching Problems (B)]

programs within one class. We had to rely, therefore, on supervisors' assessments of children's background as a basis for equating the treatment groups. This eventually proved to be not completely accurate.

Group comparison on CATM and Complex Figure pre- and postteaching scores was carried out at the end of the intervention and in a follow-up phase 1 year after the end of the program. A MANOVA of treatment (experimental vs. comparison) by phase (pre- vs. postteaching) and by grade (*K* vs. grade 1) was carried out on the CATM scores. The analysis revealed a significant triadic interaction of treatment by grade by pre-/postteaching, $F_{(2, 69)}=4.27$, $p<.02$. The interaction is portrayed in Fig. 13.3. For comparative reasons, the CATM scores at the start of the program are also plotted in Fig. 13.3; however, the analysis is based only on students who participated in the follow-up.

Figure 13.4 shows both static and DA results. The static tests results are portrayed in CATM scores before and after the intervention (set A, preteaching). The findings show that children in the experimental group made higher improvement on the CATM scores (set A) from preintervention (*K*-Pre-A) to postintervention (*K*-Post-A) phase. When the CATM was administered in a DA procedure, the findings were intriguing. While at the end of the program (*K*) the comparison children improved their performance from the pre- to postteaching phase of the DA test more than did the experimental children, in the follow-up year (grade 1) the trend was reversed! The experimental group showed higher improvement from pre- to postteaching than did the comparison group.

These results in grade 1 were interpreted as an indication for a "snowball" effect of the "learning to learn" treatment. According to the "snowball effect," treatment effects gain power with time without any additional treatment, which is to be expected when the treatment is designed to enhance "learning to learn" skills. Further support for the "snowball effect" was found when cognitive modifiability indices were taken as the dependent variable. Cognitive modifiability indices were calculated by regression analysis in which the residual postteaching scores were derived after controlling for the preteaching score (see Embreston 1987, 1992).

A MANOVA of treatment by grade (2×2) applied on the CATM cognitive modifiability indices revealed a significant overall interaction of treatment by grade, $F_{(2,69)} = 10.08$, $p < .0001$. This finding indicates higher improvement of the cognitive modifiability scores in the experimental than in the comparison group, from kindergarten to first grade.

13.7 Criticism on Dynamic Assessment

A frequent criticism mentioned in the literature is that DA takes more time to administer and requires more skill, better training, more experience, and greater effort than static testing do (Frisby and Braden 1992). It is true that the professional skill necessary to do DA effectively is not currently taught in typical graduate psychology programs, so practitioners must be trained in intensive workshops long after they have been indoctrinated in the "laws" of static, normative testing (Haywood and Tzuriel 2002). Even with excellent training, DA examiners must exercise considerable subjective judgment in determining (a) what cognitive functions are deficient and require mediation, (b) what kinds of mediation to dispense, (c) when further mediation is not needed, and (d) how to interpret the difference between premediation and postmediation performance. It seems somehow disingenuous to complain that DA requires special knowledge and special skills when its benefits are directly related to such knowledge and skills and in turn have benefits for the children.

Another criticism is that the extent to which *cognitive modifiability* is generalized across domains (i.e., analogical, numerical) needs further investigation. Related to this criticism is the question of how to translate the DA findings into effective instruction and intervention. This aspect is considered as a major educational advantage over static testing.

The relative lack of reliability is another major criticism. Establishing reliability and validity of DA is much more complex than validation of static testing because it has a broader scope of goals. The question of reliability is a pressing one, especially so given that one sets out deliberately to change the very characteristics that are being assessed. At least a partial solution is to insist on very high reliability of the tasks used in DA when they are given in a static mode, i.e., without interpolated mediation. Another solution is to use interjudge reliability based on observations. This aspect has been studied to some extent (e.g., Tzuriel and Samuels 2000) but not yet sufficiently.

Another persistent problem is how to establish the validity of DA. Ideally, one would use both static testing and DA with one group of children and static, normative ability tests with another group. The essential requirement would be that a subgroup of the DA children would have to be given educational experiences that reflected the within-test mediation that helped them to achieve higher performance in DA. The expectation would be that static tests would predict quite well the school achievement of both the static testing group and that subsample of the DA group that did not get cognitive educational follow-up. Static tests should predict less well the achievement of the DA-cognitive education group; in fact, the negative predictions made for that group should be defeated to a significant degree (Haywood and Tzuriel 2002).

One of the criticism raised by Frisby and Braden (1992) is that the literature is replete with evidence showing a strong relation between IQ and school achievement ($r=.71$). The question therefore is why applying a DA approach if so much of the variance in school learning is explained by standardized testing? The last point means that nearly 50% of the variance in learning outcomes for students can be explained by differences in psychometric IQ. My answer to the last point, being loyal to a meditational approach of inquiring and probing, is by asking three extremely important questions (Tzuriel 1992). These questions are graduated from light to heavy:

(a) What causes the other 50% of achievement variance?
(b) When IQ predicts low achievement, what is necessary to defeat that prediction?
(c) What factors influencing the unexplained variance can help to defeat the prediction in the explained variance?

13.8 Why DA Is Not Applied on a Larger Scale?

One might well ask why, if DA is so rich and rewarding, it is not more widely applied? Here are some possible answers (Karpov and Tzuriel 2009):

- One apparent reason is that it is not taught in graduate school yet.
- School psychologists often have "client quotas" to fill, and DA is far more time-consuming that is static testing, so their supervisors do not permit it.
- The school personnel who ultimately receive the psychologists' reports typically do not expect DA and do not yet know how to interpret the data or the recommendations, and psychologists have not been good enough about helping them on that score.
- There is a certain inertia inherent in our satisfaction with being able to do what we already know how to do and to do it exceptionally well. Even so, as we have observed before, "what is not worth doing is not worth doing well!"

The question of what should be done is complex as the answer depends on a myriad of interrelated factors. Haywood (2008) suggested that the most urgent task is to explore and incorporate new models of the nature of human ability. He suggested, as one such model, a "transactional" perspective on human ability with three major dimensions: intelligence, cognitive processes, and motivation, especially task-intrinsic motivation. The concept of intelligence, then, is not seen as useless or as antithetical to the notion of cognitive processes, structures, or strategies but as a construct that does not explain all that we know about individual differences in learning and performance effectiveness. We can supplement its explanatory value by adding the dimensions of cognitive processes and motivation. One should proceed from some such model of the nature of ability to define what it is that we wish to assess and only then to construct instrument for assessing individual differences in that set of variables.

References

Carlson, J. S., & Wiedle, K. H. (1992). Principles of dynamic assessment: The application of a specific model. *Learning and Individual Differences, 4*, 153–166.

Embretson, S. E. (1987). Toward a development of a psychometric approach. In C. S. Lidz (Ed.) *Dynamic assessment* (pp. 141–170). New York: Guilford Press.

Embreston, S. E. (1992). Measuring and validating cognitive modifiability as ability: A study in the spatial domain. *Journal of Educational Measurement, 29*, 25–50.

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device: Theory, instruments, and techniques*. Baltimore: University Park Press.

Feuerstein, R., Rand, Y., Hoffman, M. B., & Miller, R. (1980). *Instrumental enrichment*. Baltimore: University Park Press.

Frisby, C. L., & Braden, J. P. (1992). Feuerstein's dynamic assessment approach: A semantic, logical and empirical critique. *Journal of Special Education, 26*, 281–301.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*, 45–56.

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development, 62*, 1–22.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*, 1115–1124.

Guthke, J., & Al-Zoubi, A. (1987). Kulturspezifische differenzen in den Colored Progressive Matrices (CPM) und in einer Lerntestvariante der CPM [Culture specific differences in the Colored Progressive Matrices (CPM) and in learning potential version of the CPM]. *Psychologie in Erziehung und Unterricht, 34*, 306–311.

Guthke, J., & Stein, H. (1996). Are learning tests the better version of intelligence tests? *European Journal of Psychological Assessment, 12*, 1–13.

Guthke, J., & Wingenfeld, S. (1992). The learning test concept: Origins, state of the art, and trends. In H. C. Haywood & D. Tzuriel (Eds.), *Interactive assessment* (pp. 64–94). New York: Springer.

Hamers, J. H. M., Hessels, M. G. P., & Van Luit, J. E. H. (1991). *Learning potential test for ethnic minorities: Manual and test*. Lisse: Swets & Zeitlinger.

Haywood, H. C. (1968). Motivational orientation of overachieving and underachieving elementary school children. *Journal of Personality, 30*, 63–74.

Haywood, H. C. (1971). Individual differences in motivational orientation: A trait approach. In P. I. Day, D. E. Berlyne, & D. E. Hunt (Eds.), *Intrinsic motivation: A new direction in education* (pp. 113–127). New York: Holt, Rinehart, & Winston.

Haywood, H. C. (2008). Twenty years of IACEP, and a focus on dynamic assessment: Progress, problems, and prospects. *Journal of Cognitive Education and Psychology, 7*, 419–442.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York: Cambridge University Press.

Haywood, H. C., & Tzuriel, D. (Eds.). (1992). *Interactive assessment*. New York: Springer.

Haywood, H. C., & Tzuriel, D. (2002). Applications and challenges in dynamic assessment. *Peabody Journal of Education, 77*, 38–61.

Haywood, H. C., Brooks, P. H., & Burns, M. S. (1986). Stimulating cognitive development at developmental level: A tested, non-remedial preschool curriculum for preschoolers and older retarded children. In M. Schwebel & C. A. Maher (Eds.), *Facilitating cognitive development: Principles, practices, and programs* (pp. 127–147). New York: Haworth Press.

Hessels, M. G. P. (2000). The learning potential test for ethnic minorities: A tool for standardized assessment of children in kindergarten and the first years of primary school. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 109–131). New York: Elsevier.

Hessels, M. G. P., & Hamers, J. H. M. (1993). A learning potential test for ethnic minorities. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment* (pp. 285–311). Lisse: Swets & Zeitlinger.

Hessels-Schlatter, C. (2002). Dynamic test to assess learning capacity in people with severe impairments. *American Journal on Mental Retardation, 107*(5), 340–351.

Holyoak, K. J., & Thagard, P. (1997). The analogical mind. *American Psychologist, 52*, 35–44.

Karpov, Y. & Tzuriel, D. (2009). Dynamic assessment: Progress, problems, and prospects. *Journal of Cognitive Education and Psychology, 8*, 228–237.

Lidz, C. S., & Elliott, J. G. (Eds.). (2000). Advances in cognition and educational practice. In *Dynamic assessment: Prevailing models and applications*. Oxford: Elsevier.

Raven, J. C. (1956). *Guide to using the Colored Progressive Matrices, sets A, Ab, and B*. London: Lewis.

Resing, W. C. M. (1997). Learning potential assessment: The alternative for measuring intelligence? *Educational and Child Psychology, 14*, 68–82.

Resing, W. C. M., Tunteler, E., de Jong, F. M., & Bosma, T. (2009). Dynamic testing in indigenous and ethnic minority children. *Learning and Individual Differences, 19*, 445–450.

Shamir, A., & Tzuriel, D. (2004). Children's mediational teaching style as a function of intervention for cross-age peer-mediation. *School Psychology International, 25*, 58–97.

Shochet, I. M. (1992). A dynamic assessment for undergraduate admission: The inverse relationship between modifiability and predictability. In H. C. Haywood & D. Tzuriel (Eds.), *Interactive assessment* (pp. 332–355). New York: Springer.

Skuy, M., & Shmukler, D. (1987). Effectiveness of the learning potential assessment device for Indian and "colored" South Africans. *International Journal of Special Education, 2*, 131–149.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. New York: Cambridge University Press.

Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C., Jukes, M., & Bundy, D. A. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence, 30*, 141–162.

Tzuriel, D. (1989). Inferential cognitive modifiability in young socially Israeli-born and advantaged children. *International Journal of Dynamic Assessment and Instruction, 1*, 65–80.

Tzuriel, D. (1992). The dynamic assessment approach: A reply to Frisby and Braden. *Journal of Special Education, 26*, 302–324.

Tzuriel, D. (1997). A novel dynamic assessment approach for young children: Major dimensions and current research. *Educational and Child Psychology, 14*, 83–102.

Tzuriel, D. (2000). Dynamic assessment of young children: Educational and intervention perspectives. *Educational Psychology Review, 12*, 385–435.

Tzuriel, D. (2001). *Dynamic assessment of young children*. New York: Kluwer Academic/Plenum Press.

Tzuriel, D. (2002). Dynamic assessment of learning potential. In J.W. Guthrie (Ed.), *Encyclopedia of education* (2nd ed., pp. 127–131). New York: McMillan Press.

Tzuriel, D. (2011). Revealing the effects of cognitive education programs by dynamic assessment. *Assessment in Education: Principles, Policy and Practice, 18*(2), 113–131.

Tzuriel, D., & Caspi, N. (1992). Dynamic assessment of cognitive modifiability in deaf and hearing preschool children. *Journal of Special Education, 18*, 113–131.

Tzuriel, D., & Eiboshitz, Y. (1992). A structured program for visual motor integration (SP-VMI) for preschool children. *Learning and Individual Differences, 4*, 103–124.

Tzuriel, D., & George, T. (2009). Improvement of analogical reasoning and academic achievements by the Analogical Reasoning Program (ARP). *Educational and Child Psychology, 29*, 71–93.

Tzuriel, D., & Kaufman, R. (1999). Mediated learning and cognitive modifiability: Dynamic assessment of young Ethiopian immigrants in Israel. *Journal of Cross-Cultural Psychology, 30*, 359–380.

Tzuriel, D., & Kaufman, R. (1998). Mediated learning and cognitive modifiability: Dynamic assessment of young Ethiopian immigrants in Israel. *Journal of Cross-Cultural Psychology, 13*, 539–552.

Tzuriel, D., & Klein, P. S. (1985). Analogical thinking modifiability in disadvantaged, regular, special education and mentally retarded children. *Journal of Abnormal Child Psychology, 13*, 539–552.

Tzuriel, D., & Samuels, M. T. (2000). Dynamic assessment of learning potential: Inter-rater reliability of deficient cognitive functions, type of mediation, and non-intellective factors. *Journal of Cognitive Education and Psychology, 1*, 41–64.

Tzuriel, D., & Shamir, A. (2007). The effects of peer mediation with young children (PMYC) on children's cognitive modifiability. *British Journal of Educational Psychology, 77*, 143–165.

Tzuriel, D., & Shamir, A. (2010). Mediation strategies and cognitive modifiability in young children as a function of peer mediation with young children (PMYC) program and training in analogies versus math tasks. *Journal of Cognitive Psychology and Education, 9*, 48–72.

Tzuriel, D., Samuels, M. T., & Feuerstein, R. (1988). Non-intellective factors in dynamic assessment. In R. M. Gupta & P. Coxhead (Eds.), *Cultural diversity and learning efficiency: Recent developments in assessment* (pp. 141–163). London: NFER-Nelson.

Tzuriel, D., Kaniel, S., Kanner, E., & Haywood, H. C. (1999). The effectiveness of Bright Start program in kindergarten on transfer abilities and academic achievements. *Early Childhood Research Quarterly, 114*, 111–141.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wiedl, K. H. (2003). Dynamic testing: A comprehensive model and current fields of application. *Journal of Cognitive Education and Psychology, 3*, 93–119.

Chapter 14
Exploiting Computerized Adaptive Testing for Self-Directed Learning

Chia-Ling Hsu, Yue Zhao, and Wen-Chung Wang

14.1 Introduction

Self-directed learning enables learners to take responsibility for and to design their own learning-related activities. Frequent assessments facilitate self-directed learners to monitor learning regularly. Often, quizzes with multiple-choice (MC) items are a quick way of gathering information on the strengths and weakness of learners. Although with limitations, well-designed MC items can assess both low-level (e.g., fact recall and comprehension) and high-level (e.g., application, analysis and evaluation) thinking. MC items have played a major role in both classroom and large-scale assessments for some time. Formats such as MC, true-false, and fill-in-the-gap items are particularly suitable for online computerized testing, largely because they can be scored by computers. This chapter introduces the theory and practice of computerized testing.

Tests, such as achievement tests, licensure exams, and attitude questionnaires, are an efficient approach to obtaining information about test-takers. For example, educational and personnel tests are commonly used for admission, placement, or promotion, and licensure exams are usually used to evaluate the qualifications of

C.-L. Hsu • Y. Zhao
Assessment Research Centre, The Hong Kong Institute of Education (HKIEd),
Tai Po, Hong Kong
e-mail: clhsu@ied.edu.hk; myzhao@hku.hk

W.-C. Wang (✉)
Department of Psychological Studies, and Assessment Research Centre,
The Hong Kong Institute of Education (HKIEd), Tai Po, Hong Kong
e-mail: wcwang@ied.edu.hk

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_14, © Springer Science+Business Media Dordrecht 2013

professionals. In a broad sense, clinical inquiry or medical equipment assessments can be regarded as "tests" to help physicians diagnose patients for follow-up treatments. Whether in performance assessment of students' learning outcomes, job assignment for employees, medical diagnoses of patients, or qualifications for professionals, reliable and valid assessment instruments always play an essential role.

Paper-and-pencil (P&P) testing is the most widely used testing approach. Normally, all test-takers are administered the same test at the same time in the same location and under the same conditions. P&P testing is convenient for administration; however, it has certain limitations. First, P&P testing may lack quality. P&P testing cannot measure test-takers with three-dimensional space, sound, speed, and interaction. Second, P&P testing may lack flexibility. All test-takers must be present at the same time at the designated place (e.g., college entrance or national licensure examinations), which exacts a huge social cost. There may be serious consequences if test-takers have accidents or are in poor health prior to or during testing. Third, P&P testing may lack efficiency. Test-takers must answer all items in the test. More capable test-takers may waste a lot of time to answer items that are too easy, while less capable test-takers may be forced to answer items that are too difficult. Modern testing should be able to provide high quality, flexibility, and efficiency.

Due to the advancement of computer technology and the Internet, the three above-mentioned limitations have been greatly ameliorated. Computers can successfully make use of three-dimensional space, sound, and speed, and they allow for human-computer interaction. Such testing is called computer-based testing (CBT). Well-known CBT includes the Test of English as a Foreign Language and the Graduate Record Examinations, both of which are designed and administered by the Educational Testing Services (ETS), the General Educational Development tests, and the Minnesota Multiphasic Personality Inventory. The current Hong Kong "student assessment," which has been implemented from primary 1 to secondary 3 and covers Chinese, English, and Mathematics, is also an example of CBT. Because of the combination of multimedia and the Internet, CBT has several advantages over P&P testing. First, CBT allows the presentation of test items to be more diverse. Items presented with sound or animation are more realistic and interesting than written items. Hence, CBT appears to be more capable of measuring authentic abilities, and this increases test validity. Second, CBT enables the standardization of the testing environment. The entire administration process (e.g., announcing directions, delivering items, and scoring items) is controlled and monitored by computers, which avoids confounding human factors. Thus, higher test reliability can be achieved. Third, due to network technology advances, test-takers are not required to take tests at a fixed time and location. CBT allows test-takers to take tests at a convenient time and place as long as examination systems can monitor them, which provides higher flexibility. Finally, there is a shorter waiting time for score reports. Usually after completing the test, test-takers can obtain their test results immediately.

Although CBT increases test quality and flexibility substantially, it does not improve test efficiency because all test-takers are still administered the same test. Recent developments in computerized adaptive testing (CAT) solve this problem. In CAT, an item is selected from an item bank that adapts or is tailored to the test-taker's ability levels, such that fewer items are needed for CAT to achieve the same degree of precision as nonadaptive testing (e.g., P&P testing or CBT). Basically, test-takers do not need to waste their time answering very difficult or very easy items in relation to their ability levels. Therefore, CAT not only possesses the quality and flexibility advantages of CBT but also improves test efficiency. Well-known large-scale CAT programs include the Graduate Management Admission Test, the general test of the Graduate Record Examinations, and the Armed Service Vocational Aptitude Battery.

14.2 Computer-Based Testing

14.2.1 Advantages of CBT

Due to advances in computer technology, more people have become computer literate, and the age at which individuals start using computers is decreasing. With the help of multimedia capabilities, fast processing power, and large data storage, the development of CBT has made great progress in recent years. Currently, many large-scale assessments have developed CBT versions. Compared to conventional P&P testing, the main advantages of CBT include standardized testing environment, diversified item presentations, prompt scoring, collecting responding behavior, and lack of space and time limits. Each of these is discussed in more detail below.

14.2.1.1 Standardized Testing Environment

Test directions, time control, scoring, and score reports are all processed by computers, which creates a standardized testing environment. In contrast, in P&P testing, different test-takers may receive different messages from different test administrators who read the test directions in different tones or at different speeds; they also may have different testing times and may receive different ratings from different raters.

14.2.1.2 Diversified Item Presentation

Because of multimedia features, test items are no longer limited to written text; rather, they can be presented together with image, sound, or animation, etc. On the learning side, multimedia features can help students better acquire in-depth knowledge and skills. On the assessment side, multimedia features can help evaluate the authentic abilities of test-takers. A variety of item formats is illustrated in Table 14.1 (Chen 2005).

Table 14.1 Common item presentations in CBT

Item type	Presentation
Text	Same as P&P testing, except that items are presented on computer screens, for example, Chinese (or English) reading passage questions and short answer questions
Colored image	Items with high-resolution color images or photographs. For example, in natural sciences tests, test-takers are asked to judge what the animals and plants are, to determine what the rock structure is, or to identify what cloud formations are in the photos
Video or animation	Items shown in a movie or animation show, for example, an animated presentation of an item regarding crossing the road and traffic lights judgments. This format enables test-takers to better understand the subject content
Audio	Test-takers listen to an audio tape and then respond. This format is widely used in language testing
Interaction or manipulation	Given an item, test-takers are required to interact with computers. For example, in Chinese (or English) typing tests, test-takers need to type in their answers via computer keyboards; or in a simulated driving test for pilots, test-takers are asked to act through a series of instructions

14.2.1.3 Prompt Scoring

Computers' fast computing speed makes it possible to calculate test scores immediately after test-takers finish tests. When appropriate, test scores can be released at that very moment.

14.2.1.4 Collecting Responding Behavior

Not only are test-takers' responses recorded but their responding behavior, such as response time and response expressions, can be recorded as well, which may help facilitate the interpretation of test scores.

14.2.1.5 Beyond Limits of Time and Space

With the Internet, CBT can offer great flexibility for test-takers to choose when and where they would like to take tests, as long as the computers are equipped and the network connection speed is guaranteed and stable.

14.2.2 Test Delivery Models

Six test delivery models are commonly used in operational testing programs and intensively evaluated in the literature: (a) computerized fixed testing, (b) linear-on-the-fly testing, (c) item-level CAT, (d) testlet-based CAT, (e) structured multistage testing, and (f) computerized classification testing. This section discusses each of these in some detail.

14.2.2.1 Computerized Fixed Testing

In computerized fixed testing, different (but parallel) test forms are administered to different groups of test-takers via computers. These test forms are parallel in terms of content representations and psychometric properties, such as item discrimination and difficulty. For example, there may be five different test forms, and each form may consist of 50 MC items. In addition, test-takers who are taking the same test form answer exactly the same set of items, except that the presentation sequence may be scrambled or randomized at runtime. In other words, computerized fixed testing is analogous to P&P but administered through the mode of computers. In addition to the general advantages of CBT over P&P, computerized fixed testing may prevent certain types of cheating because of its scrambled or randomized item presentation sequence.

14.2.2.2 Linear-On-The-Fly Testing

In computerized fixed testing, a few test forms are assembled prior to test administration. Suppose, for example, that there are five test forms, and each form has 50 MC items. In this case, which 50 MC items go on which test form is determined before actual testing. In contrast, in linear on-the-fly testing (LOFT), test forms are automatically assembled from a large item bank in advance of or immediately prior to test administration (Drasgow et al. 2006). In LOFT, each test-taker can be administered a unique test form (Folk and Smith 2002). The term “linear” means that the items are administered in a linear way (i.e., nonadaptively), and the term “on-the-fly” means that items are assembled on the spot. Automatic test assembly from a large item bank can take into account content coverage and the psychometric properties of test forms. Item exposure can be well controlled. When test forms are assembled in advance, content and measurement experts will have a chance to review each test form for quality assurance. However, because of its linear nature, LOFT does not consider the ability levels of test-takers in test assembly.

14.2.2.3 Item-Level Computerized Adaptive Testing

Across category boundaries, item-level CAT is one of the most widely studied delivery models. Each test-taker takes a different set of items. The selection of items adapts (tailors) the item difficulty to each test-taker’s ability level at the item level. In CAT, in addition to the consideration of content requirements and psychometric properties, the choice of a successive item presented to a test-taker is based on an ability estimate, which is calibrated from previously administered items. CAT offers more precise estimates of test-takers' ability levels by implementing a shorter test length compared to nonadaptive testing. More details regarding how CAT works are described in the following sections.

14.2.2.4 Testlet-Based Computerized Adaptive Testing

A testlet is a set of items that are connected with a common stimulus (e.g., a reading passage or a figure) (Lewis and Sheehan 1990; Sheehan and Lewis 1992; Wainer and Lewis 1990). Generally, items in the same testlet are administered as a whole. An item-level CAT can be directly adapted to a testlet-based CAT, where a testlet is viewed as a "virtual" item. Conceptually, the selection of a testlet from a testlet pool is carried out by adapting the testlet difficulty to each test-taker. When a testlet is selected, all items in the testlet are administered linearly (nonadaptively).

14.2.2.5 Structured Multistage Testing

Analogous to the testlet-based CAT, structured multistage testing uses a set of items as building blocks to form a test (Zenisky et al. 2010). The set of items is usually described as a module or testlet (Luecht and Nungester 1998; Wainer and Kiely 1987). Like in testlet-based CAT, once a module is selected for administration, all items in the module are administered in a linear way. Of course, different test-takers may receive different sets of modules in different sequences. The term "stage" refers to an administrative division of the test to facilitate the adaption to test-takers. At each stage, a test-taker is presented with a module that is adapted in difficulty to the test-taker's ability estimate, based on his or her performance on the modules that were administered at the previous stage. In other words, the adaptation in structured multistage testing is at the module level, rather than at the item level. Within each stage, there are typically a few modules that differ from one another in terms of average difficulty (Jodoin et al. 2006).

An example of structured three-stage testing, which has three stages, is illustrated in Fig. 14.1. All test-takers are administered the same module at the first stage. There are two modules (easy and hard) at the second stage and three modules (easy, medium, and hard) at the third stage. In other words, there are six combinations of modules (also called routes) that a test-taker may receive. For a low-ability test-taker, he or she may be presented with the following route: stage 1, followed by the easy module at stage 2, and then the easy module at stage 3. In contrast, a more capable test-taker may be administered the test through a different route, for example, stage 1, followed by the hard module at stage 2, and then the hard module at stage 3.

Structured multistage testing is a variant of CAT whose adaptation level is at the module level. It preserves the same measurement efficiency as the item-level or testlet-level CAT because of its adaptive nature. Compared to the item-level CAT, it

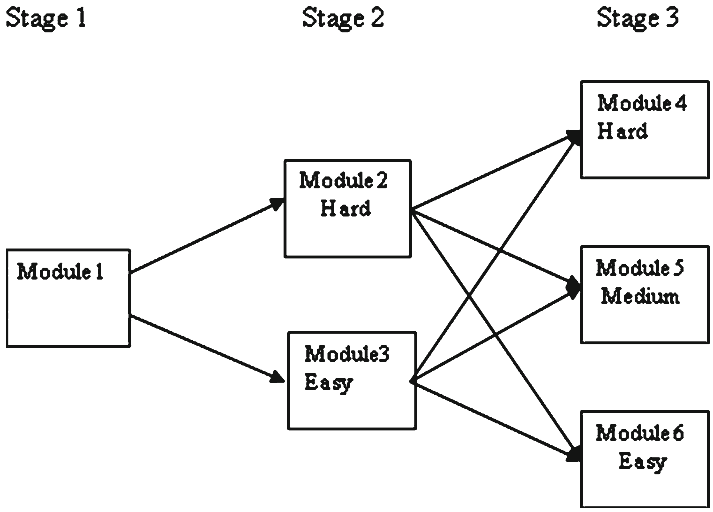


Fig. 14.1 Example of structured three-stage testing

has lower measurement precision, but it requires smaller item banks, so the cost of developing, testing, and implementing new items is reduced. Furthermore, content experts can be recruited to review items within each module (Drasgow et al. 2006). Nowadays, many commercial test delivery vendors have incorporated structured multistage testing in their testing systems, including the Uniform Certified Public Accountant Examination.

14.2.2.6 Computerized Classification Testing

The goal of CAT is to obtain precise estimates of ability levels of test-takers in an efficient way. Sometimes, test users are not very interested in precise estimates of ability levels but rather may be more interested in classifying test-takers into a limited number of categories (e.g., fail or pass; advanced, proficient, or basic; normal, marginal, or abnormal). CAT can be modified to attain this classification goal. The modified CAT is referred to as computerized classification testing (CCT) (Eggen 2010; Wang and Liu 2011; Wang and Huang 2011). CCT is commonly used in placement tests, certification, and licensure examinations. CCT usually requires fewer items than CAT to achieve its classification goal. This advantage is especially apparent for test-takers with extreme ability levels, because there may not be sufficient items in an item bank that can adapt to test-takers with extreme ability levels. In these cases, CAT would become much less efficient than CCT.

In CCT, a test-taker is to be classified into categories according to some prespecified cut scores. For example, ability levels below a certain cut score are classified as "basic," while those above another cut score as "advanced," and those between these two cut scores as “proficient.” The sequential probability ratio test (Wald 1947) and the ability confidence interval (Thompson 2009) are two commonly used classification criteria. The sequential probability ratio test formulates the decision process as the testing of a hypothesis in which the examinee’s ability estimate is equal to a specified point above the cut score or another specified point below the cut score. On the other hand, the ability confidence interval, originally termed adaptive mastery testing (Kingsbury and Weiss 1983), terminates the test when a confidence interval for the test-taker’s ability estimate is completely above or below the cut score. In general, when the sequential probability ratio test is used, only a small subset of items in the item bank will be administered, and many items will never or seldom be used. In contrast, when the ability confidence interval is used, a large subset of items will be administered, which makes the item bank more cost-effective (Wang and Liu 2011).

14.3 Computerized Adaptive Testing

CAT is a testing technology that combines CBT and adaptive testing. It includes all the advantages of CBT. In addition, CAT can achieve the same measurement precision as nonadaptive testing with notably fewer items. The next item from an item bank is to be selected if its difficulty best matches the provisional ability estimate of the test-taker. In another words, the adaptation is realized by tracing a test-taker’s performance on each item and then using this information to select the most optimal item to administer next. The item selection criterion in CAT aims to maximize the information about a test-taker’s ability level and thereby minimize the measurement error of a test-taker’s ability level. Typically, CAT requires only about half of the total number of items in P&P testing to reach the same measurement precision (Wainer 2000).

In P&P testing, all test-takers must respond to all items in the same test form. High-ability test-takers have to spend time answering very easy items, which might bore them or reduce their motivation in testing. Conversely, low-ability test-takers have to answer very difficult items, which might frustrate them. More importantly, items that are too difficult or too easy provide little information about the ability levels of test-takers. For example, if easy items are administered to high-ability test-takers, it is very likely that almost everyone will obtain close to full score points. It is thus hard to distinguish their ability differences. Similarly, if difficult items are administered to low-ability test-takers, it is very likely that they will receive close to zero score points. In theory, the most effective item for a test-taker is the one with medium difficulty (for example, the probability of being correct for the test-taker is 50%). When all test-takers take the same test, it is inevitable that high-ability test-takers will have to answer easy items and low-ability test-takers will have to answer difficult items, such that measurement efficiency cannot be improved.

To improve measurement precision, a test has to be tailor-made for every test-taker. Since test-takers often have very different ability levels, a test has to consist of a large number of items with a wide range of difficulty. Such a large test is called an item bank. A particular set of items are then drawn from the item bank to best match a test-taker's ability.

In CAT, testing time can be substantially reduced because CAT requires notably fewer items than nonadaptive (e.g., P&P) testing to achieve the same measurement precision. Sometimes, saving testing time may not generate a great economic value in school settings; it may, however, be very beneficial in other cases. For example, in the medical setting, the test-takers are patients, and proxies (nurses or doctors) may be needed to help patients take a test. Saving patients' or proxies' time is often very desirable. Time-saving in self-directed learning is also desirable because learners often need to take tests frequently in order to monitor their learning progress, and reducing testing time can ease their fatigue and maintain their motivation in assessment and learning.

14.4 Theory and Practice

14.4.1 Common Models

Both CAT and CCT are possible because of item response theory (IRT), which is a measurement theory that uses a monotonically increasing function (called an item characteristic function or item characteristic curve) to describe the relationship between a test-taker's item performance (e.g., correct or incorrect answer to an item, agree or disagree with a statement) and his or her latent trait underlying item performance (Lord 1980). A latent trait can be ability (e.g., mathematical problem-solving skills or English proficiency) or a nonability (e.g., attitude, value, or personality). In this chapter, we use "ability" to represent any kind of latent trait.

There are many IRT models. Among them, the Rasch simple logistic model (Rasch 1960) appears to be the most well known. Let P_{ni} denote the probability of being correct, θ_n denote test-taker n's ability level, and b_i denote item i's difficulty. The Rasch simple logistic model can be expressed as

$$P_{ni} = \frac{\exp(\theta_n - b_i)}{1 + \exp(\theta_n - b_i)}, \tag{14.1}$$

where exp() is a conventional way to write down the exponential function e^x and e is the base of the natural logarithms and equals approximately 2.718. When a test-taker's ability equals an item's difficulty (i.e., $\theta_n = b_i$), the probability of being correct is equal to 0.5. When $\theta_n > b_i$, the probability of being correct is greater than 0.5. When $\theta_n < b_i$, the probability of being correct is smaller than 0.5. One may view θ_n and b_i as two sports teams. When $\theta_n > b_i$, meaning that θ_n is stronger than b_i, then the probability of θ_n beating b_i will be greater than 0.5. In contrast, when $\theta_n < b_i$, meaning that θ_n is weaker than b_i, then the probability of θ_n beating b_i will be smaller than 0.5. When $\theta_n = b_i$, meaning that θ_n is the same as b_i, then the probability of θ_n beating b_i will be equal to 0.5.

Since each item has only one parameter (i.e., b_i), the Rasch model is also called the one-parameter logistic model (1PLM) (Hambleton and Swaminathan 1985). This model is simple and has good measurement properties. When real data behave as the Rasch model expects (i.e., when there is good model-data fit), the item difficulty parameters and the ability parameters can be separated, which is referred to as “specific objectivity” (Rasch 1960). In other words, the estimation of ability does not depend on item difficulty, and the estimation of item difficulty does not depend on ability. Furthermore, test raw scores and ability estimates have a one-to-one correspondence, that is, the rank orders of test raw scores and those of ability estimates are identical. Its applications to CAT are also relatively simple.

Another commonly used IRT model is the two-parameter logistic model (2PLM) (Birnbaum 1968), which can be expressed as

$$P_{ni} = \frac{\exp\left[a_i(\theta_n - b_i)\right]}{1 + \exp\left[a_i(\theta_n - b_i)\right]}, \tag{14.2}$$

where a_i is often called the item discrimination parameter or item slope parameter; the others are defined as in Eq. 14.1. Under the 2PLM, there are two kinds of item parameters: a_i and b_i. When $a_i = 1$ for all items, Eq. 14.2 becomes Eq. 14.1. The addition of a_i makes the 2PLM more flexible and easier to fit real data than the 1PLM. However, the good measurement properties of the 1PLM no longer hold for the 2PLM. If measurement quality is of great concern, then the 1PLM should be pursued. If model-data fit is the major concern, then the 2PLM is generally preferred.

It is possible to add another item parameter to Eq. 14.2 to form the so-called three-parameter logistic model (3PLM) (Birnbaum 1968):

$$P_{ni} = c_i + (1 - c_i) \times \frac{\exp\left[a_i(\theta_n - b_i)\right]}{1 + \exp\left[a_i(\theta_n - b_i)\right]}, \tag{14.3}$$

where c_i is often called the pseudo-guessing parameter or asymptotic parameter; the others are defined as in Eq. 14.2. When $c_i = 0$ for all items, Eq. 14.3 becomes Eq. 14.2. The 3PLM has been widely used to describe item responses to MC items. Since most CAT programs use MC items, the 3PLM appears to be more popular than the 2PLM and 1PLM in the CAT context.

In practice, most ability levels θ_n would be within the range of (−3, 3), and the larger the value, the higher the ability. Most b_i would be within (−2, 2), and the larger the value, the more difficult the item. Most a_i would be within (0.5, 2), and the larger the value, the higher the discrimination power the item has. Furthermore, a_i should be positive, suggesting that the higher the ability, the higher the probability of being correct. Most c_i would be within (0, 0.3), and the closer to zero, the smaller the guessing effect on the item. c_i is always not negative by definition. Figure 14.2 shows the item response functions of three hypothetical items A, B, and C. The a_i, b_i, and c_i are 1, 0, and 0 for item A, respectively; 2, 0, and 0.1 for item B, respectively; and 1, 0, and 0.3 for item C, respectively. Item B has a higher a_i than item A, indicating

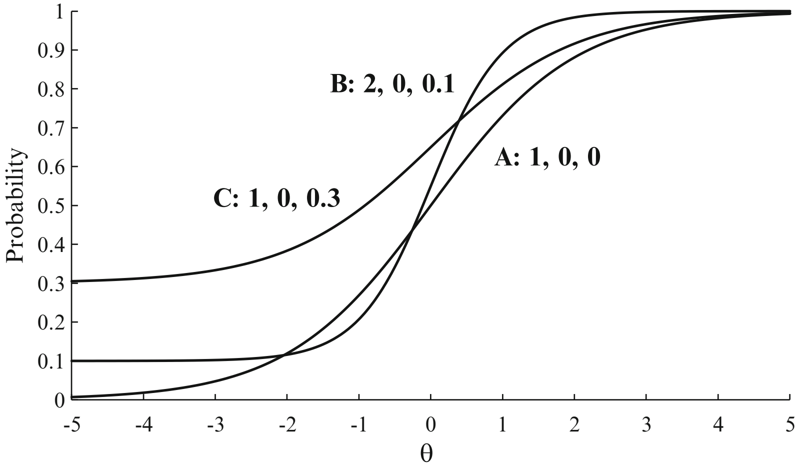


Fig. 14.2 Item response functions of three hypothetical items

that the item response function of item B is steeper than that of item A. Item A has a c_i of zero, indicating the probability of being correct approaches zero when the ability approaches negative infinity. Item B has a c_i of 0.1, indicating the probability of being correct approaches 0.1 when the ability approaches negative infinity.

Although these three items have the same b_i of zero, their item response functions are very different. When c_i is 0, b_i is the location along the ability scale where the probability of being correct is 0.5. When c_i is not zero, b_i is the location where the probability of being correct is $(1+c_i)/2$. The probabilities of being correct for items A, B, and C are 0.5, 0.55, and 0.65, respectively. For MC items, it has been argued that a test-taker can get a correct answer simply by random guessing. This is why the 3PLM has been widely used for MC items. The c_i parameter appears to represent the probability of being correct with random guessing.

Equations 14.1, 14.2, and 14.3 are developed for dichotomous items (i.e., those with two categories). Often, responses to educational or psychological tests may have more than two categories. For example, responses to essay or open-ended items may be given partial credit, and responses to rating scale items (e.g., seldom = 0, sometimes = 1, often = 2) or Likert items (strongly disagree = 0, disagree = 1, agree = 2, strongly agree = 3) are polytomously scored. Several IRT models have been developed for polytomous items. One of the most widely used models is the partial credit model (Masters 1982):

$$P_{nix} = \frac{\exp[\sum_{j=0}^{x}(\theta_n - b_{ij})]}{\sum_{r=0}^{m_i}\{\exp[\sum_{j=0}^{r}(\theta_n - b_{ij})]\}}, \tag{14.4}$$

where P_{nix} is the probability of scoring x on item i for test-taker n, θ_n is test-taker n's ability level, and b_{ij} is the jth step difficulty of item i; $x = 1,\ldots,m_i$, where m_i is the maximum score of item i and $\sum_{j=0}^{0}(\theta_n - b_{ij})$ is zero by definition.

The partial credit model is suitable for open-ended items where each item is marked with its own scoring rubric. In contrast, rating scale items or Likert items in a test are often marked with the same scoring rubric (e.g., all items use the same rating scale structure: seldom, sometimes, and often). The rating scale model (Andrich 1978) was specially developed for this case:

$$P_{nix} = \frac{\exp\left\{\sum_{j=0}^{x}\left[\theta_n - (b_i + e_j)\right]\right\}}{\sum_{r=0}^{m}\exp\left\{\sum_{j=0}^{r}\left[\theta_n - (b_i + e_j)\right]\right\}}, \tag{14.5}$$

where b_i is overall difficulty of item i, e_j is the jth threshold for all items, and m is the maximum score for all items; the others are defined as in Eq. 14.4. $\sum_{j=0}^{0}\left[\theta_n - (b_i + e_j)\right]$ is zero by definition. The choice between Eqs. 14.4 and 14.5 lies in whether all items are marked with the same scoring rubrics. If the answer is yes, then Eq. 14.5 is preferred.

Both Eqs. 14.4 and 14.5 do not have slope parameters. Along with Eq. 14.1, they belong to the family of Rasch models. Where appropriate (e.g., to increase model-data fit), it is straightforward to add slope parameters to Eqs. 14.4 and 14.5. The reader is referred to Chap. 7 (this book) for more information about common IRT models.

14.4.2 Major Steps

In general, CAT and CCT contain five major steps: item bank construction, test starting point, item selection, ability estimation, and test termination. Each of these is discussed in more detail below.

14.4.2.1 Item Bank Construction

To achieve the advantages of adaptive testing, an item bank from which items are selected must contain high-quality items that are applicable to different ability levels. The higher the quality of the item bank, the better the performance of the adaptive testing. A general plan for item bank development usually includes the following (Wainer 2000):

1. Create sufficient numbers of items in each content category, according to previously established test specifications.

2. Recruit test specialists to review item quality and sensitivity.
3. Perform an initial pretesting of the newly written items. In spite of the potential problems with conversion to computer format from P&P format, the initial item bank may have to be created in P&P format because sufficient numbers of computer testing stations may not initially be available.
4. Select a subset of items, both on the basis of conventional item analysis statistics and IRT criteria.
5. Compare the content balance of the resulting item bank with that of the previous test specifications, and evaluate the functioning of the system by conducting a simulation of the behavior of test-takers at various proficiency levels.
6. Convert the surviving items to computerized form in preparation for equating the P&P and CAT version of the tests.

14.4.2.2 Test Starting Point

How does a CAT start? If there is prior information about a test-taker's ability level (e.g., GPA), then the first item can be selected accordingly. For example, a difficult item may be administered to a test-taker with a high GPA. Often (if not always), no such prior information is available, so the first item is usually given by either (a) random selection of an item from the item bank or (b) random selection of an item with medium difficulty from the item bank. Actually, when the test length is long (e.g., more than 25 items), how the first item is selected does not affect the test-taker's final ability estimate (Lord 1977). When test-takers' ability levels have a mean of zero, then an item with difficulty of zero is commonly selected as the first item.

14.4.2.3 Ability Estimation

Once an item is administered to a test-taker, his or her ability can be reestimated. Two commonly used strategies for estimating a test-taker's abilities are the maximum likelihood estimation (MLE) (Lord 1980) and the Bayesian estimation (Bock and Mislevy 1982). The principal of MLE is to find the most plausible value of ability, given the test-taker's responses to administered items. The Bayesian estimation strategy considers not only the test-taker's responses to administered items but also the prior information about the distribution of test-takers. There are two major estimators in the Bayesian estimation strategy: the maximum a posteriori estimator and the expected a posteriori estimator. Descriptions of these estimation methods are available elsewhere, and the reader is referred to van der Linden and Glas (2010) for further details.

MLE has several statistical advantages. It is consistent, efficient, and asymptotically normally distributed. Also, the standard error for the estimate is readily available. However, limitations do exist. One of the serious limitations is its inability to yield estimates for test-takers who obtain all correct or all incorrect scores in dichotomous items, or, more generally, test-takers who obtain the highest score category or lowest score category for all items. This limitation is especially serious in adaptive testing. At the beginning of the testing, when the first item is administered, every test-taker either gives a correct answer or an incorrect answer. Thus, MLE is not feasible. Even when two items are administered, it is very likely that a great number of test-takers still have all correct or all incorrect scores on the two items, which means that MLE is not feasible for them.

In contrast, the Bayesian estimation can yield ability estimates for any kind of response patterns, including all correct or all incorrect scores. Thus, after the first item is administered, the Bayesian estimation is already available. This feature is especially important in adaptive testing, because at the beginning of the testing (e.g., with only one or two items administered), it is very likely that test-takers would have all correct or all incorrect scores. However, it should be noted that the efficiency of the Bayesian estimation depends on the correct specification of the distribution of test-takers (called prior distribution). If the specification of prior distributions is far from correct, then the Bayesian estimation can yield poor estimates. Furthermore, with the use of prior information, the ability estimate is biased toward the mean of the prior distribution, and the bias is especially serious for those with extreme ability levels.

Given the major advantages and disadvantages of MLE and Bayesian estimation methods presented above, the question is how to choose a method. To answer the question, there are several considerations. First, the purpose of the test should be taken into account (Keller 2000). For instance, does the CAT aim to determine the ability of test-takers or to place test-takers into categories? Is the standard error of ability estimate to be minimized or is the classification accuracy of test-takers to be maximized? Second, measurement issues of standard errors, bias, and efficiency should be considered. If we consider licensure exams as an example, the major concern is to minimize both the standard errors and bias because the testing is typically to distinguish those with mastery and those with nonmastery rather than to rank test-takers. Unfortunately, there is no best method for all conditions. Trade-offs may be involved in choosing an appropriate estimation method.

14.4.2.4 Item Selection

After the ability estimate is given, denoted as $\hat{\theta}$, the next step is to select an optimal item from the item bank to administer. The principal of item selection is to select an item that provides the maximum information about the location of θ, the true ability level. In IRT for dichotomous items, the item information function can be quantified as

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)}, \tag{14.6}$$

where $P_i(\theta)$ is the probability of being correct and $Q_i(\theta)$ is the probability of being incorrect on item i for a test-taker with ability level of θ, with $P_i'(\theta)$ as the first derivative of $P_i(\theta)$. It can be shown that the item information functions for the 1PLM, 2PLM, and 3PLM are

$$I_i(\theta) = P_i(\theta)Q_i(\theta), \tag{14.7}$$

$$I_i(\theta) = a_i^2 P_i(\theta)Q_i(\theta), \tag{14.8}$$

$$I_i(\theta) = \frac{a_i^2 \left[P_i(\theta) - c_i\right]^2}{(1-c_i)^2} \times \frac{Q_i(\theta)}{P_i(\theta)}, \tag{14.9}$$

respectively. Let us take the item information function of the 1PLM as an example. It has the maximum value of 0.25 when $P_i(\theta)=Q_i(\theta)=0.25$, which occurs when item difficulty b_i is equal to the test-taker's ability θ. In other words, an item with difficulty equal to the test-taker's ability would provide the maximum information (with an amount of 0.25) and is therefore the one that should be selected. For the 2PLM, in addition to the b_i parameter, the item information function is also affected by the a_i parameter—the larger the a_i parameter, the higher the item information. For the 3PLM, in addition to the b_i and a_i parameters, the item information function is also affected by the c_i parameter—the smaller the c_i parameter, the higher the item information.

The item information functions can be summed across items to form the test information:

$$I(\theta) = \sum_{i=1}^{L} I_i(\theta), \tag{14.10}$$

where $I(\theta)$ is the test information function and L is number of items being administered. The greater the test information function, the higher the measurement accuracy is—in other words, the smaller the standard error of measurement is. The relationship between the test information and the standard error of measurement is

$$SE(\theta) = 1/\sqrt{I(\theta)}. \tag{14.11}$$

When CAT stops, Eq. 14.11 is used to yield standard error for that particular test-taker if MLE is used for ability estimation. When Bayesian estimation method is used for ability estimation, the standard error will be slightly different from that rendered by Eq. 14.11 because the prior distribution has to be considered in the calculation of standard error (Bock and Mislevy 1982).

14.4.2.5 Test Termination

The last step in CAT is to terminate the test. In general, CAT stops when the precision of ability estimation is adequate or when a predetermined maximum number of items have been administered. In CCT, the test is terminated either when the test-taker can be classified into a category or when the maximum test length is reached (under such a case, the test-taker is forced to be classified). Following this logic, two termination rules are applied in CAT: the fixed-precision rule and the fixed-length rule.

Just as in conventional P&P testing, the fixed-length rule is to administer a fixed number of items (e.g., 30 items) to all test-takers in a CAT environment. This rule is easy to implement. However, the measurement precision (or standard error) would be different across test-takers. Often, test-takers with extreme ability levels (very high or very low) will have a larger standard error than those with medium ability levels, simply because there is no sufficient number of items with appropriate difficulty levels for those extreme ability levels.

The fixed-precision rule appears to be preferred. The test stops when the prespecified measurement precision is reached, such that all test-takers have similar degrees of measurement precision. However, different test-takers often receive different test lengths. In general, test-takers with extreme ability levels would require longer tests than those with medium ability levels because item banks often consist of more items of medium difficulty than items with extreme difficulty (i.e., very easy or very difficult). Sometimes there is no sufficient number of items at an appropriate difficulty level, and CAT cannot stop in a reasonable amount of time. In such a case, CAT has to be forced to stop at a maximum test length.

14.4.3 Operational Issues with CAT

Often, a test consists of multiple domains. For example, a mathematics test may consist of numbers, algebra, geometry, and data and chance, and a science test may consist of biology, chemistry, physics, and earth science. It is important to ensure that all test-takers receive similar distributions of domains, for example, 20% in numbers, 20 % in algebra, 30 % in geometry, and 30 % in data and chance. That is, the contents should be balanced according to some prespecified test specification. If a test-taker receives items all on numbers and another all on algebra, this diversity would cause a great challenge in score comparability.

Another important operational issue in CAT is how to maintain test security, which is especially critical in high-stakes tests, such as college entrance exams or license exams. If tests are compromised, test scores are no longer valid, and the subsequent decisions are misleading.

14.4.3.1 Content Balancing

When a test consists of multiple domains, target percentages of contents should be specified before CAT proceeds. Procedures should put into place to ensure that the target percentages are guaranteed. There are two commonly used procedures for content balancing: the constrained CAT and the modified multinomial model.

The Constrained CAT

In the constrained CAT procedure, the selection of an optimal item from an item bank is restricted to the domain with the current exposure rate the farthest below its target percentage (Kingsbury and Zara 1989). For example, assume the target percentages of four domains are 20%, 20%, 30%, and 30%, respectively. If a test-taker has received five items, of which one was from domain 1, one from domain 2, one from domain 3, and two from domain 4, then the percentages for the four domains are 20%, 20%, 20%, and 40%, respectively. Domain 3 thus has the rate farthest below its target percentage of 30%. Thus, the next item will be selected from domain 3. The major advantage of this procedure is its simplicity, while the major concern is that the sequence of domains becomes highly predictable, which may threaten test security.

The Modified Multinomial Model

The modified multinomial model was developed to resolve the problem of the highly predictable sequence in the constrained CAT procedure (Chen and Ankenmann 2004). A cumulative probability is generated based on the target percentages of the content areas that add up to 1.0. For example, if the target percentages for the four domains are 20%, 20%, 30%, and 30%, respectively, then the cumulative distribution will be 0.2, 0.4, 0.7, and 1.0. For simplicity, let the fixed-length rule be used and the maximum test length be 20 items. Of the 20 items, 4, 4, 6, and 6 items should come from the four domains, respectively. Then, a random number selected from the uniform distribution U(0,1) is employed to determine the corresponding content area in the cumulative distribution from which the next optimal item will be selected. If the random number is 0.12, then an item from domain 1 should be administered, because 0.12 falls within the range of 0 and 0.2, which belongs to domain 1. Once the first item from domain 1 is administered, there are 19 items to be administered; among them 3, 4, 6, and 6 items should be selected from the four domains, respectively. Therefore, the updated percentages for the four domains are 0.158 (=3/19), 0.211, 0.316, and 0.316, respectively, and the cumulative probabilities are 0.158, 0.368, 0.684, and 1, respectively. Then a new random number is selected from U(0,1). If it is 0.55, then an item from domain 3 will be selected, because 0.55 is within the range of 0.368 and 0.684, which belongs to domain 3. After this item is administered, there are 18 items to be administered; among them 3, 4, 5, and 6 items should be selected from the four domains. These steps repeat until all 20 items have been administered. When a domain has reached its target percentage, a new multinomial distribution is generated by adjusting the unfulfilled percentages of the remaining domains. This procedure not only guarantees the target percentage of each domain is met but also ensures an unpredictable sequence of domains and, thus, test security.

14.4.3.2 Test Security

In CAT, an item with the maximum information about the ability level will be selected for administration. Because most test-takers have medium levels of ability, items with medium difficulties will be exposed intensively, whereas those of extreme difficulty would seldom be administered. Overexposure makes it possible that some test-takers will obtain prior knowledge of a certain set of items from other test-takers who took the test beforehand. In this scenario, the test-takers' responses will no longer reflect their true ability levels. While overexposed items threaten test security, underexposed items are not cost-effective because the construction of quality items is very expensive. It is desirable that all items in an item bank have a similar chance of being administered.

There are dozens of methods for item exposure control, which can be classified into four categories (Stocking 1993; Georgiadou et al. 2007; Way 1998): (1) randomization, (2) conditional selection, (3) stratified, and (4) combined methods.

Randomization Methods

In randomization methods, items are selected based on both item information and randomness. As CAT proceeds, randomness plays a less and less important role in item selection. The 5–4–3–2–1 method (McBride and Martin 1983) and the progressive method (Revuelta and Ponsoda 1998) are two representative methods.

Conditional Selection Methods

In conditional selection methods, item exposure control parameters are employed to control item exposure. The Sympson and Hetter method (1985), the Davey and Parshall method (1995), the Stocking and Lewis multinomial method (1995), and the shadow test method (van der Linden and Veldkamp 2004) are representative methods. Often, intensive simulations prior to the actual CAT are conducted to obtain these item exposure control parameters.

Stratified Methods

In stratified methods, items are stratified into several strati according to their a_i or b_i parameters, such that the item bank is partitioned into several sub-banks. For example, in the a-stratified method (Chang and Ying 1999), items with low a_i parameters are administered at the early stage, whereas items with high a_i parameters are administered at the later stage. Other stratified methods include the a-stratified with b-blocking method (Chang et al. 2001) and the 0–1 stratification method (Chang and van der Linden 2003).

Combined Methods

In combined methods, multiple methods are incorporated to facilitate item exposure control, such as the progressive restricted method (Revuelta and Ponsoda 1998) and a combination of the *a*-stratified and the Sympson and Hetter method (Leung et al. 2002).

14.4.4 CAT Construction

Constructing CAT typically involves hardware, software, and professionals. These three elements are discussed in this section.

14.4.4.1 Hardware

Computers and computer accessories (e.g., monitors, keyboards, mice, speakers, headsets, printers) are required in constructing CBT or CAT. If online testing is desirable, then the Internet is needed. In order to achieve a standardized testing condition and to reduce bias in test scores due to differences in testing conditions, these devices should have approximately the same quality.

1. Item presentation: Item presentation can be text, image, audio or video, etc. Presentation can be in color or in black and white.
2. Test duration: Items should be presented in real time. That is, after a test-taker responds to an item, the next item should appear without delay. The response time that a test-taker spends on an item is usually recorded.
3. Input instrument: The most frequently used instruments for test-takers to input their responses are keyboards, mice, microphones, and touch screens.
4. Output equipment: Score reports are available immediately after CBT or CAT. When necessary, printing devices need to be set up.

14.4.4.2 Software

Software is needed to record, score, and monitor test-takers' responses. The computation in CAT is very intensive because after a test-taker responds to an item, the test-taker's ability should be reestimated, and another item should be selected from the item bank for the test-taker to respond to. If technical problems occur during administration, the monitoring system should notify test administrators for further action.

14.4.4.3 Professionals

1. Test developers. A group of test developers with content knowledge and expertise are responsible for item writing and test development.

2. IRT and CAT specialists. Psychometrics and measurement specialists with expertise in IRT and CAT are essential to ensure that CAT works well from a statistical and psychometrical perspective, including appropriateness of ability estimation and item selection, as well as operation issues, such as item exposure control and content balancing.
3. Information technology professionals. Information technology professionals are needed to ensure that both the computer hardware and the software work well in item presentations, test-takers' reaction recording, scoring, automatic item selection, ability estimation, networking, data management, and system maintenance, among others.
4. Test administrators. Administrative staff should be in charge of CAT administration and test quality and security.

14.5 Conclusion and Future Developments

Self-directed learners need frequent assessment to monitor their learning paces and to adjust learning strategies. CBT and CAT provide them with better tools than P&P testing because of their high quality, flexibility, and efficiency. With the development of computer technology, CBT has become more and more popular. It enables diversity in item presentation and testing environment standardization, allows test-takers to take tests anytime and anywhere that computers are connected to examination systems, and delivers score reports immediately. These properties improve test quality and flexibility substantially. CAT (or CCT) further improves test efficiency through its ability to adapt. However, as indicated in this chapter, constructing CAT requires a joint effort that combines computer hardware, software, and professionals with expertise in content matters, measurement, information technology, and administration. The biggest challenge in CAT construction is developing and updating item banks. Usually, an item bank consists of hundreds of items, and it should be updated regularly. In addition, the speed and stability of network systems also limit whether CBT and CAT are able to effectively break through the barriers of space and time.

Both CAT and CCT involve adaptation. CAT aims to estimate the ability levels of test-takers, whereas CCT aims to classify test-takers into a few categories. In both CAT and CCT, item exposure control and content balancing are two important practical issues, especially for high-stake examinations.

Most CAT is limited to MC or short answer items because computers need to complete scoring and select the next item to administer from item banks in a very short time. In recent years, there has been increasing interest in other item types, such as speaking and writing in language testing, which brings a big challenge to real-time automatic scoring with computers. Automated scoring is an application of computers in assessment and analysis of open-ended items. In addition to improving the validity of test scores, automated scoring could reduce the cost and effort involved in using human graders. One of the most well-developed automated scoring systems is the e-rater® (Burstein et al. 1998a), which uses a combination of statistical and neurolinguistic programming techniques to extract linguistic features from the essays to be graded. It identifies and extracts linguistic features from stored electronic text or speech and predicts essay scores based on features related to writing quality, including grammar, usage, mechanics, style, organization, and development (ETS 2011a). The e-rater® has been used by the ETS for automated essay scoring since 1999 and is currently embedded in Criterion, a Web-based real-time version of the system developed by the ETS Technologies. Criterion enables students to use the e-rater® engine's feedback to evaluate their essay-writing skills and to identify areas that need improvement. It also allows teachers to help students develop their writing skills independently and receive automated and constructive feedback (ETS 2011b).

In a study that compared the performance of the e-rater® for scoring essays in the Graduate Management Admission Test with that of expert readers, there was 87–94 % agreement between e-rater®'s scores and expert readers' scores on 13 different essay prompts. The e-rater® was also evaluated on two essay prompts from the Test of Written English. The e-rater® achieved agreement rates between 93 % and 94 % with expert readers (Burstein et al. 1998b). In addition to the e-rater®, there are other automated scoring technologies, including the c-rater™ system, the m-rater engine, and the SpeechRater engine. The c-rater system was developed for automatic analytic-based content scoring of short free-text responses, ranging in length from a few to approximately 100 words. The m-rater engine scores computer delivers constructed-response mathematics items for which a response is either a mathematical expression or equation or a graph. The SpeechRater engine provides automated scoring of spoken English proficiency, as demonstrated through spontaneous speaking tasks like those found on the Test of English as a Foreign Language.

Within the past 30 years, CBT and CAT have been studied extensively in educational achievement tests from both theoretical and practical perspectives. Test security is always of great concern. A promising direction is to revisit CBT and CAT in other types of assessment, including low-stakes diagnostic or self-assessment, personality and attitude assessments, and medical and clinical tests. However, as noted by Wainer (2000), there are controversial issues in the construct validity of computerized measures of personality and attitudes, because responses to personality and attitude items may be very sensitive to item ordering and context. In medical and clinical assessment, test security might not be as critical and consequential as in high-stakes achievement tests or license assessment. Instead, test efficiency could be significantly improved, which is of great benefit to patients and their proxies.

Last but not least, current operational CAT or CCT is based on IRT models. Recently, cognitive diagnostic modeling has become a new field of psychometric research in education. Cognitive diagnostic models aim to diagnose test-takers' mastery status within a group of discretely defined skills or attributes, thereby providing them with detailed information regarding their specific strengths and weaknesses, rather than a summative ability estimate as in IRT models (Junker and Sijtsma 2001; Rupp et al. 2010; and Chap. 5 this book). Such combination of cognitive diagnosis with computer adaptive assessments has emerged as an important field (Cheng 2009; McGlohen and Chang 2008).

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–563.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998a, August). *Enriching automated scoring using discourse marking*. Paper presented at the workshop on discourse relations & discourse marking conducted at the annual meeting of the Association of Computational Linguistics, Montreal, Canada.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998b). *Computer analysis of essays*. Retrieved on November 19, 2011, from http://www.ets.org/Media/Research/pdf/erater_ncmefinal.pdf

Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in a-stratified computerized adaptive testing. *Applied Psychological Measurement, 27*, 262–274.

Chang, H.-H., & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211–222.

Chang, H., Qian, J., & Ying, Z. (2001). a-Stratified multistage computerized adaptive testing with b-blocking. *Applied Psychological Measurement, 25*, 333–341.

Chen, P.-H. (2005). Computer adaptive testing theory and application. *National Elite, 1*, 57–173.

Chen, S.-Y., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41*, 149–174.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619–632.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. *Education and Psychological Measurement, 69*, 778–793.

Eggen, T. J. H. M. (2010). Three-category adaptive classification testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 373–387). New York: Springer.

ETS. (2011a). *About the e-rater® Scoring Engine*. Retrieved on November 19, 2011, from http://www.ets.org/erater/about

ETS. (2011b). *ETS automated scoring and NLP technologies*. Retrieved on November 19, 2011, from http://www.ets.org/Media/Home/pdf/AutomatedScoring.pdf

Folk, V. G., & Smith, R. L. (2002). Models for delivery of computer-based tests. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah: Lawrence Erlbaum Associates.

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*, 1–38.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams. *Applied Measurement in Education, 19*, 203–220.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Keller, L. A. (2000). *Ability estimation procedures in computerized adaptive testing* (AICPA technical report). Ewing: The American Institute of Certified Public Accountants, AICPA.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York: Academic.

Kingsbury, G. G., & Zara, A. R. (1989). Procedure for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 259–375.

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2002). Item selection in computerized adaptive testing: Improving the *a*-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement, 26*, 376–392.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367–386.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95–100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah: Lawrence Erlbaum Associates.

Luecht, R. M., & Nungester, R. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 239–249.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Pychometrika, 60*, 523–547.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 224–236). New York: Academic.

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*, 808–821.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests. Copenhagen: Institute of Educational Research.* (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311–327.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16*, 65–76.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Technical Report RR 3–2). Princeton: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995a). *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95–25). Princeton: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing.* In Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973–977). San Diego: Navy Personnel Research and Development Centre.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 118–193.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York: Springer.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273–291.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer*. Mahwah: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1–14.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wang, W.-C., & Huang, S.-Y. (2011). Computerized classification testing under the one-parameter logistic response model with ability-based guessing. *Educational and Psychological Measurement*. doi:10.1177/0013164410392372.

Wang, W.-C., & Liu, C.-W. (2011). Computerized classification testing under the generalized graded unfolding model. *Educational and Psychological Measurement, 71*, 114–128.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*(4), 17–27.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York: Springer.

Part III
Case Studies of Self-Directed Learning Oriented Assessment in the Region

Chapter 15
Learning Assessment Reform in Thailand

Somwung Pitiyanuwat and Tan Pitiyanuwat

15.1 Why Learning Assessment Matters

Concerning value judgments and the school context, there are three distinct aspects of evaluation or assessment in education (Terwilliger 1971). First is a need for value judgments concerning the merit of methods and materials used in education, that is, assessment of the entire curriculum, a specialized technique, or resource material. Second is value judgments on the merit of the personnel responsible for the education enterprise, that is, job performance of school administrators, supervisors, teachers, and others. The third aspect of assessment in education concerns value judgments about the individual student. The merit of performances, actions, and achievements by the student within the school setting can be informally judged by administrators, counselors, and other students, but the great majority of value judgments relating to students are formally made by teachers. It is primarily the teacher who facilitates and transmits the knowledge, skills, and values of the society, and it is the teacher who judges the extent to which these have been acquired and processed by the student. Therefore, value judgments in education by teachers and other stakeholders play significant roles in effective education especially for reforming teaching and learning processes.

S. Pitiyanuwat (✉)
Chairman, National Institute of National Testing Service (NIETS) and Royal Associate,
The Royal Institute, Bangkok, Thailand
e-mail: Somwung.P@Chula.ac.th

T. Pitiyanuwat
Interior Design Department, Faculty of Architecture, Kasem Bundit University,
Bangkok, Thailand
e-mail: tanpitiyanuwat@yahoo.com

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_15, © Springer Science+Business Media Dordrecht 2013

Learning assessment is the major driving force and can be viewed specifically as a way to improve students' learning habits, support their learning, and improve the learning environment. Assessment of students' learning achievement is fundamental to a teaching and learning quality assurance system. It provides an appraisal for improving the instruction and assessment designed for students to reach their full potentials in terms of ethical and moral enhancement, inspiration, and enjoyment of lifelong learning.

In this chapter, "learning assessment" will be taken to mean the kind of classroom and school assessment that informs the type of support needed to enhance student learning and determines the level of grades and certificates awarded to students, including national education standard assessment. It does not include the kinds of assessments used to satisfy the accountability demands of an external authority.

15.2 Good Practice in Learning Assessment

Good practices in learning assessment discussed in this section are derived from research by Trilling and Fadel (2009), Kallick (2000), Shepard (2000), and Terwilliger (1997). Specifically, Kallick wrote:

> "… we need assessments that are designed for learning, not assessments that are used for blaming, ranking, and certifying. That, in turn, requires deep shifts of attitudes about testing and learning for parents, educators, and students themselves…." (Kallick 2000)

Summarizing from these previous studies, good practices in learning assessment should have the following qualities:

1. Learning assessment should be conducted on the principle of assessment *for* learning, *as* learning, and *of* learning.
2. Learning assessment should match the curriculum, that is, a student-centered curriculum, and clearly depict what we expect of our students.
3. Learning assessment must reinforce and model the approach to problems that we expect from students.
4. Learning assessment must look at more than just knowledge content. It should cover the learning processes and outcomes, as stated in the National Qualification Framework.
5. Learning assessment must provide explicit feedback on students' performance within the desired model. It should be an ongoing process and integrated with instruction.
6. Learning assessment must provide some form of tracking of students' development as they progress through the course and the curriculum.
7. Learning assessment must assess teaching as well as student learning.
8. Learning assessment must employ multiple methods for assessing performance on standards. Alternative assessment procedures should be adopted in combination with more traditional forms of assessment as new evidence of the educational and

psychometric value of such alternatives becomes available. (Terwilliger 1997; Tangdhanakanond et al. 2005; Jankarn 1987)

9. Learning assessment must encourage students to be active in assessing their own work.
10. Learning assessment must provide information to teachers, students, and all stakeholders.

15.3 Selected Research in Learning Assessment

Kaovichit (1981) conducted research in learning assessment relevant to discussion in this chapter. The purpose of the research was to determine the effects of three learning evaluation systems upon the learning achievement in mathematics of upper secondary school students. The three evaluation systems used were norm- and criterion-referenced, criterion-referenced, and norm-referenced. Using an experimental design, the sample consisted of 3 classes each with 30 students, and the students' past mathematics achievement and achievement motivation did not vary significantly across the three classes. Their mathematics learning achievement scores, measured by achievement tests, were analyzed by means of one way analysis of variance and Newman-Keuls tests.

It was found that (Kaovichit 1981):

1. Students learned under the norm- and criterion-referenced learning evaluation system had higher mathematics learning achievement (71.70%) than students learned under the norm-referenced learning evaluation system (56.97%) or the criterion-referenced learning evaluation system (66.33 %).
2. The criterion-referenced learning evaluation system (66.33 %) yielded higher mathematics learning achievement than under the norm-referenced learning evaluation system (56.97 %).

Interesting research concerning alternative assessment, or sometimes labeled as authentic assessment, was carried out by Tangdhanakanond et al. (2006). Constructionism as an educational concept asserts that students are more likely to construct knowledge and form new ideas when they engage in building tangible objects. Such objects are often products of a group project. Learning under constructionism is therefore project-based.

Twelve students from Darunsikkhalai School served as sample for the study reported by Tangdhanakanond et al. (2006). Darunsikkhalai School in Bangkok, Thailand, provides total project-based education for its students. Students' portfolios are used to assess students' academic and nonacademic development. Their portfolios were assessed 3 times during a 9-week project period. The results indicated significant improvement in both academic and nonacademic outcomes. It was also found that academic gain was larger than nonacademic gain and that gain from the second to the third assessment was larger than the gain from the first to the second assessment.

The above research findings (Tangdhanakanond et al. 2006) give empirical support to the application of criterion- and norm-referenced assessment. In the study, such assessment is conducive to student learning and optimally facilitates students' learning achievement. Further, the study demonstrated that alternative or authentic assessment procedures in combination with traditional forms of assessment contribute to standard-based assessment.

15.4 Learning Assessment: Looking Backwards

The purpose of this section is to trace the revolution and development of learning assessment in Thailand over a period of 126 years of public education and learning assessment. Learning assessment indeed has a deep root in the history of Thai education, as highlighted by Pitiyanuwat and Sukamolson (1985) and Suwankul (1975). In 1884, two significant historical phenomena took place. One was the establishment of the first public school, Wat Mahunparam School, and second, on March 27, the first examination was conducted for the purpose of certification.

Since then, the development of learning assessment in Thailand can be categorized into six eras, the first covering times prior to 1884:

15.4.1 The First Era "The Preschool Period (1283–1883)"

The year of 1283 stands out in Thai education as the year that King Ramkhamhaeng the Great invented Thai alphabets. These greatly facilitated the process of Thai education and have given all Thais great pride in their identity for having a language of their own. Learning in the Sukhothai period focused mainly on reading, writing, career training, and moral practice. Due to a very limited number of students, the teacher could get to know his students well and was aware of different abilities in all subjects. Thus, he could subjectively judge their ability levels and certify their fulfillment according to his own criteria; in modern parlance he carried out teacher-referenced assessment. Learning assessment testing techniques used in this period were oral tests.

Later on, in both the Ayutthaya and early Rattanakosin periods, the subjects of study were spelling and writing with emphasis on grammar. Normally, if students could not write by themselves their teacher would help them by means of guiding their hands with his. At higher levels, the students would learn mathematics and translation of "Tripodike." Learning assessment was dependent on the teacher's consideration on his students' ability. The teacher would independently judge which students should advance to the higher grade or those who should repeat the same grade (Chongkol 1984). In short, learning assessment was carried out regularly by the teacher and was completely by means of an oral examination.

15.4.2 The Second Era "The First Official (Noncurriculum) School Period (1884–1891)" or the "Premodernization Period"

King Rama V the Great's visionary idea was that the way to develop the country and to make it more civilized was to provide educational opportunity to all people. Thus, he established the first official, but noncurriculum, school in 1871 within the palace compound as a model school. Under the first school principal, Luang-sara-pradit who was later entitled Praya-srisunthorn-wohan (Noi Arjariyang-koon), Thai, foreign languages, and some others were the subjects being taught. The teachers were common people, not monks, and the purpose of this school was to educate military officers and civilian officials.

As previously mentioned, Wat Mahunpararm School for common people was established in 1884. That year became the first time that the content of study was divided into military and civilian fields, following a common period of studying the fundamental subjects. However, since only a few students could complete the whole "6-text set" and additional qualified graduates were needed, those who had not completed the whole set needed to be categorized in some way.

Consequently, some form of examination was called for to test the different levels of the students' capability and to convey this information to appropriate government sectors. Thus, the first examination covered the subject matter in the "6-text set," and those who passed got the first certificate. In the same year, there was a second examination for the second certificate which consisted of the following eight subjects:

1. Handwriting, both in fast and slow styles
2. Formal writing
3. Text editing based on fast handwriting
4. Text copying and passage summarizing
5. Letter writing
6. Prose composing and correcting
7. Mathematics
8. Accountancy

These examinations were administered once a year. In order to promote education and make examination procedure publicly recognized, an Examination Act was issued in 1890. From then, the examination was given twice a year, in October and March.

In summary, learning assessment in the second era began to develop its own principles and regulations for testing and assessment more formally. From an oral examination given by the teacher, it was then administered by central officials; in 1884, an essay test was introduced as the examination paper, and a year later, it was taught in classes. In 1890, an Examination Act was issued which significantly emphasized and recognized the examination system. Thus, examination has been central from the beginning of the Thai education system, even before the introduction of any formal Thai curriculum.

15.4.3 The Third Era "The Formal School Curriculum During the Absolute Monarchy Period (1892–1932)"

This was also regarded as the "initial stage of the modernization" period. Under King Rama V the Great, this period was marked by the establishment of a national curriculum (1892) and a secular education system wherein schools were separated from temples (1898). Between 1892 and 1932, eight curricular subjects for primary and secondary schools were developed. As regards assessment of students' learning achievement, seven types of examination (Chongkol 1984) were developed:

1. School examinations based on a normal curriculum
2. Competition examinations for grants and certificates issued by the Ministry of Education
3. Competition examinations for king scholarships to study abroad
4. Competition examinations for being civilian officials
5. Teaching certificate examinations for being school teachers
6. Examinations for the monks
7. Any other special examination for being admitted into a college and a special school

Between 1913 and 1921, a new curriculum was developed. For assessment, examinations were administered only for those who were to finish their primary or secondary education. For primary education, any occupational subjects could be tested in school; all other examinations came under the responsibility of the Ministry of Education. To finish primary education, students had to pass both general and occupational subjects. In 1928, an upper secondary education curriculum was implemented. Thai language was required for all students, and failure in Thai would lead to failure in secondary education as failed students were not allowed to sit other subject examinations.

15.4.4 The Fourth Era (1933–1977)

It is also known as the period of "modernization" and presented a change from an absolute monarchy to a constitutional monarchy. The government in the period of modernization advocated universalization of education, and the curriculum emphasized responsibility towards country, society, family, and oneself. Based on the first National Scheme of 1951, school curriculum was developed to reflect the four H Principles: head, heart, health, and hand or, in other words, intellectual, moral, physical, and practical education. Between the years 1933 and 1977, six school curricula were implemented. In 1960, a lower and an upper primary school curriculum were implemented along with the ones for lower and upper secondary. Amendments to these broad field curricula were the outcome of the Chachoengsao Project (Pipyajan 1958). For learning assessment purposes, at lower and upper primary levels, points were allocated to the domains of student character development, year-round class work or assignments, and final examination. The first two domains were executed by the schools, while the third by district (Amphur) and province (Changwat) authorities. At the lower and upper secondary levels, only the last two domains remained under the school's authority.

During the period of 1969–1975, learning assessment was norm-referenced. The teacher's handbook stated that the standard score and T-score would be used. In 1975, the percentage system was changed to a grade-point system. Learning assessment at the upper secondary level was the responsibility of each school and the school district, with a suggestion that the cognitive, affective, and psychomotor domains were to be assessed.

15.4.5 The Fifth Era or Period of "Postmodernization" (1978–1998)

The fifth era (1978–1998) presented the new National Scheme of Education, which changed the education system from 7–5–4 to 6–3–3 years and introduced new curricula in both primary and secondary schools. In conjunction with the curricula, regulations for the assessment of students' learning achievement were enforced.

At primary level, the school was now responsible for assessment of students' learning achievement across all grades. Both formative assessment and summative assessment were suggested. The teacher was responsible for assessing the prior behaviors of the students, and while giving instruction, she/he was expected to assess student achievement and ability in specified instructional objectives. Measuring tools were chosen on the basis of content coverage and the congruency of the given instructional objectives. Remedial teaching measures were conducted where weaknesses were identified.

In conclusion, the concept of automatic promotion was fully implemented in the new primary education curriculum, whereby schools and teachers play significant roles in assessing students' learning achievement. Regulation in this period with regard to assessment of students' learning achievement indicated that criterion- or objective-referenced assessment was being implemented.

At the lower secondary level, assessment of students' learning achievement followed the same principle as the primary curriculum. Differences seemed to exist only in the method of assessment of students' learning achievement. A total score should consist of formative scores and summative scores. Formative scores should consist of scores from quizzes, assignments, work styles and habits, and the development of students' attitudes, interests, and/or personality. Finally, a summative score should be the final examination score which reflects the degree in which the students achieve the essential instructional objectives.

At the upper secondary level, regulation concerning the assessment of student learning achievement was essentially the same as the one used in 1975. That is, schools were responsible for assessing student learning achievement, and objective-referenced assessment was being adopted as two types of assessment: ongoing school term assessment and final school term assessment.

15.4.6 The Sixth Era or the "Modernized and Developed Education Period"

This begins in 1999 and continues to date. Based on the 1999 National Education Act, Section 6 stipulates that education aims to achieve the full development of the Thai people in all aspects: physical and mental health; intellect; knowledge; morality; integrity; and desirable living leading to a life in harmony with other people and as a national and global citizen (Office of the National Education Commission 2001).

The current Thai formal education system comprises two levels: basic education and higher education. Basic education is divided into three levels. They are preschool, primary education, and secondary education levels. The basic education cycle covers 12 years of core student-centered curriculum, divided into four levels (see below), and the 2-year preschool curriculum is separated from the basic education curriculum.

- First level – primary education grades 1–3
- Second level – primary education grades 4–6
- Third level – secondary education grades 1–3
- Fourth level – secondary education grades 4–6

While each grade level has the same goals and objectives, each develops a different emphasis. Overall, the substance consists of a body of knowledge, skills or learning processes, values or virtues, morality, and correct behavior. This substance is assembled into eight subject learning groups: Thai language, mathematics, science, social studies, religion and culture, health and physical education, art, career and technology, and foreign languages. Higher education is divided into two levels: predegree and degree level.

As indicated in Section 22 of the 1999 National Education Act, education is based on the principle that all learners are capable of learning and self-development. The teaching-learning process aims at enabling learners to develop at their own pace and to maximize their potential. The concept of learner-centered learning has been generally accepted in the teaching-learning process to facilitate learner development at various stages, and to provide a learning environment that allows for freedom, relaxation, and enjoyment, so that a child's intellect can be developed to its full potential. Considerable efforts have been made to reform the teaching-learning process, including a shift from teacher-centered to more learner-centered approach. The development of new learning media, technologies, and the training of teacher are promoted.

Furthermore, the 1999 National Education Act, Section 26 stipulates that "education institutions shall assess learners' performance through observation of their development; personal conduct; learning behavior; participation in activities and results of the tests accompanying the teaching-learning process commensurate with the different levels and types of education. In addition, educational institutions shall use a variety of methods for providing opportunities for further education and shall also take into consideration results of the assessment of the learners' performance referred to the first paragraph" (Office of the National Education Commission 2001).

15.5 Contemporary General Assessment at the School and National Level in Thailand

At basic education level, learning content and standards are applied as criteria to determine the quality of learners after graduation. Each subject group has a standard according to its substance. At the school level, classroom assessment in each subject group is conducted to assess whether learners have actively gained knowledge and skills and whether moral behavior and desirable values have been instilled. Learning assessment is conducted by schools to check learning advancement in each class, grade level, and year. Schools can stipulate assessment principles and criteria with approval of the school committee.

At the national level, national learning assessment is carried out by the Ordinary National Education Tests (O-NET) in each subject group implemented by the NIETS (National Institute of Educational Testing Service) at the end of each terminal grade, that is, primary grade 3, primary grade 6, secondary grade 3, and secondary grade 6. In addition to O-NET, NIETS administers the General Aptitude Test (GAT) and the Professional and Academic Aptitude Test (PAT) to secondary grade 6 learners. National test results are normally required for entry to universities.

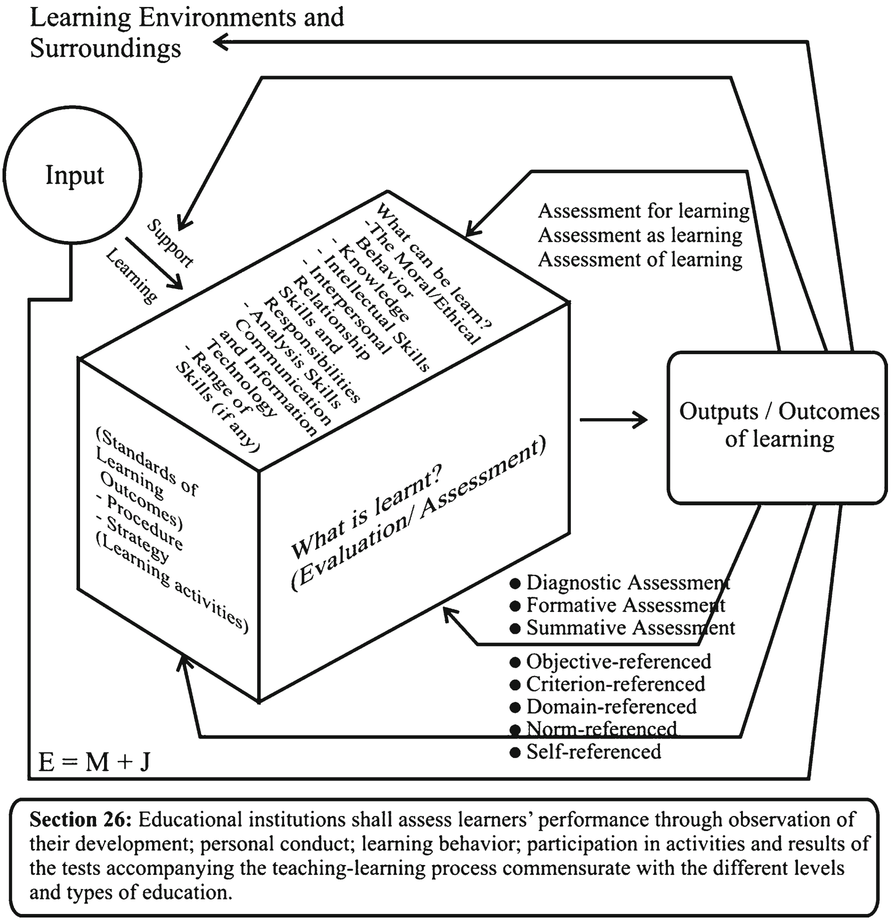
In conclusion, the present standards-based curricula place more emphasis on students' cognitive and noncognitive development as indicated in the 1999 National Education Act. To be responsive to criterion- or objective-referenced assessment, an imported and advanced technology is being enforced nationwide. However, the shift from norm-referenced assessment, with teacher-centered curricula, to criterion-referenced assessment, with the student-centered curricula, has not materialized. It is hoped that formative assessment and effective remedial teaching will help the students to progress. Automatic promotion, as in the fourth era, is still fully implemented at the basic education level with the Ministry of Education delegating power to the schools to be fully responsible for learning assessment and decision-making regarding assessments of students' learning achievement. The objective tests, specifically multiple choice tests, are very popular in students-based learning assessment at classroom, school, and national levels.

15.6 Learning Assessment: Looking Forward

Some issues and future trends of learning assessment in Thailand are highlighted below.

15.6.1 A Holistic Learning Assessment Framework

According to Dr. Kowit Worapipat (2000), the former Permanent Secretary General of Ministry of Education, who said, "What and how students learn and how teachers teach largely depends on how to assess students' learning," reform of learning



Section 26: Educational institutions shall assess learners' performance through observation of their development; personal conduct; learning behavior; participation in activities and results of the tests accompanying the teaching-learning process commensurate with the different levels and types of education.

Fig. 15.1 Learning Assessment Framework in accordance to Section 26 of the 1999 National Education Act

assessment requires assessing all dimensions of students' learning, not just knowledge and content of subject matter and the adoption and implementation of Section 26 of the 1999 National Education Act. A learning assessment framework is shown in Fig. 15.1.

15.6.2 Multiple Assessment Methods

In order to match learning assessment to the student-centered curriculum, we have to combine more tradition forms of assessment with alternative assessment techniques. Multiple assessment methods may include evaluations of student portfolios of project work, classroom observations and performance rubrics, online quizzes

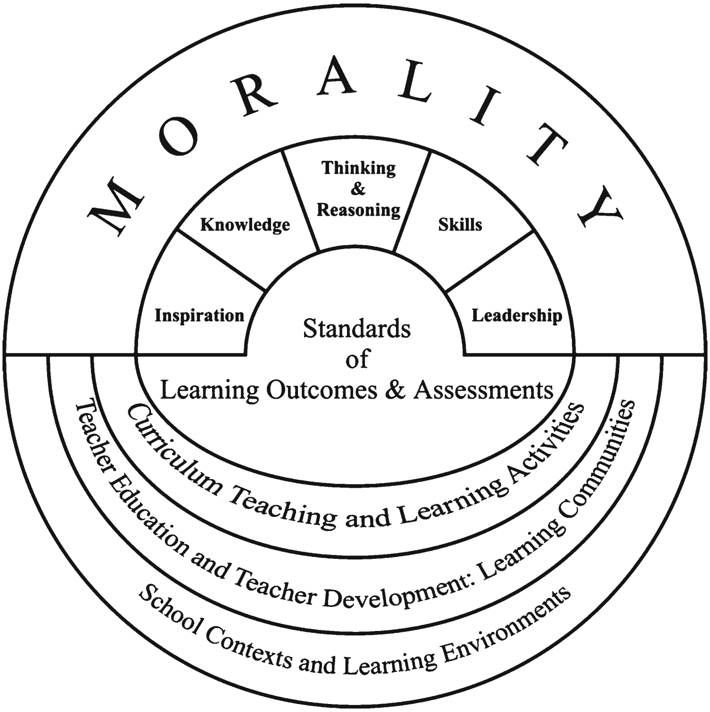

MORALITY
Thinking & Reasoning
Knowledge
Skills
Inspiration
Leadership
Standards of Learning Outcomes & Assessments
Curriculum Teaching and Learning Activities
Teacher Education and Teacher Development: Learning Communities
School Contexts and Learning Environments

Fig. 15.2 Learning outcomes and assessment framework for Thai basic education

and simulation-based assessments, juried presentations, and juried exhibits or performances (Trilling and Fadel 2009).

15.6.3 Learning Outcomes and Assessment Framework

The Basic Education Commission, Ministry of Education, should encourage and support the development of the Thailand Qualifications Framework (TQF) for the basic education level. From Fig. 15.2, learning outcomes consist of six domains as follows: M (I + K + T + S + L)

1. Morality, ethical, and moral development
2. Inspiration and imagination
3. Knowledge
4. Thinking and reasoning
5. Skills: numerical, communication, IT, interpersonal, career, and life
6. Leadership

With a TQF, it should be possible to design the standards and learning assessments with curriculum and instruction, including teacher education and teacher development

15.6.4 Reforming Principles for the Ordinary National Education Test (O-Net)

The Basic Education Commission and former Education Minister Dr. Wichit Srisa-arn have agreed to use the Ordinary National Education Test (O-Net) scores of high school students together with grade-point average (GPA) to determine whether a student is eligible to graduate. Consequently, O-Net score and the GPA have significant implications for high school graduation.

High school graduates under the new guidelines will be evaluated based on their O-Net score, which will help to certify the quality of each student's education nationwide. Evaluating students under the new agreement will be 70 % based on their GPA and 30 % on their O-Net score.

O-Net is a standardized test, which aims to be a comprehensive evaluation of a student's learning achievement. In order to achieve this and measure a student's success in high school, the test must cover the entire curriculum of basic education, which is eight subject areas, and also cover seven national standards of learners as the following (Pitiyanuwat 2007):

- Standard 1: The students have integrity, moral conduct, and beneficial values.
- Standard 2: The students have ability to think both analytically and synthetically, discursive thinking, creativity, and vision.
- Standard 3: The students demonstrate essential knowledge and skills of the curricula.
- Standard 4: The students have self-initiated inquiry and love of lifelong learning.
- Standard 5: The students have a positive attitude towards work, have skills to work independently and cooperatively with others, and value ethics in the workplace.
- Standard 6: The students have good sanitary habits and good physical and psychological health.
- Standard 7: The students develop aesthetics and physical fitness through appreciation of the fine arts, music, and activities.

The current test does not cover all eight subject areas and the seven national standard of learners, and for this reason O-Net cannot be fairly used in evaluating potential graduates in both secondary grade 3 and secondary grade 6. In fact, only five subject areas were part of the O-Net test in 2005 and 2006. The many subject areas to consider in revising the test were discussed by the University Presidents Council. The committee suggested O-Net scores would be most useful to universities if the test covered eight subjects: Thai, social studies, foreign language, mathematics, science, health, arts, and technology. One of the primary functions of the O-Net score is its usefulness in evaluating prospective students applying to universities.

Following this agreement on using the test, the next step, which urgently needs to be taken by the Ministry of Education (MOE), is to propose reform principles for the O-Net exams effective for secondary grade 6 graduates starting in 2011 and for secondary grade 3 graduates starting in 2012. This will require a committee be formed of all relevant organizations and a plan setup to implement the new policy.

The MOE should also prepare to present these changes in policy in a clear and effective way to the public.

Implementing this policy will encourage the management of schools to hold their students to a higher educational standard. It will also allow Thailand to better certify the quality of its graduates, which enhances Thai basic education without adding a single new employee.

15.7 Conclusion

Assessment is the major aspect of education and learning in Thailand. Learning assessment can be viewed as a part of instruction and used to support and enhance learning. The concept of assessment has been gradually changed from "to prove" to "to improve." Learning assessment should be conducted on the principle of assessment *for* learning, *as* learning, and *of* learning. In the first part of this chapter, the significance of learning assessment is described, and good practice in assessment is presented in the second section. The third part of this chapter concerns research findings on the effects of learning assessment systems and learning assessed by student portfolios. The development of learning assessment in Thai education is shown in the fourth section cataloging it over six developmental eras: the preschool period, the period of premodernization, the initial stage of the modernization, the period of modernization, the period of postmodernization, and the period of modernized and developed education. Finally, some issues and selected future trends of learning assessment in Thailand are explored and suggested.

References

Chongkol, S. (1984). *A century of the Thai measurement*. A paper presented at the conference on a century of Educational Evaluation in Thailand, Faculty of Education Chulalongkorn University.

Jankarn, T. (1987). *A study of the quality of MEQ test for measuring mathematics problem solving ability*. Master thesis, Chulalongkorn University Graduate School.

Kallick, B. (2000). Assessment as learning. In P. M. Senge, N. H. Cambron-McCabe, T. Lucas, B. Smith, J. Dutton, & A. Kleiner (Eds.), *Schools that learn: A fifth discipline fieldbook for educators, parents, and everyone who cares about education*. New York: Doubleday-Currency.

Kaovichit, S. (1981). *Effects of learning evaluation system upon the learning achievement in mathematics of the upper secondary school students*. Master Thesis. Chulalongkorn University Graduate School.

ONEC. (2001). National Education Act B.E.2542 (1999). Bangkok: Prig wan Graphic Co, Ltd.

Pipyajan, S. (1958). Educational Improvement Project, wittayajarn, volume 12.

Pitiyanuwat, S. (2007). School assessment in Thailand: Roles and achievement of ONESQA. *Educational Research Policy and Practice, 6*, 261–279.

Pitiyanuwat, S., & Sukamolson, S. (1985). In P. Samphanpanich (Ed.), *A century of educational evaluation in Thailand: The state of the art*. Bangkok: The Khurasastra-Sampan Association.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Suwankul, L. (1975). *The development of elementary and secondary school curriculum in Thailand*. Master Thesis, Chulalongkorn University Graduate School.

Tangdhanakanond, K., Pitiyanuwat, S., & Archwamety, T. (2005). Constructionism: Student learning and development. *Academic Exchange, 9*(3), 259–266.

Tangdhanakanond, K., Pitiyanuwat, S., & Archwamety, T. (2006). Assessment of achievement and personal qualities under constructionist learning environment. *Education, 126*(3), 495–503.

Terwilliger, J. S. (1971). *Assigning grades to students*. Glenview: Scott, Foresman and Company.

Terwilliger, J. (1997). Semantics, psychometrics, and assessment reform: A close look at “authentic” assessments. *Educational Researcher, 26*(8), 24–27.

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. San Francisco: Wiley.

Worapipat, K. (2000). The Last words (in Thai & Unpublished).

Chapter 16
Concerns of Student Teachers: Identifying Emerging Themes Through Self-Assessment

Pauline Swee Choo Goh and Bobbie Matthews

16.1 Background to the Study

Malaysia, like countries such as Australia, New Zealand, Germany and the United Kingdom, holds the traditional view that a university education should provide an in-depth study of a particular discipline that equips graduates with appropriate higher-order thinking skills. The Education Commission in Hong Kong, in a report entitled Learning for Life, learning through life: Reform proposals for the education system in Hong Kong (2000), criticized this approach and said that in universities that followed the British system of education, students often had very little experience outside their specialized area of study. This form of education is contrary to the expectations of a lifelong learning society and posed a serious challenge to the Hong Kong education system (Kember and Leung 2005).

There has been similar criticism of higher education institutions in Malaysia of the need to achieve greater integration between university learning and learning in the 'real world'. Teacher education institutions have not been immune to this demand; but at the very least, the teaching practice (or more popularly known as the practicum) holds the promise that this may still be achieved.

Malaysian student teachers undergoing the practicum gain first-hand experience and knowledge about the students and the school environment, and these can provide a frame of reference for future teaching skills they are building. Teacher educators have argued that campus-based programmes do not duplicate the real school situation, and as such the practicum is considered the most important part of a teacher education programme. The practicum denotes the entire processes and

P.S.C. Goh (✉)
Faculty of Education and Human Development, Universiti Pendidikan Sultan Idris (Sultan Idris Education University), Tanjong Malim, Perak Darul Ridzuan 35900, Malaysia
e-mail: goh.sc@fppm.upsi.edu.my

B. Matthews
School of Nursing and Midwifery, Flinders University, Bedford Park, Australia

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_16, © Springer Science+Business Media Dordrecht 2013

range of actual school teaching experiences, including (a) training in critical thinking, (b) self-managed or lifelong learning, (c) adaptability, (d) problem solving, (e) communication with staff and other students and (f) the ability to work in groups and develop interpersonal relationships (Kember and Leung 2005).

Price (1987) suggested that the practicum is also an opportunity for student teachers to apply those theoretical principles previously gained during campus learning. Schön (1996) viewed the constructions of teaching practice as the core of any teacher education curriculum, and he suggested that it is during practice that student teachers link the theoretical and practical components of a programme. It appeared that student teachers are naturally optimistic about the issues of teaching until they are confronted with the reality and shock as they begin teaching (Hoy and Woolfolk 1990). The role of a practicum experience is, therefore, important not only to provide the 'lived-in' reality but also to raise concerns which can be used to prompt the investigation between related theory, knowledge and practice (Schön 1996).

Researchers have tried to understand what it is that turns experience into learning and enables a student teacher in practicum to gain the maximum benefit from the situations they are in. Authors such as Boud et al. (1985) have identified that it is crucial that individuals be given the opportunity to reflect on experiences in the light of their current knowledge and understanding. This is the position which the study has taken. The notion is that Malaysian student teachers involved in practicum would begin to think of themselves as professional teachers and use the practicum periods to reflect upon real experiences and concerns. The purpose of this study is to allow Malaysian student teachers to self-assess their thoughts and feelings, hindrances experienced, and fears and worries that plagued their experiences in practicum through a process of reflective thinking. The more that is known about the concerns faced by student teachers during their practicum, the greater the possibility of reducing stress, improving their success and maximizing the benefits of the practicum for them. An intention of the study is that it would inform current and future teaching practice in Malaysia.

16.2 Concerns of Practicum Students

Lock (1977) suggested that the types of concerns student teachers encountered should be given more attention to enable better preparation of new teachers and that the study of problems faced by student teachers was warranted. There was a better chance of eliminating problems encountered by student teachers if more was known about the difficulties they faced and the source of their concerns. On the other hand, Briggs and Richardson (1992) cautioned that the many problems faced by student teachers during their practicum could possibly have been an omen of future conflicts. Similarly, Chan and Leung (1998) advocated that it was necessary to focus on the concerns expressed by student teachers during teaching practice as areas of importance for future development in teacher education.

More recently, studies on teaching practice have given focus to the challenges faced by student teachers and how they might have affected various aspects of teacher education. For example, Smith and Lev-Ari (2005) highlighted the theory to practice and overall school context concerns faced by student teachers and how they successfully managed to gain invaluable experiences during their practicum. Yourn (2000) cautioned that the concerns faced by beginning teachers are real and these concerns do have the ability to limit and frustrate their already complex teaching situation. These issues need to be addressed at the institutional level. From a Malaysian context, Ong et al. (2004) discovered that pressures felt during student teachers' practicum prevented them from positively engaging in theory and practice. Student teachers could have been overwhelmed by the numerous realities of the classroom, students' expectations of spoon-feeding and the challenges of mixed-ability classes (Kabilan and Raja Ida 2008). Although the practicum serves as a bridge that would have provided the student teachers with the experience to develop their own personal competence and professional identity as teachers, the practicum experience is also fraught with difficulties and concerns which might have influenced the development of student teachers.

16.3 Theoretical Framework: Self-Assessment Through Reflective Practices

Wolf and Siu-Runyan (1996) stated. "Reflection is what allows us to learn from our experiences: it is an assessment of where we have been and where we want to go next" (p. 36). Reflection is a process that permits the evaluation of actions taken. The utilization of reflection encourages individuals to consider what they have done in particular situations and plan subsequent activities. Reflection is an expected undertaking in many Western countries and is now considered to be a critical part of self-development in the practicum experience involved in the training of Malaysian student teachers. It is, therefore, a strategy that encourages self-assessment by a reflective practitioner (Schön 1996). Introduced by Schön, 'reflective practice' is often used in education pedagogy. Reflective practice allows individuals to step back and deliberate on their own thoughts and actions. It is a conscious self-appraisal that reviews events that have occurred and gives meaning to feelings, thoughts and actions by questioning motives and attitudes (Dewey 1933).

Schön (1996) succinctly explained that reflective practitioners would thoughtfully consider their own experiences in applying knowledge to practice while being assisted by mentors in the discipline. In addition, reflective practitioners would also be self-directed towards a deeper understanding of their own teaching styles and ultimately achieve greater effectiveness as a teacher. It is about allowing the individuals to 'recapture their experience, think about it, mull it over and evaluate it' (Boud et al. 1985, p.19). Schön advocated that it was not only useful for student teachers on teaching practice but for all teachers to reflect on their classes in terms of class management, content and teaching and learning strategies to improve themselves and enable the transference of knowledge to their students.

16.4 Method

16.4.1 The Participants and the Context

From a total of 16 student teachers invited to participate in the study, 14 accepted: all are female. These student teachers are undertaking the Bachelor of Education in Science degree from a teacher education university in Malaysia. The degree prepares them to teach in secondary schools. Student teachers are required to attend courses from both the Faculty of Science and Technology for subject content and the Faculty of Education and Human Development for general education subjects such as teaching models, methodology and strategy, assessment and classroom and organizational skills. They have a 2-week school orientation programme spread over the seven semesters of their education. Practicum occurs in the eighth semester where the student teachers are placed in selected secondary schools for 14 weeks. Each student teacher is assigned a university-based supervisor and a school-based mentor who are experienced teachers to support and guide them during their practicum. At the time of the study, English is the language of instruction in the teaching of science and mathematics. However, the language of instruction in the teaching of science and mathematics will revert to *Bahasa Melayu* (Malay Language) from the year 2012.

16.4.2 Data Collection and Procedure

This study is based on the intention to reproduce a ‘lived-in’ reality for the participants. It assumes that the participants are inextricably related to the contexts in which they ‘experience, conceptualize, perceive and understand various aspects of, and phenomena in the world around them’ (Martön 1986, p.31) and that a qualitative approach best allows the researchers to share and experience their realities. An approach using the participants’ capacity for reflective practice allows them to determine what an experience means to them as they interact with their social realities (in this case, the schools, the students and other teachers with whom they interact). To achieve this purpose, reflective practice written as journals is used to capture experiences and thoughts of the participants during their practicum.

Journal writing encourages participants to record their thinking through narration and so ‘by writing about experiences, actions and events, student teachers will reflect on and learn from those episodes’ (Loughran 1996, p. 8). Further, it can ‘clarify and extend individual thoughts and concerns and provide supervisors with a means of consistently supporting interns’ inquiry into their development as learners and teachers’ (Collier 1999, p. 174). It is a way to stimulate reflective thinking and provides one of the best methods for participants to self-direct and assess their own teaching-learning issues (Zeichner and Liston 1987). Hall and Bowman (1989) have found that through a reflective journal, participants are able to reflect on socialization and professional growth issues that they would not normally be aware of. Previous studies (e.g. Yinger and Clark 1981; Hatton and Smith 1995) advocated the use of reflective journal writing as a technique that can promote and document reflective thinking.

Each of the 14 participating student teachers attended a pre-practicum briefing on a one-to-one basis with the first author to discuss what their participation would entail. The participants were asked to maintain a reflective journal throughout their practicum to document their teaching experiences, concerns and their confidence to teach. There were no fixed number of entries, but the participating students were advised to write as often as they felt necessary. Some guiding questions to assist the participants in the reflection process included:

- What are you reflecting on? You don't necessarily have to reflect on the entirety of something. You can choose certain aspects. For example, a single lesson, the school environment, staff meetings or meetings with your school-based mentor.
- Give a description of the circumstances, situation or issues related to what has been selected: *Who* was involved? *What* were the concerns, issues or worries? *When* did the event happen?
- Self-assessment occurs at this stage as you interpret the activity or evidence and evaluate its appropriateness and impact. Self-assess your experience, your piece of evidence or the activity.

Upon completion of the practicum, the participants visited the first author to submit their written journals and for those who were unable to do so, sent their journals through the post. Each of the 14 journals received was given a code name.

16.4.3 Analysis

The analysis of the student teachers' reflective journal consisted of a series of steps:

- Step 1: The journals were read and reread using a method of 'free' and 'open' coding to find common themes that emerged which pertained to student teachers' concerns experienced during their practicum.
- Step 2: A more careful analysis was conducted where each text was compared using an iterative reading and rereading to establish similarities and differences in the written documents. 'Chunks' of text with similar or different themes were highlighted with pens of different colours.
- Step 3: Highlighted texts were then retyped into separate documents, representing emerging themes. Each document was read in totality to obtain a 'picture' that was written by the student teachers. Each theme was again divided into different derived concerns. Specific comments were sought to provide quotations that represented each derived concerns. Those data that were written in *Bahasa Melayu* (Malay language) were translated as closely as possible into English so that the original intention of the writer was not lost.

Table 16.1 Themes and derived concerns from student teachers' journals

Major themes	Derived concerns
Institutional and personal adjustments	Adjustments to the role as teachers
	Meeting expectations of school-based mentor
	Impressing school-based mentor
	Working harmoniously with the school staff
Classroom management and discipline	Classroom management
	Students' discipline issues
	Students' behavioural problems
Methods and strategies	Appropriate use of teaching methodology and strategies
	Organization of teaching activities
	Using English as the medium of instruction
	Mastery of the subject matter
	Teaching other subjects
	Availability of adequate or appropriate teaching aids
	Adequate time to cover the curriculum
Student achievement	Students' understanding of the subject matter
	Students' affective, emotional and social growth
	Attracting student's interest and attention
	Effect change in students' behaviour

A total of four themes were identified: (a) institutional and personal adjustments, (b) classroom management and discipline, (c) methods and strategies and (d) student achievement. Table 16.1 showed each major theme that was further divided into three to seven derived concerns. Figure 16.1 showed a tabulation of the occurrence of each of the derived concerns in the teachers' reflections.

In Fig. 16.1, the bar graphs depict the number of occurrences for each of the 18 derived concerns listed in Table 16.1.

16.5 Results and Discussion

16.5.1 Institutional and Personal Adjustments

Many participants were concerned about their transition from being a student teacher to being a teacher. The journals indicated that there were adjustment concerns which were either of an institutional type or of a personal nature. Institutional adjustments were centred on their adaptation to the norms of the school and their relationship with other teachers in the school. They were worried about adjusting to the school environment indicated by comments such as '… not being able to uphold my responsibilities well', 'being accepted by the other teachers' or about 'the school environment and if the other teachers could help'. Hayes (2003) described it as an 'anticipatory emotion' prior to a school placement. Participants' emotions were those of excitement and enthusiasm but threaded with both agitation and doubt. The feeling of fear of the unknown and uncertainty generated both feelings of excitement and anxiety.

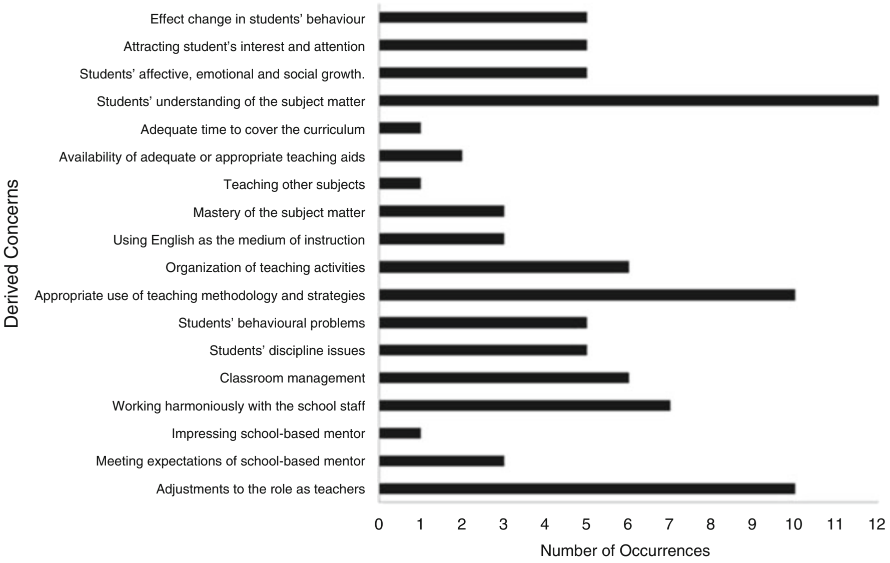


Fig. 16.1 Number of occurrences for each derived concern from student teachers' journals

Personal adjustments were emotional concerns about the perceptions of the school staff to them as trainee teachers and the acceptance by students of them as teachers. They were concerned about their adequacy and competency as teachers. Zaitun wrote that she was 'worried that if I became too strict, the students would hate me'. Participants felt the need to make a good impression on the school staff. They were concerned about meeting their school-based mentor's expectations of them as teachers or having an overly strict mentor teacher who was hard to impress or please. Rose felt that although she had no issues with her school-based mentor, she was not able to work harmoniously with the other school staff as she perceived there was some prejudice towards her as a trainee in the way she was treated: '… my mentor teacher was very helpful and I 'clicked' with her … but the other staff did not seem friendly and I felt that they were biased towards trainee teachers like me'. Wong (2009) in her interviews of 120 new teachers found that recognition, support and affirmation of teaching competencies were important concerns. New teachers needed to know that they were recognized in their teacher roles and accepted as autonomous professionals. Unfortunately, Rose did not elaborate on the nature of the prejudice in her journal.

16.5.2 Classroom Management and Discipline

Participants reported that classroom management was their most worrisome issue. However, their reflections showed that they were not clear about the differences between classroom management and lack of student discipline in class and tended to use the terms interchangeably. Classroom management was related to events that occurred in a classroom such as maintaining order and cooperation to prevent problems from arising; whereas disciplinary problems were those that occurred in the act of handling and managing students' behavioural problems (Levin and Nolan 2000). Examples of some misconceptions were the following: '… among my concerns, the worst was in controlling my class from the point of class management and students who were too noisy'; another wrote that '… aspects of classroom management especially the behaviour of the students'; and another said '… I was worried that I was not able to control the class because I have a kind and lenient personality'. Katy was worried about managing her science practical sessions. She wrote that her training did not prepare her well to do so: 'I was not introduced to proper methods to run a practical session or to handle the situation if something untoward were to happen'.

Many wrote about their attempts to control disruptive behaviour so that lessons could be carried out by using psychology and understanding the emotional make-up of their students. Sharifah wrote: 'I did not feel confident in my teaching as I felt unable to control some of the students in the classroom', while Lina shared that 'I must use positive psychology and ways to approach the different behaviour of my students'. Others needed to alternate between being a 'strict disciplinarian' to being 'an understanding teacher' depending on the behaviour of their students. Many felt they were quite unprepared for the plethora of disruptive behaviours that could occur and that could disrupt their well-planned lessons. Some were surprised that the students did not seem to want to behave. Nurul was quite traumatized with her students' bad behaviour that she cried in her first week: 'I cried because I have failed and was worried that I cannot control my class…'. She wrote that '… the practicum changed my confidence somewhat, before practicum, I felt confident to teach, but after experiencing students who were disrespectful of teachers, the experience made me feel otherwise'. Page (2008) suggested that discipline has been regarded as one of the most prevalent problems experienced by new teachers and, therefore, was considered a serious problem in most schools.

16.5.3 Methods and Strategies

The participants detailed concerns about the limitations and frustrations of their teaching situation. Many participants' written evidence showed that during their practicum experiences, many were worried about 'how' and 'what' to teach. Some of the participants wrote about their concerns of 'choosing the correct methodology and techniques that were appropriate', while others wrote about the need to use newer and creative ways of teaching. Lina wrote: 'I prepared my content for the day well and prepared all sorts of alternative teaching aids to ensure my learning outcomes were met … once when I was badly prepared, I was confused and nervous'. Nora wrote that 'conducting experiments was a challenge to me'. As teaching involved many instructional skills, it could be an arduous task even for experienced teachers, it was therefore not surprising that teachers who were just beginning to get a taste of the actual classroom situation would be anxious about handling new teaching methods and strategies (Freiberg and Driscoll 2005).

Many participants documented concerns about their own teaching activities and performances. They wrote about trying to improve their teaching performance and the need for adequate preparation. Participants were particularly concerned about using English to teach. Statements that showed these concerns were: '… there would be times during teaching, I got lost as I forgot the English words that were equivalent to Bahasa Melayu' or 'I was worried that I would use the English words inaccurately …'. Some indicated that they were not confident during the first few weeks to teach in the English language and were concerned about not being able to find the correct words or use the correct grammar. Lina reflected that having a dialogue with students in English was also a worry because of her lack of proficiency in the language.

Another cause for concern was the participants' mastery of their subject matter, whether they had enough knowledge of the content and adequate teaching aids and materials. Lina shared that 'I was concerned about my mastery of biology and science … I had to answer challenging questions given to me by my students'. Amelia was concerned about the subject she was being given because it was other than what she was trained to teach: 'I could not perform well because I was given another subject that was not a science subject, but I tried my best'. Participants worried about adequate teaching aids, materials and equipment to assist them in the teaching: '… the science lab in the school did not provide me with the equipment and material that I needed, these situations affected my teaching effectiveness'.

Another concern that appeared to impede students was the lack of time to complete the curriculum. There were worried about completing the required syllabus within the 14 weeks time frame. Many wrote about the helpfulness of their school-based mentors in going over their teaching plans and advising changes. It would seem that the students gained confidence if greater support was given by their mentor teachers.

16.5.4 Student Achievement

Aspects of student achievement were participants' concerns for their students' understanding of the subject matter and the concerns for their students' personal growth and moral development. Participants questioned whether they had made an impact in the lives of their students. Fuller and Boun (1975) suggested that teachers who had more concerns for their students than about themselves have reached a level they called 'impact concerns'. Teachers at this level were more concerned about the needs of their students and the effect of their teaching/learning process upon their students' achievement. They questioned whether their students were getting the preparation to be successful in their lives.

Some of the participants detailed concerns for their students' understanding and their developmental needs. To enable the participants to grasp their students' understanding better, some wrote that they encouraged questioning and Amelia tried to ‘relate what was learnt with the reality of the students’ environment’. She indicated that she attempted to instil interest and ‘wanting to know’ among her students by being creative in her teaching and in the process of doing it: ‘increased my own motivation toward becoming a teacher who is dedicated and encouraged to assist my students’. On the other hand, Alina wrote that ‘I know that when I faced an academically weak student, I would endeavour to make my lessons interesting to enable me to attract their attention in the hope that they would develop from being weak academically to being moderately strong academically… that way I knew I would have done a good job in helping my students to be more effective learners at the end of this practicum’. Another concern was the students’ tendency to play truant and this caused concern among the participants as such habits jeopardized understanding of the subject. All the students participated fully in any extra tuition classes organized by their respective schools that were seen as being an opportunity to assist further academically weak students.

Besides academic needs, some participants also said that success in a student’s life was not always about academic achievement but students must also be successful affectively, emotionally and socially. Alina succinctly reflected:

> In my opinion, teaching is a process of delivering knowledge to students who are taught. The knowledge I impart must also be real and able to be realized within the students’ own world. However, academic knowledge alone is not enough; knowledge should also encompass students’ physical, emotional, and spiritual needs.

A few of the participants related that they felt an emotional attachment and a greater connection with their students on more personal levels as the weeks progressed. Amelia wrote that she formed a strong teacher-student relationship by ‘deeply knowing and understanding my students to effect change in their learning’. Many participants expressed the need for their students to succeed. Some wrote that they derived pleasure knowing that their students could grasp difficult concepts.

Alina summarized her feelings and probably that of her fellow student teachers when she wrote:

> Upon completion of this practicum, I am optimistic that I shall use this experience as my ‘provision’ to fall back on when I become a teacher in the near future, God willing and thanks be to God. But I know I shall need to improve continuously my knowledge so that I can face future challenges and concerns ahead.

On the other hand, Lina summed it up with: ‘I am like a budding flower in this area (teaching) and should work hard to learn many things from shaping students to imparting knowledge about something to my students’.

In every written reflection, there is always a large amount of data that cannot be comprehensively reported. However, providing an avenue for the student teachers to write freely and reflect on their teaching tasks has given invaluable insights into how trainee teachers in their instructional experiences managed their practicum, but more importantly for the student teachers to develop their own ‘voice’ while on their professional quest for growth.

16.6 Implications and Recommendations

The transition from being a student being taught to being a teacher teaching is not an easy one and adjustments are to be expected. To help lessen the anxiety of this transition, teacher educators should inform student teachers, either formally or informally, about the changing landscape of teaching today, the diversity that exists in the classroom, giving them a realistic view of today's classrooms and the dynamics of the profession. Some form of support network to allay fears and anxiety for practicum students should be initiated. Hayes (2003) cautioned that the emotional welfare of teacher trainees should not be overlooked as it could have an impact on their success and failure as future teachers. In addition, instead of a one-off 14 weeks practicum, as currently employed in the university, the period student teachers are placed in schools could be extended and more visits to schools arranged to allow student teachers greater opportunity to become familiar with school routines, to work with and to observe experienced teachers.

The concerns that are prominent among all practicum students are those that involve managing students' behaviour and discipline and aspects of classroom management. Although there are discussions and observation of behavioural issues during the student teachers' school orientation programme, student teachers do not appear to be able to draw upon their knowledge to find solutions. As such, greater emphasis should be placed on these. With the student population in classrooms are becoming more diverse in both abilities and needs, and Malaysia is no exception, student teachers should be assisted to better understand the concepts of discipline as overcoming student problems versus classroom management as maintaining order within a class. Student teachers should be assisted to understand better the concept of discipline as overcoming student problems versus classroom management as order within a class enabling a conducive learning environment. According to Freiberg and Driscoll (2005), classes that are not managed well will generally lead to student discipline problems and can inhibit effective instructional approaches from occurring.

Encounters such as difficulty in choosing and using teaching strategies and techniques are also important concerns and are perceived as important for successful teaching in order to achieve positive learning outcomes. Special attention should be given to exposing student teachers in education institutions to varieties of teaching methods and the way these methods can be used and effectively implemented. Probably, the programme in teaching institutions should be more practical with greater emphasis placed on how to translate theory to practice. Assignments should engage student teachers in real school issues and actual teaching problems.

It is evident from the reflections received that the concerns of the student teachers were felt sincerely. They were passionate in their writings as they related their 14-week experiences – both personal and professional concerns associated with their role as 'trainee teachers' – as they grappled to understand their working environment better. However, because of the practicum and student teachers' involvement in the study and the requirement to be reflective in their writings, they have engaged not only in analysing their experiences but have also come to terms with some of the conflicts and dilemmas of teaching. They confronted their own attitudes and values about their teaching. This was evidenced from, for example, 'teaching required not only skills but a lot of patience to succeed as a good teacher'. Another wrote: 'I felt that teaching was very challenging because it tested both your physical and mental strength'. Yet another realized her strengths and weaknesses, another saw the holistic process of teaching which according to her: 'was not simply imparting the content of a subject, but a combination of proper class management, controlling behavioural issues, proper sequencing of lessons, and above all instilling a sense of fun among pupils as they attained knowledge'.

16.7 Concluding Discussion

It is not the intention of the researchers to provide generalizations for all practicum students as the findings have been limited to a group of student teachers from one university. Rather, the study has extended the request of the Malaysian Ministry of Education (MOE) to study and evaluate teaching practices of student teachers as an avenue to identify and examine concerns experienced by teachers during their practicum. The Ministry contends that any findings can further enhance and improve the education programmes in teacher training institutions and ultimately lessen the concerns and worries of student teachers going out for their teaching practice (IPT et al. 2005). In addition, and perhaps more importantly, the study has provided insights into the formative experiences practicum students have in learning how to teach from a distinctly Malaysian perspective. However, it is not enough to simply identify and categorize the problems practicum students face, more importantly – it is to provide ways to prevent and manage those areas of concern that must be integrated into future education courses. How to integrate the theoretical aspects learnt at university and the practical reality of the classroom needs to be established in order to assist student teachers 'survive' the practicum experience. Education courses need to be more applicable to actual school settings and environments. A systematic way for teacher educators to periodically review course content to ensure that problem areas are included in the curriculum should also be established.

There also seems to be some merit in allowing practicum students an avenue to explore reflectively their own experiences in a meaningful way that would help promote the independence and critical thinking necessary for the challenges ahead as future in-service teachers and as part of the espoused 'lifelong learning' call. Incorporating a structured approach for student teachers on practicum to self-assess their learning would provide them with the opportunity to develop their reflectivity and accept responsibility for their own professional development.

The focus of this study has been to listen to the 'voice' of the student teachers during their practicum through their thoughtful and careful self-assessment of their teaching practice and experiences. The value of the study was in the pursuit of using student teachers' capacity to self-assess and appraise their circumstances as a research area in teaching to reveal a higher complexity of learning in their specific professional domain. Further, it showed how the understanding of learning to teach could be enriched through their own awareness of the circumstances surrounding them. However, two questions still need to be answered: (a) how can teacher education programmes harness and encourage the development of such self-assessment or 'growth experience' among trainee teachers on their practical experiences towards creating a high but realistic level of confidence and optimism in Malaysian students aspiring to be teachers and (b) could learning to appraise and understand why an event or activity occurs be an important strategy to transform student teachers' progression from student to master teacher? These are questions for a further study.

Acknowledgements This study was supported by a grant from Sultan Idris Education University (Malaysia) given to the first author. Both authors gratefully acknowledged each of the 14 student teachers involved in this study and who have kindly contributed with integrity and honesty.

References

Boud, D., Keogh, R., & Walker, D. (1985). *Reflection: Turning experience into learning*. London: Kogan Page.

Briggs, L. D., & Richardson, W. D. (1992). Causes and sources of student concerns for student teaching problems. *College Student Journal, 26*(2), 268–272.

Chan, K. W., & Leung, M. T. (1998). *Hong Kong preservice teachers' focus of concerns and confidence to teach – A perspective of teacher development*. Retrieved from AARE the Association for Active Educational Researcher, website: http://www.aare.edu.au/98pap/cha9836.htm

Collier, S. T. (1999). Characteristics of reflective thought during the student teaching experience. *Journal of Teacher Education, 50*(3), 173–181.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston: Heath.

Education Commission. (2000). Learning for life, learning through life: Reforms proposals for the education system in Hong Kong. Hong Kong: Government Printer.

Freiberg, H. J., & Driscoll, A. (2005). *Universal teaching strategies* (4th ed.). Boston: Pearson.

Fuller, F. F., & Boun, O. H. (1975). Becoming a teacher. In K. Ryan (Ed.), *Teacher education: Seventy-fourth yearbook of the National Society for the Study of Education* (pp. 25–52). Chicago: University of Chicago Press.

Hall, J., & Bowman, A. (1989). *The journal as a research tool: Preservice teacher socialization*. Paper presented at the annual meeting of the Association of Teacher Educators in St. Louis.

Hatton, N., & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education, 11*(1), 33–49.

Hayes, D. (2003). Emotional preparation for teaching: A case study about trainee teachers in England. *Teacher Development, 7*(2), 153–172.

Hoy, W., & Woolfolk, A. (1990). Socialization of student teachers. *American Educational Research Journal, 27*(2), 279–300.

IPT, BPG, & MOE. (2005). *The evaluation of teaching practice in teacher education programs in Malaysia*. Penang: School of Educational Studies, Universiti Sains Malaysia.

Kabilan, M. K., & Raja Ida, R. I. (2008). Challenges faced and the strategies adopted by a Malaysian English Language teacher during teaching practice. *English Language Teaching, 1*(1), 87–95.

Kember, D., & Leung, D. Y. P. (2005). The influence of the teaching and learning environment on the generic capabilities needed for a knowledge-based society. *Learning Environments Research, 8*, 245–266.

Levin, J., & Nolan, J. F. (2000). *Principles of classroom management: A professional decision-making model*. Boston: Allyn and Bacon.

Lock, C. R. (1977). Problems of secondary school student teachers. *The Teacher Educators, 13*(1), 30–40.

Loughran, J. J. (1996). *Developing reflective practice: Learning about teaching and learning through modelling*. London: The Falmer Press.

Marton, F. (1986). Phenomenography – A research approach to investigating different understanding of reality. *Journal of Thought, 21*, 28–49.

Ong, S. K., Ros, A. S., Azlian, A. A., Sharnti, K., & Ho, L. C. (2004). *Trainee teachers' perceptions of the school practicum*. Paper presented at the conference of the National Seminar on English Language Teaching 2004. Bangi, Malaysia.

Page, M. L. (2008). *You can't teach until everyone is listening: Six simple steps to preventing disorder, disruption, and general mayhem*. California: Corwin Press.

Price, D. A. (1987). The practicum and its supervision. In K. J. Eltis (Ed.), *Australian teacher education in review* (pp. 105–133). Bedford Park: South Pacific Association for Teacher Education.

Schön, D. A. (1996). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. San Francisco: Jossey-Bass, Inc.

Smith, K., & Lev-Ari, L. (2005). The place of the practicum in pre-service teacher education: The voice of the students. *Asia-Pacific Journal of Teacher Education, 33*(3), 289–302.

Wolf, K., & Siu-Runyan, Y. (1996). Portfolio purposes and possibilities. *Journal of Adolescent and Adult Literacy, 40*(1), 30–37.

Wong, Y. F. I. (2009). *Toward an agenda for helping the beginning teacher: Perceptions of concerns and best help strategies*. Retrieved from AARE the Association for Active Educational Researcher, website: http://www.aare.edu.au/03pap/won03819.pdf

Yinger, R., & Clark, C. (1981). *Reflective journal writing: Theory and practice*. East Lansing: Institute for Research on Teaching, Michigan State University.

Yourn, B. R. (2000). Learning to teach: Perspectives from beginning music teachers. *Music Education Research, 2*(2), 181–192.

Zeichner, K. M., & Liston, D. P. (1987). Teaching student teachers to reflect. *Harvard Educational Review, 57*(1), 23–48.

Chapter 17
Informing Learning and Teaching Using Feedback from Assessment Data: Hong Kong Teachers' Attitudes Towards Rasch Measurement

Chi Ming Ho, Anthony Wai Chi Leung, Magdalena Mo Ching Mok, and Paisley Tsz Mei Cheung

17.1 Introduction

17.1.1 Assessment Reform in Hong Kong

Assessment for learning has been a central tenet of the Hong Kong SAR government's education reform since 2000 (Curriculum Development Council 2001). This emphasis on assessment for learning is unambiguously articulated in a recent address by Dr K. K. Chan, principal assistant secretary for education:

> Assessment is an inextricable element of learning. With the introduction of the concept of 'Assessment for Learning', we all agree that the objectives of assessment in education are to understand students' learning progress, recognise their individual achievements and develop their diverse potential, so as to enhance their whole-person development. It also serves as a basis for improving teaching and learning. (Chan 2008)

C.M. Ho
Formerly Centre for Assessment Research and Development,
The Hong Kong Institute of Education, Tai Po, Hong Kong
e-mail: cmho2009@yahoo.com.hk

A.W.C. Leung
HHCKLA Buddhist Wisdom Primary School, Sheung Shui, Hong Kong
e-mail: waichi97@hotmail.com

M. Mo Ching Mok (✉)
Department of Psychological Studies, and Assessment Research Centre,
The Hong Kong Institute of Education, 10 Lo Ping Road,
Tai Po, N.T., Hong Kong
e-mail: mmcmok@ied.edu.hk

P.T.M. Cheung
Education Assessment Services, Hong Kong Examinations
and Assessment Authority, Wanchai, Hong Kong
e-mail: paisleycheung@gmail.com

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_17, © Springer Science+Business Media Dordrecht 2013

Reform initiative towards assessment for learning is a response to a number of developments both in the region and internationally. Developments in theories of learning and knowledge are one powerful factor for curriculum change worldwide. The shift from an understanding of learning as a static process to a view that learning is a dynamic process whereby knowledge is constructed through the active involvement of the learner urges educators and policymakers to rethink the role of assessment in the learner's knowledge construction process (Klenowski 1998; Mok et al. 2003). The enormous impact of the economic crisis at the turn of the century has forced Asia-Pacific country leaders to face the deep-rooted problems in their education systems and given strong impetus for education reform. Policymakers are being challenged to find answers to 'Where to?' and 'How?' for this reform. The review by Black and Wiliam (1998) gives one possible direction. The findings refocused assessment 'on classroom processes' (Black and Wiliam 2003, p. 628) and highlighted quality feedback as the crucial ingredient in assessment. The purpose of this type of assessment is to generate information to support and advance learning using feedback generated from the assessment process, hence the name 'formative' assessment.

The success or otherwise of the implementation of assessment for learning relies heavily on the knowledge, values and skills of teachers in using assessment data, not only for diagnostic purposes but also for fine-tuning their instruction and providing essential feedback to foster student reflection and improve subsequent learning (Wiliam and Thompson 2008). Teachers' attitudes influence their decision-making in the classroom and their perceptions and evaluations of outcomes, choice of instructional methods and student achievement (Hofer and Pintrich 1997; Priestley 2005; van der Schaaf et al. 2008). Teachers' attitudes, associated with teacher efficacy, hold the key to implementing reform (Bruce and Ross 2008; Tierney 2007; Wiliam and Thompson 2008).

This chapter focuses attention on teachers' attitudes towards Rasch measurement. In particular, it looks at teachers' beliefs about the potential benefits of using reports produced from Rasch analysis as tools for assessment for learning and the challenges facing them in implementing this method of assessment. The authors are guided by the following research questions:

1. What were teachers' attitudes towards the desirability of Rasch measurement?
2. What were teachers' attitudes towards the feasibility of Rasch measurement?

17.1.2 The Role of Feedback in Assessment

Feedback is an essential component of assessment because it can often generate opportunities for student reflection which, in turn, results in enhanced motivation and learning (Marriott 2009; Poulos and Mahony 2008). As Leahy et al. (2005) pointed out, effective feedback should inspire thinking and move learners forwards. The significance of feedback in the assessment process can best be summarized in

Marriott's (2009) words: 'feedback is…[a] conduit for facilitating student self-assessment and reflection, encouraging positive motivational beliefs and self-esteem' (pp. 238–239). Consequently, the strategic and well-planned use of feedback is critical to both learning and teaching (Ellery 2008; Marriott 2009; Poulos and Mahony 2008).

Nevertheless, not all feedbacks are effective. A number of factors that contribute to the effectiveness of feedback in the learning process have been highlighted in the literature. These factors are timeliness, frequency and appropriateness, as well as the mode of feedback delivery and the actual practice of its use (Marriott 2009; Poulos and Mahony 2008). Furthermore, feedback should be well constructed so that it can also communicate critical information between students and teachers. The information can show students their specific strengths and weakness and so help students to remedy their deficiency and improve their performance in the future (Black and Wiliam 1998; Marriott 2009; Poulos and Mahony 2008).

17.1.3 Assessment for Learning and Teachers' Attitudes

Assessment for learning is a topical issue in Hong Kong. To ensure quality learning, valid and reliable assessments should be developed that show the learning progress and needs of students as well as the effectiveness of teaching. Nevertheless, assessment reform efforts are often focused on the curriculum rather than on either students or teachers (Edwards et al. 2008, p. 683). The current vision of assessment for learning is that teachers use insights from ongoing assessment of students' strengths and weakness to advance curriculum and instruction. However, teachers are finding the implementation of this vision a great challenge (McNamee and Chen 2005; Watering et al. 2008). Factors contributing to teachers' success or failure in implementing the reform vision deserve to be explored. Among these factors, teachers' attitudes towards assessment, especially what and how effective feedback information is provided, are critical. As Wong (2006) pointed out, because teachers are the key to the success of any implementation of educational policy, their attitudes towards their daily practice need to be discussed and explored.

In order to enhance learning and teaching, feedback based on formative assessment should be timely and specific (McTighe and O'Connor 2005). Several authors (e.g. Campbell 2008; Lim and Chai 2008; Robertson 2008) have recommended that educators glean from recent advancements in information technology to support teacher implementation of reform initiatives. These studies highlighted three factors as important in teachers' effective use of technology and as an integral component in their pedagogy: the teachers' dispositions towards new technology, their values regarding innovation and their self-efficacy in using technology. Indeed, teachers' knowledge, experience and attitudes towards teaching by means of technological tools might have significant potential impacts on technology integration (ChanLin et al. 2006; Kessler 2007).

A growing body of literature has focused on exploring teachers' attitudes towards and perceptions of their educational practice. For example, Kessler's (2007) study demonstrated that the preparation of informal computer-assisted language learning was closely linked to teachers' attitude towards technology. The study by Spiropoulou et al. (2007) explored in-service primary teachers' perceptions about environmental issues and attitudes towards education for sustainable development. The results revealed that teachers' inexperience in new methodological approaches to promoting environmental matters led to them having less interest in the environmental programmes, and, as a consequence, the implementation rate of these programmes was relatively low. A study by Flowers et al.(2005) suggested that the increase in paperwork and demands on time had the most significant impact on teachers' use of alternative assessment. ChanLin et al. (2006) identified factors influencing teachers' integration of information technology into their teaching. These factors could be classified into four categories: environmental, personal, social and curricular. Environmental factors related to computer facilities, support and management of resources (including staffing), and in-service training. Personal factors were about a teacher's personality and beliefs. Social factors referred to the level of support received from colleagues and senior administration, as well as from students, parents and the community. Curricular issues meant what factors teachers took into consideration when both teaching and assessing achievement of learning objectives. Wong's study (2006) indicated that teachers' daily practices were highly influenced by the examination system and the massive workload. Finally, Ekiz's study (2006) showed that the primary factor motivating teachers to undertake educational research was to do what was best for their students.

The results of these and similar studies serve as reference for educators or teachers when advocating or implementing educational innovations. As the implementation of assessment for learning requires teachers to change their ways of thinking (Black et al. 2003), it is worth exploring teachers' attitudes towards using feedback from assessment data to inform learning and teaching. With this in mind, the purpose of the current study was to understand teachers' attitudes towards the value of the Rasch model of analysing assessment data. Through the study, it was intended to unearth what kind of information teachers considered to be useful for effective feedback and what are the factors that affect the teachers' use of feedback generated from the Rasch analysis.

17.1.4 Rasch Measurement and Assessment for Learning

Rasch measurement (Wright and Masters 1982) is a tool being used increasingly by educational researchers in large-scale testing (Alagumalai et al. 2005; Callingham and Bond 2006). There are a number of Rasch measurement models (Wright and Mok 2004). The basic Rasch model (Rasch 1960) for dichotomous items (i.e. item responses are either right or wrong) is a probabilistic model that describes the probability of getting an item correct in terms of a simple logistic function of the difference between the person's ability and the item difficulty: the higher the ability of the person compared with the item difficulty, the higher the probability of getting the item correct. The converse is also true: the lower the ability of the person compared with the item difficulty, the lower the chances of the person getting the item right. If the ability of the person is the same as the difficulty level of the item, then the probability of getting the item right is 50%. The logistical transformation converts ordinal data from educational measurement into linear measures, where the unit of measurement is called logit (log-odds unit) (Bond and Fox 2007).

The basic Rasch model was extended to the polytomous Rasch model by Andrich (1978). The polytomous Rasch model can handle partial credits in achievement scoring; for example, achievement items can be scored 0, 1, 2 or 3 marks to reflect different levels of achieving the standard answer (Masters 1982). It can also be used with rating-scale responses, such as Likert-type questionnaire items with possible responses 'strongly disagree', 'disagree', 'agree' and 'strongly agree' to reflect different levels of agreement with an item (Andrich 1978).

The most important feature of the Rasch model for this study is the ordered conjoint measurement scale of both person and item, which enables teachers to inspect the relative positions of students and assessment items on the same measurement scale. Figure 17.1 is an example of such a scale produced by the Winsteps software (Linacre 2011), using data from an assessment of Chinese reading comprehension.

In Fig. 17.1, students are placed on one side of the measurement scale and assessment items on the other. On the left, students with higher comprehension levels (i.e. students with ID numbers 16, 27, 38 and 44) are placed at the top end of the scale and students with lower comprehension levels (i.e. students with ID numbers 17, 23 and 39) at the low end of the scale. On the right, the more difficult assessment items (e.g. Q10 [Elaboration]) are placed at the top of the scale and the less difficult items (e.g. Q5 [Recall]) at the bottom. If a student and an item are at the same level, then the ability of the student is the same as the difficulty level of the item. In such cases, the student has 50 % chance of getting the item correct. Conceptually, this is Vygotsky's zone of proximal development (ZPD) (Vygotsky 1978) of the student. The same student has more than 50 % chance to answer correctly those items below their level, and the further the student lies above the items, the higher the probability they have of getting these items correct. Mastery is where the student has close to 100 % chance of getting the item right. Similar arguments can be used for those items at locations well above the student's ability level. If an item is so advanced compared with the current ability level of the student, then there is close to zero chance of the student getting a right answer for that item.

If the items are designed so that they are aligned with the curriculum, then the teacher will have valuable information on regions of mastery, ZPD and 'regions beyond the current ability' of individual students. By identifying the ZPD of each student, the teacher can provide appropriate scaffolding to support their learning. Furthermore, by comparing where the student currently is and where they could be on the scale, the potential of each student can be established, and, based on this information, the teacher can then design individualized instruction for each of their students. On a broader scale, by inspecting the pattern of distribution of items versus that of the whole group of students, the teacher can develop a pretty clear idea of the strengths and weaknesses of the group and modify teaching instructions accordingly. Using the item–person map produced by Rasch measurement, the assessment can

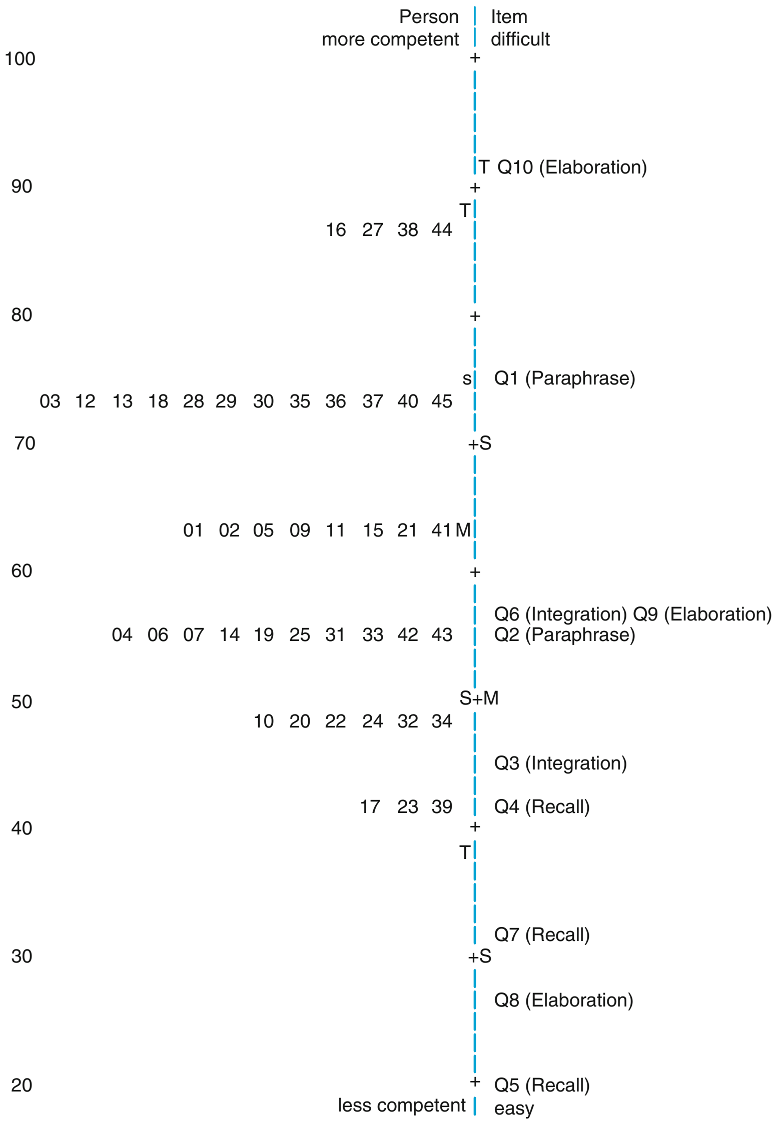
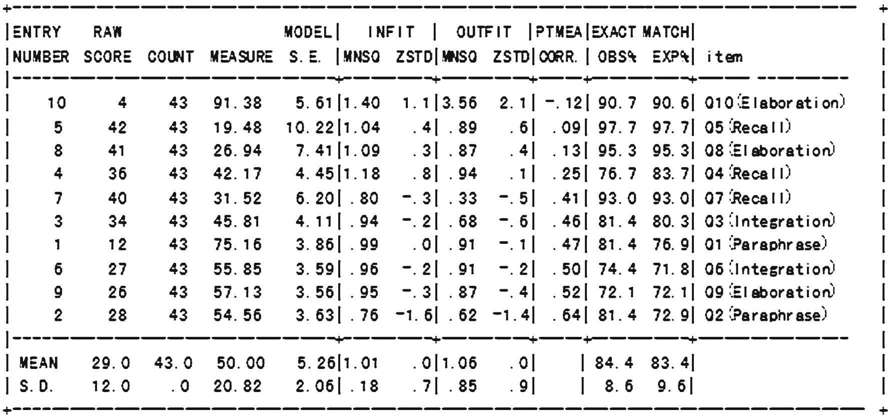


Fig. 17.1 Item–person map

be truly formative in the sense of Black and Wiliam (1998) because feedback information can be used to guide subsequent teaching and learning.

It is important to note that the item–person map can also be used by students to help them gain a deeper understanding regarding their own performance and that of their peers. By identifying where they are and where they could be, with appropriate scaffolding from the teacher, the students can assume responsibility for their own learning and build up a sense of control.

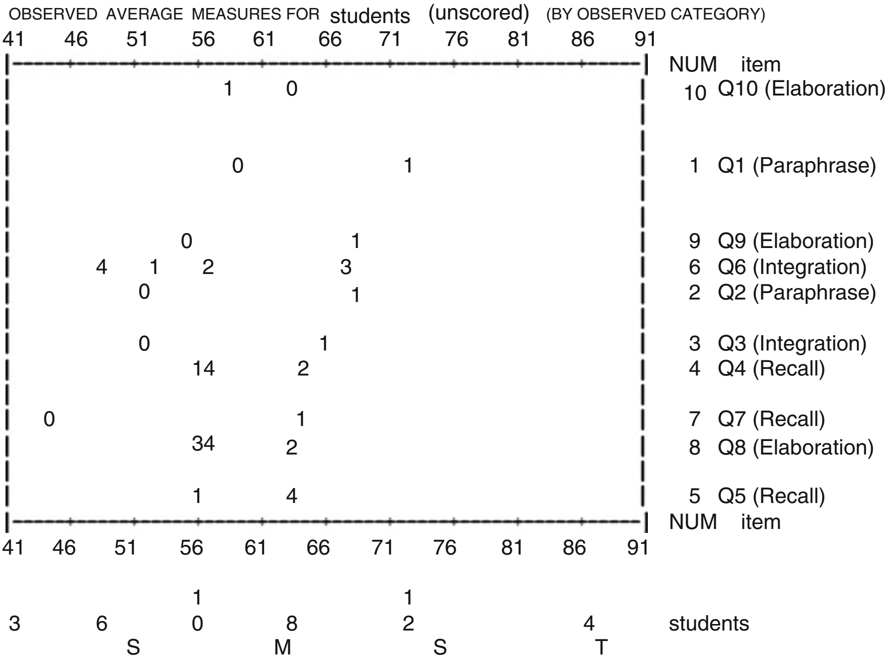


```
+--------------------------------------------------------------------------------------------------+
|ENTRY   RAW                          MODEL|   INFIT   |  OUTFIT   |PTMEA|EXACT MATCH|                |
|NUMBER  SCORE  COUNT  MEASURE   S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| OBS%  EXP%| item            |
|------------------------------------+----------+----------+-----+-----------+---------------  |
|   10      4     43    91.38    5.61|1.40   1.1|3.56   2.1| -.12| 90.7  90.6| Q10(Elaboration) |
|    5     42     43    19.48   10.22|1.04    .4| .89    .6|  .09| 97.7  97.7| Q5(Recall)       |
|    8     41     43    26.94    7.41|1.09    .3| .87    .4|  .13| 95.3  95.3| Q8(Elaboration)  |
|    4     36     43    42.17    4.45|1.18    .8| .94    .1|  .25| 76.7  83.7| Q4(Recall)       |
|    7     40     43    31.52    6.20| .80   -.3| .33   -.5|  .41| 93.0  93.0| Q7(Recall)       |
|    3     34     43    45.81    4.11| .94   -.2| .68   -.6|  .46| 81.4  80.3| Q3(Integration)  |
|    1     12     43    75.16    3.86| .99    .0| .91   -.1|  .47| 81.4  76.9| Q1(Paraphrase)   |
|    6     27     43    55.85    3.59| .96   -.2| .91   -.2|  .50| 74.4  71.8| Q6(Integration)  |
|    9     26     43    57.13    3.56| .95   -.3| .87   -.4|  .52| 72.1  72.1| Q9(Elaboration)  |
|    2     28     43    54.56    3.63| .76  -1.6| .62  -1.4|  .64| 81.4  72.9| Q2(Paraphrase)   |
|------------------------------------+----------+----------+-----+-----------+---------------  |
| MEAN   29.0   43.0    50.00    5.26|1.01    .0|1.06    .0|     | 84.4  83.4|                  |
| S.D.   12.0     .0    20.82    2.06| .18    .7| .85    .9|     |  8.6   9.6|                  |
+--------------------------------------------------------------------------------------------------+
```

Fig. 17.2 Item polarity table

In addition to the item–person map, Rasch analysis using Winsteps (Linacre 2011) also produces an item polarity table (Fig. 17.2) which can be used by teachers to refine their items. Specifically, the item polarity table contains two columns with fit mean-squared values and PTMEA correlations. The fit mean squared is a statistic which reflects how well the data fits the Rasch measurement model. Conventionally (Bond and Fox 2007, pp. 240–251), fit mean-squared values between 0.75 and 1.3 are considered a good match between the data and the Rasch model. Values beyond this range suggest discrepancies between the data and the Rasch model. As far as items in the assessment are concerned, the implication is that the response patterns to these items have too much noise, and these items have to be carefully inspected as to whether or not they are measuring the same construct as the other items in the batch. For example, item 10 in Fig. 17.2 has a large fit mean-squared value of 1.40. In response to this statistic, teachers should read the item more carefully and see how to improve its validity or consider whether even to include it in the next test administration.

PTMEA correlation (Fig. 17.2) is a shorthand notation for point-measure correlation. It is a statistic that reflects the correlation between the measure of one item and the total measure for the whole test. If the PTMEA correlation is low (e.g. item Q5 in Fig. 17.2), the item is likely to be measuring some different traits from the rest of the test. For instance, if the test is assessing Chinese reading comprehension but item Q5 in the test requires mathematical skills to understand the question, then item Q5 will have a low PTMEA correlation. Items with a negative PTMEA correlation (e.g. item Q10 in Fig. 17.2) are measuring something opposite to the rest of the test and so should be reviewed carefully. Items with PTMEA correlations lower than 0.4 (e.g. items Q10, Q5, Q8 and Q4 in Fig. 17.2) should also be reviewed carefully to understand the reasons behind the anomaly. For example, items that are either too easy (where almost everybody gets the right answer) or too difficult (where almost everybody gets the wrong answer) tend to have low PTMEA correlations since changes in their scores (which tend to be small) are unlikely to be associated with changes in the overall score.



```
 OBSERVED AVERAGE MEASURES FOR students (unscored) (BY OBSERVED CATEGORY)
41   46   51   56   61   66   71   76   81   86   91
|----+----+----+----+----+----+----+----+----+----|  NUM  item
|                1    0                           |   10 Q10 (Elaboration)
|                                                 |
|                                                 |
|                  0           1                  |    1 Q1 (Paraphrase)
|                                                 |
|                                                 |
|             0            1                      |    9 Q9 (Elaboration)
|      4   1   2          3                       |    6 Q6 (Integration)
|         0                1                      |    2 Q2 (Paraphrase)
|                                                 |
|         0              1                        |    3 Q3 (Integration)
|              14      2                          |    4 Q4 (Recall)
|                                                 |
|  0                   1                          |    7 Q7 (Recall)
|              34     2                           |    8 Q8 (Elaboration)
|                                                 |
|              1      4                           |    5 Q5 (Recall)
|----+----+----+----+----+----+----+----+----+----|  NUM  item
41   46   51   56   61   66   71   76   81   86   91

               1               1
3      6       0      8        2             4     students
          S          M            S           T
```

Fig. 17.3 Empirical item–category measures

The empirical item–category measures (Fig. 17.3) are another helpful output from Rasch analysis using Winsteps (Linacre 2011). Teachers can use these as a tool to improve the assessment items, particularly for multiple-choice items. The empirical item–category measures in Fig. 17.3 are expressed in a two-dimensional grid. The measurement scale used in the item–person map is re-expressed as two scales. The horizontal axis on the top is the scale to indicate the ability of the students. The vertical axis on the right is the scale to indicate the item difficulty. For example, students choosing option 3 in item Q6 in Fig. 17.3 are estimated to have an ability rating of 68, while students choosing option 2 in the same item would have an estimated ability rating of 57.

The empirical item–category measures table from Winsteps gives two pieces of important information. First, it informs the teacher whether students choosing the right option (e.g. option 3 of item Q6 in Fig. 17.3) are actually more able than students choosing the wrong options (e.g. other options of item Q6 in Fig. 17.3) as expected. If this is not the case, then the item and its response options should be modified because the item is tricking more able students to make wrong responses. Second, the empirical item–category measures table informs the teacher whether or not there is discrimination among the options. For instance, students choosing options 1 and 4 of item Q4 in Fig. 17.3 have very similar abilities.

The empirical item–category measures table can also be applied to items involving partial credit, where items are given partial scores such as 1 mark for a partly right answer and 2 marks for a fully correct answer, or to attitude questionnaire items. The analysis is exactly the same; for example, if the possible scores for an item are 0, 1, 2 and 3 but the empirical item–category measures table shows that students scoring 3 actually have ability lower than those scoring 1, then the scoring guidelines for this item need to be closely inspected. Alternatively, if the abilities of students scoring 1 or 2 marks in this item are not distinguishable, then teachers might consider whether the scoring system for this item should be collapsed into a 3-point partial credit item (with possible scores of only 0, 1 or 2) instead of a 4-point partial credit item (with possible scores of 0, 1, 2 or 3). As it can be more time-consuming for a teacher to use a multi-point partial credit scoring assessment with more scoring points, then the empirical item–category measures table can be used by teachers to help them design scoring systems that will make grading more efficient.

17.2 Research Design

17.2.1 Sample

The participants of this study were 25 primary school teachers and 24 secondary school teachers. All of them were current teachers of Chinese language. Among them, 30 were also responsible for administrative duties, such as acting as department head of the curriculum subject at their school.

17.2.2 Data Collection

Three to six teachers from each of the ten participating schools were invited to take part in focus-group interviews. Teachers from the same school were interviewed together in the same focus group. The focus-group interview had three components: (1) demonstration of Rasch analysis and presentation of results from an analysis, (2) a questionnaire survey on attitudes towards Rasch measurement and (3) the focus-group interview on the teachers' attitudes towards the costs and benefits of Rasch measurement as a means to support assessment for learning. These are described in detail below.

At the start of the focus-group interview, the teachers were presented with a short demonstration on how to analyse assessment data by Rasch measurement using the Winsteps computer software (Linacre 2011). The analysis took about 10–15 min. The analysis yielded the following information for the focus group:

(a) An item–person map (Fig. 17.1) showing the competence of students and difficulty of items on the same scale.

(b) An item polarity table showing the fit mean-squared value and PTMEA correlation for each item so that the teachers could check for positive correlations.
(c) Empirical item–category measures showing the most probable response on the latent variable in multiple-choice questions so that the teachers could check whether correct answers and higher category values corresponding to 'more' of the variable are to the right.

After the demonstration of a Rasch analysis and presentation of results, a questionnaire was given to each teacher before the focus-group interview. The questionnaire had a section on teachers' background information, followed by another section with rating-scale items to solicit their attitudes towards Rasch measurement. The attitude items included in the questionnaire are:

(a) What was your first impression of this analysis method?
(b) Given marks ranging from 1 to 7, how did you rate the value (desirability) of the analysis method?
(c) What were the reasons you value or did not value the implementation of this analysis method?
(d) Given marks ranging from 1 to 7, how did you rate the feasibility of applying this analysis method in your school?
(e) If you were going to use this analysis method in your school, what supports would be necessary?
(f) What would be the obstacles to the implementation of this analysis method?

Question (a) asked for the teachers' first impression of the analysis method. Questions (b) and (c) intended to explore the teachers' attitudes towards desirability on the method, while questions (d) to (f) aimed at finding out their attitudes towards its feasibility.

The teachers were given ten minutes to complete the questionnaire. They were able to refer to their responses on their questionnaires during the focus-group interview. The focus-group interview lasted between 50 and 60 min and was videotaped. Teachers' responses in the focus-group interview were transcribed from the audiotapes, and a coding system was developed under thematic analysis.

The researchers read the teachers' anonymous transcripts separately in a separate setting. They tried to identify the 'big ideas' (Krueger 1998) from the transcripts as a whole. They then encoded these big ideas with key phrases. Afterwards, the researchers met together and compared their codings to see whether they matched. If there was any disagreement, they read the transcripts again and tried to reach a compromise. If necessary, the data was recoded. Finally, the researchers developed a common coding system with sorted categories. The researchers used this coding system to analyse all the transcripts again. The goal of this analysis was to find out the teachers' main concerns about the desirability and feasibility of the Rasch measurement method as a tool to support assessment for learning in primary and secondary schools.

17.3 Results

17.3.1 The Teachers' First Impressions of the Rasch Model

The teachers' first impressions of Rasch models were gauged using qualitative data which was extracted from analysing teachers' interview transcripts.

In response to the question 'What is your first impression of this analysis method?', 97 responses were extracted from 49 teachers' transcripts. These responses were encoded into eight themes, five of which were positive first impressions from 73 of the teachers' responses, and the other three themes were negative first impressions from 24 responses. The five positive themes were (1) understanding students' abilities and individual differences, which included responses referring to how the Rasch method enhanced the teacher's understanding of their students' learning; (2) understanding the validity and reliability of the measuring items, which included responses pointing to how the Rasch method could help teachers to identify psychometric properties of the assessment items; (3) powerful analysis of assessment data, which included comments relating to the details that could be provided by the Rasch method in relation to the assessment outcomes; (4) presentation of data analysis is user-friendly, which included teachers' remarks on the user-friendliness of the Rasch method; and (5) enhancement of teachers' professional competence, which included responses on how the Rasch method could support teachers' professional judgement and so enhance their competencies. The majority of the responses could be categorized within the first two themes, with 23 responses in each. Sample responses and the relative importance of the themes are presented in Table 17.1.

Twenty-four teachers had a more reserved or even negative first impression about the Rasch model. Their responses were encoded into three themes: (6) apprehension, which included negative responses indicating concerns or fear about the technology or the complexity of the assessment; (7) demand on resources, which included responses referring to concerns about the time and human resources required to implement the Rasch method; and (8) suspicion of the effectiveness of the assessment method, which included remarks showing teachers' suspicions about the capacity and effectiveness of the Rasch method in supporting teaching and learning. Apprehension was by far the largest of the three negative themes, with 20 teachers showing fear, concern and hesitation (Table 17.1).

17.3.2 Teachers' Attitudes Towards Desirability of Implementing the Rasch Method

Teachers' attitudes towards the desirability of implementing the Rasch models were gauged using both qualitative and quantitative data. Qualitative data was extracted from analysing teachers' interview transcripts, and quantitative data was obtained from teachers' ratings on the questionnaire.

Table 17.1 The first impression and desirability of teachers of the Rasch measurement

Coding	Definition and sample from transcripts	No. of responses first impression ($n=97$)	No. of responses desirability ($n=100$)
(a) Favourable first impressions and indications that the Rasch assessment is worthwhile		Favourable impression ($n=73$)	Worthwhile ($n=82$)
Understanding student ability and individual differences	Definition: The Rasch method could enable teacher to understand the learning performance and abilities of individual students Sample of first impression: ‘Can identify which students are more able, which are less able. That means you can clearly know the performance and ability of individual students’ Sample of desirability: ‘And the most important is to show the distribution of students’ abilities, what they know and what they don’t know’	23	28
Understanding the validity and reliability of the measuring items	Definition: The Rasch method could help teachers to identify psychometric properties of the assessment items Sample of first impression: ‘Analyse whether the design of items are good or not, and the analysis reflects item quality. I think it is invaluable’ Sample of desirability: ‘It can show item validity…We can find out whether we can trust the measures’	23	22
Powerful analysis of assessment data	Definition: The Rasch method could provide the details in relation to assessment outcomes Sample of first impression: ‘Its analysis is more in-depth and in greater detail’ Sample of desirability: ‘The analysis is comprehensive’	13	19
Presentation of data analysis is user-friendly	Definition: The Rasch method seems to be user-friendly Sample of first impression: ‘Both the difficulties of items and students’ performances are shown clearly’ Sample of desirability: ‘I think that it is clear and easy to understand’	10	7

Enhancement of teachers' professional competence	Definition: The Rasch method could support teachers' professional judgement and enhance their competencies Sample of first impression: 'I think teachers can self-reflect on both the student ability and teacher competency in item-setting. It is helpful' Sample of desirability: 'Empower teacher to reflect. That means when we design each option of answers for an item for the students, we will be more careful in reviewing which options are functioning. We will ask whether or not any unnecessary answer exists. Consequently, I think the professional development of teachers in this aspect of item setting (measuring which level) can be enhanced'	4	6
(b) Unfavourable first impressions and indications that the Rasch assessment is not worthwhile		Resistance ($n = 24$)	Not worthwhile ($n = 18$)
Apprehension	Definition: Teachers indicated negative reactions such as worry or fear about the technology or the complexity involved Sample of first impression: 'My first impression is that it is complex. From the first sight I read it, some colleagues may be resistant in using it' Sample of desirability: 'Actually, I feel scared'	20	4
Demand on resource	Definition: Teachers raised their concern with the demand on time and human resources in the implementation of the Rasch method Sample of first impression: 'Need support from experts' Sample of desirability: 'Time is a problem'	2	4

(continued)

Table 17.1 (continued)

Coding	Definition and sample from transcripts	No. of responses first impression ($n=97$)	No. of responses desirability ($n=100$)
Suspicion of effectiveness of the assessment method	Definition: Teachers showed their suspicions on the capacity and effectiveness of the Rasch method in supporting teaching and learning Sample of first impression: ‘I am suspicious about how the difficulties of items are to be defined?…Even if we adopt it, I question whether or not the reviewed items can be helpful to students. I think it depends on ones’ competence on item setting’ Sample of desirability: ‘It certainly has inadequacy. For example, some open-ended questions. That means how language teachers give marks to students. This tool may not be suitable for this kind of testing’	2	10

The eight themes were encoded and obtained from the teachers' interview transcripts. It can be seen from Table 17.1 that 100 responses were extracted from 48 teachers' transcripts. There was strong concordance between the teachers' first impression and their comments about the perceived strengths of the Rasch model, i.e. its desirability. Eighty-two responses indicated that the teachers perceived the Rasch measurement as a worthwhile undertaking. Similar to their 'first impression' responses, analysis of the teachers' 'desirability' responses found that the strongest theme was the capacity of the Rasch analysis to help teachers to better understand their students' abilities and individual differences (theme 1; $n=28$; Table 17.1). The next strongest theme was that the Rasch analysis enabled deeper understanding of the validity and reliability of the assessment items (theme 2; $n=22$; Table 17.1); this result again concurred with the teachers' 'first impression' responses. Teachers regarded information generated from the Rasch analysis on the strengths and weaknesses of students and of items to be informative and useful for giving feedback. According to the teachers, information generated from the Rasch analysis would enable teachers to give feedbacks that were more concrete and relevant to student learning. This point of view can be illustrated by two teachers' responses:

> The Rasch model allows us to focus on the types of measuring items so that we can examine the students' differences and deficiencies. It can provide feedback for teachers' reflection of students' learning difficulties in specific areas. It can also help them to reflect on their inadequacies which can in turn improve their teaching.
>
> When we have more understanding of students' level of performance, we can offer students more concrete and correct feedback.

Nineteen responses indicated that the Rasch measurement was unworthy of the anticipated effort for a variety of reasons. Whereas teachers' apprehension was the main deterrent at the first impression, analysis of teachers' beliefs about the desirability of this assessment method showed that the strongest hindrance was their suspicion of its effectiveness (theme 8; $n=10$; Table 17.1). However, more teachers thought that the Rasch measurement was worthy of implementation than the number who thought it was not worthwhile. The result was consistent with teachers' first impression on the Rasch model.

In addition to the interviews, teachers were invited to rate the desirability of the Rasch method on a 7-point Likert-type scale (1 = 'not desirable'; 7 = 'highly desirable'). The analysis of their ratings showed a very positive response, with a mean rating of 5.21 ($SD=1.29$; range = 2–7) on the 7-point scale.

17.3.3 Teachers' Attitudes Towards the Feasibility of Implementing the Rasch Method

Teachers' attitudes towards the feasibility of implementing the Rasch method were gauged using both quantitative and qualitative data. Quantitative data was obtained from teachers' rating on the questionnaire, and qualitative data was extracted from analysing the teachers' interview transcripts. Teachers were asked to rate on a

7-point Likert-type scale (1 = 'not feasible at all'; 7 = 'extremely feasible') the feasibility of implementing the Rasch method. Teachers were then interviewed about its perceived feasibility, the kind of support they might need and possible obstacles to its implementation. Analysis of the ratings found a mean rating of 3.75 (SD = 1.45, range = 1–6) on the 7-point scale, suggesting that although the averaged inclination was positive, there was reservation by many teachers regarding the feasibility of implementation.

Reasons behind teachers' reservation on implementing the Rasch model came out clearly in the interviews. Teachers spoke explicitly about their fears and concerns, possible obstacles and supports required. The results are presented in Table 17.2.

Content analysis of the 133 responses from the interview transcripts of 48 participants identified ten themes. These were, in order of importance, (1) demand on human resources, defined as concerns about anticipated extra demand on human resources, workload and time; (2) professional training, defined as expressed needs for professional development on how to interpret the data and outputs from the Rasch analysis; (3) apprehension, defined as teachers' concerns with the power of the Rasch analysis to reveal weak items and poor quality examination papers, which might be threatening to teachers. Other negative emotions were also classified under this theme; (4) effectiveness and applicability of the Rasch model, defined as teachers' perceived strengths of the method in promoting teaching and learning; (5) expert support, defined as expressed needs for a professional expert to undertake the Rasch analysis and interpretation of the results; (6) software features, defined as teachers' suggestions on the enhancement of the computer software (i.e. Winsteps) in order to facilitate their use of the Rasch method; (7) assessment resources, defined as expressed needs for an item bank and databases to reduce teachers' workloads; (8) teacher efficacy, defined as teachers' self-belief in their own efficacy in undertaking the analysis of assessment data and in interpreting outputs from the analysis; (9) computer facilities, defined as expressed needs for adequate and appropriate computer software and hardware facilities to support teachers' use of the Rasch method; and (10) school policy, defined as comments on special school policies that might conflict with the implementation of the Rasch method. Of these themes, the first theme had the most observations ($n = 42$). However, there were signs that some perceived obstacles were interrelated. For example, the heavy workload of teachers (theme 1) together with their perceived low self-efficacy in having the competence to interpret results from the Rasch analysis (theme 8) might make the teachers more cautious about accepting this new method. They also feared that the workload of incompetent colleagues will end up on their shoulders; as one respondent observed (probably from bitter experience), 'It depends on teachers' willingness. If some of them have different ideas, the rest of us will be exhausted by hard work.'

In the study of ChanLin et al. (2006), factors affecting teachers' use of technology in creative teaching could be classified into four categories: environmental, personal, social and curricular issues. Though some factors observed from their study could be found in this study, factors affecting teachers' attitudes towards the implementation of the Rasch measurement could be sorted into different ways. Five categories were

Table 17.2 The teacher's attitudes towards the feasibility of the Rasch measurement

Coding	Sample from transcripts	No. of responses ($n=133$)
Demand on human resource	Definition: Teachers expressed concerns with extra demand on human resources, workload and time Sample: 'I personally think that there will be huge work-load. Even the analysis finished, the meeting time among colleagues is limited. That means we have to share extra-time to review the analysis. It is a burden for teachers'	42
Professional training	Definition: Teachers expressed the need for professional development on how to interpret the data and out puts from the Rasch analysis Sample: 'If training is provided, teachers will have greater interest in the method. In general, I think it will benefit our school'	29
Apprehension	Definition: Teachers expressed the feeling of being threatened as they worried about the power of the Rasch model in revealing weak items and poor quality of examination papers Sample: 'At the beginning, there is no need to post all the data in the public domain. Posting publicly may reveal one's poor item setting skills (or poor quality of the exam paper), colleagues will be threatened and subsequently not willing to use the approach'	14
Effectiveness and applicability	Definition: Teachers expressed their concerns about their colleagues who were able to perceive the strengths of the Rasch measurement Sample: 'Its advantage is that teacher can know students' performance immediately. If teachers see it from the educational point of view, they are willing to spend the time'	13
Expert support	Definition: Teachers expressed the need for a professional expert to analyse the assessment data by the Rasch method as well as report and explain the results to them Sample: 'We need expert's support. The person can analyse the data and explain to us, advice us how to improve it. It will be useful only if we know the "story" behind the data. If we only get the information, it is "dead" – no meaning to us'	9
Software features	Definition: Teachers' suggestions on the enhancement of features in the computer software (i.e. Winsteps) in order to facilitate their use of the Rasch method Sample: 'Adjust the interface of the software to be more user-friendly and simpler, just show the index (figure) which should concern teachers is already adequate'	7
Assessment resources	Definition: Teachers expressed the need for an item bank and databases for their use Sample: 'Web database is very important because it can provide professional support for us, or the database may include some articles, or photo resources'	7

Table 17.3 Identified categories of factors that influence the implementation of the Rasch measurement

Identified categories	Coding	No. of responses ($n = 133$)
Resources ($n = 49$)	Demand on human resources	42
	Assessment resources	7
Professional factors ($n = 38$)	Professional training	29
	Expert support	9
Psychological factors ($n = 33$)	Apprehension	14
	Effectiveness and applicability	13
	Teacher efficacy	6
Technical factors ($n = 10$)	Computerized support	7
	Computer facilities	3
Policy ($n = 3$)	School policy	3
	Educational policy	0

suggested; these are, in order of relative importance, (1) resources, which included the themes of 'demand on human resource' and 'assessment resource'; (2) professional, which included the themes of 'professional training' and 'expert support'; (3) psychological, which included the themes of 'apprehension', 'effectiveness and applicability' and 'teachers' efficacy'; (4) technical, which included the themes of 'computerized support' and 'computer facilities'; and (5) policy, which included the theme of 'school policy'. Though not found in this study, the theme of 'educational policy' could also have been included in this category.

Of these five categories, 'resources' ($n = 49$) was regarded as the most influential in determining the feasibility of implementing the Rasch method, although 'professional' also had significant influence ($n = 38$; see Table 17.3). Both of these categories consisted of factors that could be controlled. Thus, teachers' attitudes towards the implementation of the Rasch measurement might change over time. Had appropriate measures been available, change or constancy of teachers' attitudes could have been investigated.

However, regardless of which factors might influence the implementation of Rasch model, the issues raised by these teachers are possible obstacles to the use of effective feedback from assessment data to inform learning and teaching.

17.3.4 Association Between Feasibility and Desirability of the Rasch Model

To further explore the teachers' attitudes towards the Rasch model, each participant's desirability and feasibility ratings were plotted on a scatter plot. Figure 17.4 showed the distribution of each teacher's ratings. The x-axis gives the desirability rating, and the y-axis gives the feasibility rating. Overlapping points occurred when participants had the same measure. Under such circumstances, the points were slightly

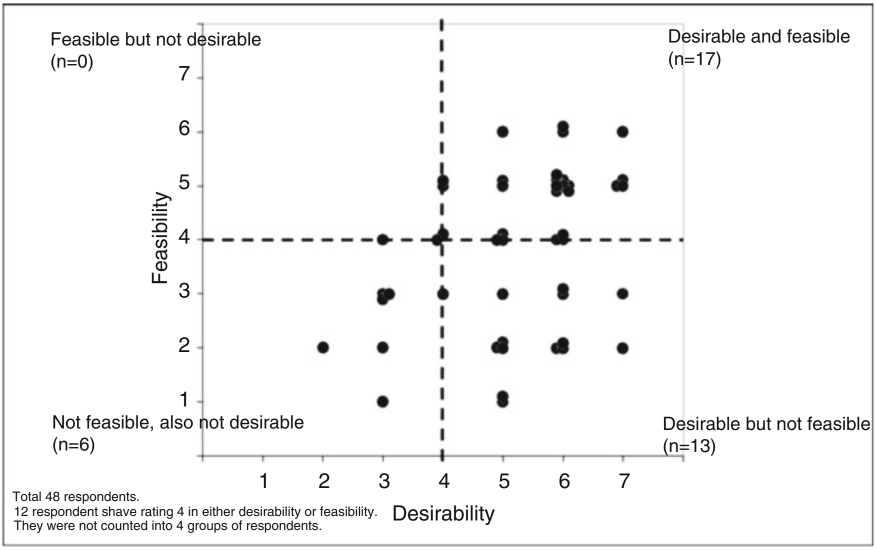


Fig. 17.4 Teachers' ratings on the desirability and feasibility of Rasch Measurement
Notes:
1. Total *n*=48 respondents
2. The desirability scale is a 7-point Likert-type scale (1=not desirable,…, 7=highly desirable)
3. The feasibility scale is a 7-point Likert-type scale (1='not feasible at all',…, 7='extremely feasible')
4. The 12 respondents who gave a rating of 4 for either desirability or feasibility were not included in any of the 4 groups

adjusted in order to display visually all the points without distorting the results or affecting the overall conclusion. Four groups of responses could be broadly identified from the teachers' pairs of ratings. The four response groups are 'feasible but not desirable', 'not feasible and not desirable', 'desirable and feasible' and 'desirable but not feasible', and these are displayed in the four quadrants of the scatter plot in Fig. 17.4. As shown in Fig. 17.4, more responses ($n=17$) fell in the quadrant that said implementation of the Rasch measurement was 'desirable and feasible' than in any other quadrant; 'desirable but not feasible' also held a large number of responses ($n=13$), but no respondent was found in the 'feasible but not desirable' quadrant.

Responses from 'not feasible and not desirable', 'desirable and feasible' and 'desirable but not feasible' groups in Fig. 17.4 were further studied in order to find out the factors contributing to these teachers' attitudes. For the 'desirable and feasible' group, most teachers considered the Rasch measurement a powerful tool for analysing assessment data and that, through this tool, they could understand their students' abilities and individual differences as well as the quality of the assessment items. They indicated that available human resources and in-service professional training would facilitate their implementation of the Rasch method. For example, one of the teachers said, 'Teachers are willing to use the Rasch measurement because it can help them design an assessment paper and find out students' learning problem. Consequently, it can help them enhance students' ability. I think teachers will value it so much' and 'we need support so that we can learn how to [use Winsteps to analyse data].'

For the 'desirable but not feasible' group, the findings highlighted again that the teachers focus mainly on learning and teaching and on the quality of assessment items. Consequently, the Rasch measurement was valued as it gave teachers feedback on the areas that interest them. For example, one of the teachers responded, 'The analysis allows us to know the strength and weakness of students. On the other hand, it can provide feedback on our teaching and on the quality of items.' In terms of feasibility, this group of teachers identified various factors obstructing its implementation. Among them, there was considerable concern about teachers' heavy workloads. For example, one of the teachers pointed out, 'The first thing is to reduce teachers' workload…'. Another teacher also mentioned, 'We're exhausted because of heavy workload. This will affect our resistance to the implementation of the Rasch measurement.' In contrast, teachers of the 'desirable and feasible' group tended to give suggestions rather than to blame factors related to human resources.

Although there was only a small number ($n = 6$) of teachers who considered the Rasch measurement to be both undesirable and not feasible, their concerns should not be ignored. Almost all of them ($n = 5$) admitted that the Rasch method could help them understand their students' abilities and the quality of assessment items. Nevertheless, their attitudes towards the method seemed to be overwhelmed by the psychological factors such as apprehension and self-efficacy. Adjectives such as 'complex', 'difficult', 'worry', 'surprise', 'strange', 'annoying' and 'problematic' were found in their responses. The following response is representative of the concern of this group of teachers:

> After I found that the paper should be improved, I worry about the follow-up works. The point of being not feasible is that it provides much information about the quality of examination paper. However, teachers may not accept it. Woo, my paper should be good. To my surprise, you told me it was bad?

The above analysis highlights teachers' main concerns on what and how assessment data can be used to gauge students' performance and to ascertain the quality of assessment tasks. Of course, in practice, a number of factors affect teachers' actual use of such feedback data.

17.4 Discussions and Suggestions

The results suggested that the teachers tended to value the implementation of Rasch measurement from the perspectives of learning, teaching and assessment. Analysis suggests that teachers in this study realized that the Rasch measurement functioned as a powerful analysis tool of assessment data, one which could provide them with invaluable feedback about both their students' learning and their own teaching. Some of the teachers in the study initially showed resistance to the measurement.

First impressions might have been overwhelmed by affective factors such as apprehension. However, when the teachers did get the opportunity to reflect more deeply, there were fewer responses showing resistance. Instead, more concerns were expressed on the effectiveness of the assessment methods, in particular whether or not the assessment could generate valuable feedback. The primary factor that determined whether the teachers believed it was desirable to implement the Rash measurement was whether or not they believed the analysis could enhance teachers' understanding of students' abilities and individual differences. Such understanding is important because students' learning can be effectively promoted when feedback is targeted at their performance and outcomes. Feedback becomes significant when it yields information that enables teachers to shape teaching and learning (Marriott 2009). The result from the current study was consistent with the findings of Ekiz (2006). In his study, Ekiz (2006) found that the majority of teachers were willing to carry out educational research, and the primary factor motivating them to undertake a research was to do the best for their pupils. However, teachers 'believed that they would have difficulties in time and necessary conditions if they wish to carry out research' (p. 399). The participants in the present study also expressed these difficulties in the interviews, although more emphasis was placed on the heavy workload or human resources, as can be seen in their responses about the feasibility of implementing the Rasch measurement.

The social cognitive theory developed by Bandura (1997) provides a useful framework for understanding teachers' concerns. In this theory, Bandura (1997, p. 2) highlighted self-efficacy as the belief 'in one's capabilities to organize and execute the courses of action required to produce given attainments'. Using this theory, it is expected that the development of self-efficacy for a new technology on assessment for learning will involve teachers considering first of all the likely outcomes to their students and to themselves and the relevance of the innovation, and then teachers will also consider the structural impediments and opportunities (Bruce and Ross 2008; Lieberman and Pointer Mace 2008). Those teachers who perceive more benefits and relevance to student learning of the new technology are more likely to espouse it – but even the most committed teacher will be deterred by the lack of social support, lack of technical support, the pressure of time and the lack of meaning. Lieberman and Pointer Mace (2008, p. 227) have cogently argued for the following conditions, which will enable teachers to engage in sustainable education reform:

> [Teachers] learn through practice (learning as doing), through meaning (learning as intentional), through community (learning as participating and being with others), and through identity (learning as changing who we are). Professional learning so constructed is rooted in the human need to feel a sense of belonging and of making a contribution to a community where experience and knowledge function as part of community property. Teachers' professional development should be refocused on the building of learning communities. (Lieberman and Pointer Mace 2008, p. 227)

Policymakers and school administrators who advocate educational reform must listen to teachers' concerns. If implementation of any reform is to be successful, then it is essential that the problems of heavy workloads and inadequate human resources be addressed. Additional funding, smaller class sizes or external supports might be helpful. Indeed, the Education Bureau (2007) has already realized that time and technical problems are major obstacles to teachers' effective use of information technology in classroom. Thus, their 'priorities are to reduce the burden on teachers in integrating IT into their core activities from lesson planning to assessment of students' (p. 21) as stated in the 'consultation document on the third strategy on information technology in education' (Education Bureau 2007). Nevertheless, the Bureau's focus was on developing digital resources for teachers to use in learning and teaching activities, as well as sharpening 'teachers' IT pedagogical skills' (p. 18). The authors of this current study believe that not enough emphasis was put on developing teachers' professional capacity in using information technology for enhancing assessment for learning. It might be considered as 'the fourth strategy on information technology in education' in Hong Kong. Furthermore, experts from universities still play a significant role in advocating assessment for learning by means of valid and reliable measurement methods accompanied with effective information technology. They can provide professional training for teachers and develop assessment item banks, etc. In short, partnership between schools and universities should continue to develop, and funding from the government is essential to enable this partnership to grow.

17.5 Conclusions

This study set out to investigate teachers' attitudes towards the costs and benefits of using reports generated from Rasch measurement to support assessment for learning. The importance of the study can be seen from recent research (e.g. Priestley 2005; Bruce and Ross 2008; Tierney 2007; van der Schaaf et al. 2008; Wiliam and Thompson 2008) which has shown that success of reform is greatly influenced by the attitudes and values of the front-line practitioners – in this case, teachers. This study found that Hong Kong teachers in general welcomed Rasch measurement as a powerful alternative to the traditional 'total score' method for providing useful and detailed feedback to support and improve student learning. They saw promises in the Rasch method and showed willingness to use the assessment tool as a helpful device. However, some teachers were deterred by contextual and technical problems, including the time required to learn the software and to the analysis, and the difficulties involved in learning and applying the technology. Through sharing in the form of this chapter, we hoped that we could provide a forum for teachers' voices to be heard and, in so doing, contribute to the better understanding of teachers' perceptions, their concerns and their needs for support in implementing assessment for learning.

Acknowledgements We wish to express our sincere appreciation to those who have contributed to this study. Special and deepest thanks are given to Dr Chen Cheng Hui who assisted with the development of the research framework, the coding of the transcripts and who gave us valuable advice. We also wish to express our gratitude to Tung Lung Chu who helped us conduct the interviews and encoded the transcripts. Without their professional expertise and kind assistance, this study would not have come to fruition. Last but not least, we wish to extend our heartfelt thanks to the teachers who took part in the group interviews. Without their support, we would not be able to obtain important information about their attitudes towards the Rasch measurement.

References

Alagumalai, S., Curtis, D. D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Dordrecht/Norwell: Springer-Kluwer.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 357–74.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (2003). In praise of educational research: Formative assessment. *British Educational Research Journal, 29*(5), 623–637.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning. Putting it into practice*. Berkshire: Open University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah: Lawrence Erlbaum.

Bruce, C. D., & Ross, J. A. (2008). A model for increasing reform implementation and teacher efficacy: Teacher peer coaching in grades 3 and 6 mathematics. *Canadian Journal of Education, 31*(2), 346–370.

Callingham, R., & Bond, T. (2006). Editorial: Research in mathematics education and Rasch measurement. *Mathematics Education Research Journal, 18*(2), 1–10.

Campbell, T. (2008). The capacity of instructional technologists to provide systemic support for science education reform. *Teacher Development, 12*(1), 67–83.

Chan, K. K. (2008, Feb). *Curriculum reform: Building on strengths for continuous intensification*. Hong Kong: Education Bureau Hong Kong SAR. Retrieved October 5, 2008, from http://www.edb.gov.hk/index.aspx?nodeid=6400&langno=1

ChanLin, L. J., Hong, J. C., Horng, J. S., Chang, S. H., & Chu, H. C. (2006). Factors influencing technology integration in teaching: A Taiwanese perspective. *Innovations in Education and Teaching International, 43*(1), 57–68.

Curriculum Development Council. (2001). *Learning to learn – The way forward in curriculum development*. Hong Kong: Curriculum Development Council.

Education Bureau. (2007). *Right technology at the right time for the right task. Consultation document on the third strategy on information technology in education*. Hong Kong: Hong Kong Government Printer.

Edwards, P. A., Turner, J. D., & Mokhtari, K. (2008). Balancing the assessment of learning and for learning in support of student literacy achievement. *The Reading Teacher, 61*(8), 682–684.

Ekiz, D. (2006). Primary school teachers' attitudes towards educational research. *Educational Science: Theory & Practice, 6*(2), 395–402.

Ellery, K. (2008). Assessment for learning: A case study using feedback effectively in an essay-style test. *Assessment & Evaluation in Higher Education, 33*(4), 421–429.

Flowers, C., Ahlgrim-Delzell, L., Browder, D., & Spooner, F. (2005). Teachers' perception of alternative assessments. *Research & Practice for Persons with Severe Disabilities, 30*(2), 81–92.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67*, 88–140.

Kessler, G. (2007). Formal and informal CALL preparation and teacher attitude toward technology. *Computer Assisted Language Learning, 20*(2), 173–188.

Klenowski, V. (1998). The use of portfolios for assessment in teacher education: A perspective from Hong Kong. *Asia Pacific Journal of Education, 18*(2), 74–86.

Krueger, R. A. (1998). *Analyzing & reporting focus group results. Focus group kit 6*. Thousand Oaks: Sage.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership, 63*(3), 18–24.

Lieberman, A., & Pointer Mace, D. H. (2008). Teacher learning: The key to educational reform. *Journal of Teacher Education, 59*(3), 226–234.

Lim, C. P., & Chai, C. S. (2008). Teachers' pedagogical beliefs and their planning and conduct of computer-mediated classroom lessons. *British Journal of Educational Technology, 39*(5), 807–828.

Linacre, J. M. (2011). Winsteps (Version 3.72.3) [Computer Software]. Chicago: Winsteps.com

Marriott, P. (2009). Students' evaluation of the use of online summative assessment on an undergraduate financial accounting module. *British Journal of Educational Technology, 40*(2), 237–254.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

McNamee, G. D. & Chen, J. Q. (2005). Dissolving the line between assessment and teaching. *Educational Leadership, 63*(3), 72–77.

McTighe, J, & O'Connor, K. (2005). Seven practices for effective learning. *Educational Leadership, 63*(3), 10–17.

Mok, M. M. C., Gurr, D., Izawa, E., Knipprath, H., Lee, I. H., Mel, M. A., Palmer, T., Shan, W. J., & Zhang, Y. (2003). Quality assurance and school monitoring. In J. P. Keeves & R. Watanabe (Eds.), *International handbook of educational research in the Asia-Pacific region* (Kluwer International Handbooks of Education, Vol. 11, pp. 945–958). Dordrecht: Kluwer Academic.

Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: The students' perspective. *Assessment & Evaluation in Higher Education, 33*(2), 143–154.

Priestley, M. (2005). Making the most of the curriculum review: Some reflections on supporting and sustaining change in schools. *Scottish Educational Review, 37*(1), 29–38.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.

Robertson, I. (2008). Learners' attitudes to wiki technology in problem based, blended learning for vocational teacher education. *Australasian Journal of Educational Technology, 24*(4), 425–441.

Spiropoulou, D., Antonakaki, T., Kontazaki, S., & Bouras, S. (2007). Primary teachers' literacy and attitudes on Education for Sustainable Development. *The Journal of Scientific Educational Technology, 16*, 443–450.

Tierney, R. D. (2007). Changing practices: Influences on classroom assessment. *Assessment in Education, 13*(3), 239–264.

van der Schaaf, M. F., Stokking, K. M., & Verloop, N. (2008). Teacher beliefs and teacher behaviour in portfolio assessment. *Teaching & Teacher Education, 24*(7), 1691–1704.

Vygotsky, L. (1978). In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Watering, G., Gijbels, D., Dochy, F., & Rijt, J. (2008). Students' assessment preferences, perceptions of assessment and their relationships to study results. *High Education, 56*, 645–658.

Wiliam, D., & Thompson, M. (2008). Integrating assessment with learning: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*. New York/London: Lawrence Erlbaum Associates.

Wong, J. L. N. (2006). Control and professional development: Are teachers being deskilled or reskilled within the context of decentralization. *Educational Studies, 32*(1), 17–37.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement. In E. V. Smith, Jr & R. M. Smith (Eds), *Introduction to Rasch measurement* (Chapter 1, pp. 1–24). Chicago: JAM Press.

Chapter 18
Accelerated Approach to Primary School English Education in China: Three Case Studies

George C. Yu

18.1 Introduction

One of the most recommended practices for developing early literacy is storybook reading (IRA and NAEYC 1998; Teale 1987; Pearson et al. 2000). Storybook reading, commonly practiced at home as well as in the early childhood classroom, seems to rely on one form of delivery exclusively: *reading aloud.*

In summarizing a huge body of research, Jim Trelease stated in the 6th edition of his Read-Aloud Handbook (2006):

> Whenever an adult reads to a child, three important things are happening simultaneously and painlessly: (1) a pleasure connection is being made between child and book, (2) both [adult] and child are learning something from the book they're sharing (double learning), and (3) the adult is pouring sound and syllables called words into the child's ear. (p. 33)

He highly recommended storybook reading, stating:

> Books which hold children's attention will garner larger learning benefits, the more interesting the book, the keener the child's attention and the more learning results. (p. 56)

Just as storybook reading provides multiple benefits to the English-speaking children, Segers et al. (2004) in the Netherlands found that listening directly to story reading by the teacher seemed to benefit the immigrant children more than the native group. In that Dutch study, the researchers found that the immigrant students gained more vocabulary listening to the teacher than from the computer whereas

G.C. Yu (✉)
Director and Senior Consultant of the Hong Kong Language and Culture Institute,
Centre for Assessment Research and Development, The Hong Kong Institute
of Education, 23D Tower 1, Harbour Place, 8 Oi King Street, Kowloon,
Tai Po, Hong Kong
e-mail: georgeyu@ied.edu.hk or gyu5279@gmail.com

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_18, © Springer Science+Business Media Dordrecht 2013

the gain for native-speaking children remained the same. There were two possible contributing factors:

1. The teacher tends to be more adaptive and friendly than a computer software.
2. The teacher comes to the aid for the children more quickly and in a more appropriate fashion with necessary elaboration and explanation based on the children's facial expressions and/or body language.

According to Mandel et al. (2002), children learning a second language follow a very similar process in which they acquire their first language. While humans are hardwired for picking up any oral language quite naturally from their early social interactions (Chomsky 1977), it is quite a different approach when it comes to the written language which is also called "decontextualized speech" where the sender and receiver of a written message do not meet at the same time and space (Roskos et al. 2003). This challenge in early literacy development calls for explicit and direct instruction (Strickland et al. 2004).

18.1.1 ELL/ESL Reading Model

Two important studies (Clarke and Silberstein 1977; Coady 1979) approached reading instruction from a psycholinguistic model. They both looked at reading as an active process of comprehending and ELL students needed to receive among other things instruction on reading strategies so as to make their English reading more efficient. Helping students to identify their goals and strategies became an important integral part of good second language instruction (i.e., Anderson 2003).

Metacognitive knowledge and skills monitoring is one of the key components of reading skills. Metacognitive knowledge may be defined as knowledge about cognition and self-regulation of cognition (Brown et al. 1986). Knowledge about cognition, including knowledge about language, involves recognizing patterns of structure and organization and using appropriate strategies to achieve specific goals (e.g., decoding words, comprehending texts, recalling information). Self-regulation strategies would include planning ahead, testing self-comprehension, checking effectiveness of strategies being used, revising strategies being used, and so on.

18.1.2 Self-Directed Learning Model

In summing up the research on self-regulated/self-directed learning for the past two decades, researchers Schunk and Zimmerman (in 2007) and Schmitz and Wiese (in 2006) have all reported strong correlation between student academic progress and the use of targeted self-regulatory processes. Significant improvement in academic performance was reported due to SRL/SDL training.

The project teachers received targeted training on theories and practices of self-directed learning including the key role of metacognition in motivating students.

18.1.3 Research Questions

Explicit teaching of reading strategies (Strickland et al. 2004) and supportive scaffolding (Gallimore and Tharp 1990) techniques are found to be particularly effective for reading comprehension. Using explicit or direct instructional approaches, teachers can demonstrate and model to their students how use of reading strategies can enhance reading effectiveness. The steps of explicit instruction typically include direct explanation, teacher modeling ("thinking aloud"), guided practice, and application (Adler 2001). On the other hand, supportive scaffolding fosters effective teaching and learning by creating multiple zones of proximal development. In this context, the form of teaching becomes dialogue which integrates listening, speaking, reading, and writing.

This project set out to investigate the following questions:

(a) How well English reading strategies would work to accelerate the English reading achievement among Chinese primary school students
(b) What impact the self-directed learning approach would play out on the Chinese primary students
(c) How other major factors (i.e., instructional time, resources, parent support, etc.) would impact on student English reading achievement

18.2 An Overview of Primary English Education in China

18.2.1 Primary English as State Mandate

Starting from 2001, China has made English instruction compulsory for grades 3 and up across the country. While the decision might appear made in haste with little preparation in teacher training and instructional materials and little room for consultation (Hu 2007), it nonetheless reflected the urgent need at the top policy level to raise the English literacy level among the country's next generation. No formal English instruction for primary pupils ever existed until the turn of the century (Lam 2002). At best was an assortment of ad hoc interest-based classes in large cities. The situation in China's landlocked interior and western regions was farther removed. Thanks to this mandate, the percentage of primary school children having access to English increased dramatically (Chen 2008):

Year	Percentage of primary students receiving English instruction (%)
2000	9.70
2002	22.10
2004	35.10
2006	60.50

18.2.2 Quality vs. Quantity

However, this quantitative increase (albeit multifold) didn't guarantee program effectiveness. Given the general low level of China's primary education infrastructure and resources investment, it would take time for the mandate to translate into effective programs across the whole swath of China. Among the quality issues plaguing the mandate is the question of teacher preparation. In contrast to student increases, the percentage of college-trained primary English teachers only went from 5.48% in 2000 to 12.65% for 2006. By the official standards, only 26.78% of the total primary English teaching force was considered qualified.[1] The country just can't produce qualified English teachers fast enough to match the rising demands.

While the shortage of qualified primary English teachers will persist for years to come, another problem has been a lack of English instructional materials (Hu 2004) suitable for primary students in China. These two factors combined have made the implementation of the mandate difficult and ineffective. Recently, the situation regarding the availability of resources suitable for primary students has improved steadily as more child-friendly materials have become available. But parents and educators continue to decry the lack of knowledgeable school personnel to organize these materials into effective, cohesive, and well-articulated courses and programs at the nation's primary schools.

18.2.3 School Curriculum and Setting

The purpose of the curriculum at most primary schools is to provide exploratory opportunities for students to establish basic understanding of English pronunciation, the alphabet, and most common daily utterances. This is pretty much in line with the primary English education standards and goals articulated by the Ministry of Education. However, there was clearly a gap between the stated goals and rationale on one hand and suitable teaching materials, teacher preparation, and pedagogical approach on the other. The prevalent instructional approach at most schools throughout the country is still that of teaching English as foreign language (EFL) with heavy reliance on textbooks and grammar-translation pedagogy (Huang and Xu 1999; Brown 2006).

[1] Based on data provided by China Basic Foreign Language Education Research & Training Centre, 2008.

18.3 The Project Implementation

18.3.1 Participants

The first cohort of participants was 97 grade 5 students from Da Guan Primary School (DG). Their English teacher was Ms. Jie Gao. The second group was 112 grade 4 students from Yinma Jingxiang Primary School (YMJX). Ms. Xiufeng Zheng was their English teacher. Both schools are located in the city of Hangzhou.

Meanwhile, all 460 grades 4–6 students plus 6 English teachers at Guangya Primary School in Guangzhou joined the project.

The amount of instructional time allocated for English at the three schools was consistent with the national norm: students in grades 3 and above (ages 9–11) typically receive two or three 40-min English lessons per week (Nunan 2003; Cortazzi and Jin 1996).

Each of the three project schools designated one teacher to coordinate the project. It turned out that all three were also acting as the English subject panel heads in their respective schools.

18.3.2 Treatment

18.3.2.1 Targeted Professional Development for the Project Teachers

First and foremost, the project put an emphasis on the professional development of the teachers. Presentations and workshops were provided on site, and a steady stream of consultation sessions was maintained via internet and long distance calls. The teachers and their principals were also invited to participate in an international assessment conference organized by the research team.

The contents delivered through professional development covered many topics but followed three major threads:

Further Enhancement of the Teachers' Expertise to Conduct English Reading Class Properly

Given the general composition of college English courses and the dominant pedagogy in China, very few graduates would be adequately prepared to teach a primary English reading class. Most teachers hungered for further studies. Therefore, it was a huge motivation for the project teachers to see this need get addressed. Here is a list of key topics covered in the training:

- Development of English literacy (from phonics, vocabulary, fluency in decoding to comprehension, etc.)
- Creative use of English sight words and word puzzles
- Storybook reading (e.g., reading aloud, guided reading, shared reading)
- Introduction of reading strategies and good classroom practices

Introduction to Formative Assessment and Rasch Measurement

It was the firm belief among the research team members that all would be lost should the teacher's mindset fail to shift to recognize the needs of their students. It was really encouraging to see that they were ready to break away from the past to embrace the concept of using assessment to guide teaching and learning. They realized that meaningful exposure to the target language is the key to maintaining student interest. Here is a list of topics explored:

- Introduction to assessment for learning
- Introduction to Rasch model and developmental scales of measurement
- Questioning techniques and learner profiling

Self-Directed Learning to Accelerate English Reading

While the concept of using self-directed learning strategies to motivate the learner is sound, its practice was viewed intimidating by the teachers. It differed markedly from the state-mandated curriculum and test-driven approach. A deeper and more holistic understanding of the learner was called for to appreciate the great variety of skills they already carry. Here are some examples introduced:

- An English Reading Log
- How to model and demonstrate the use of reading strategies
- Understanding the developmental stages of self-directed learning

18.3.2.2 Increased Amount of English Storybook Reading Among Students

While the amount of English story book reading was extremely limited within the fixed curriculum, the project teachers explored the following avenues to boost English literacy exposure:

- A half-hour English Salon at noon break to allow students in grades 3–6 to browse and read English story books every day
- After-school activities (i.e., English Reading Club, English Reading Corner, reading competitions, etc.) to boost exposure
- Parents were encouraged to create additional reading opportunities at home.

Thanks to these efforts, the amount of English reading by the students went up dramatically. On average, they read 2–3 story books every month.

18.3.2.3 English Reading Log

A special reading log with imbedded self-directed learning features was introduced to the project teachers who in turn modified it for their students. Throughout the 1-year project, nearly all students finished more than 20 log entries. While the students went through a learning curve to handle the log properly, by midpoint of the project, 80–95% could manage entries independently.

18.3.2.4 Reading Strategies, Phonics, and Sight Words

A set of English reading strategies were incorporated into the reading log. The teachers were trained to introduce to students, modeling only a few strategies each time.

Some basic phonic techniques together with the sight word lists were incorporated into the daily teaching. The teachers became more focused on addressing the primary link between meaning and sound so that students become more efficient in decoding new vocabularies.

18.3.2.5 English Reading Tips for Parents

Sound English reading tips were introduced to the parents of the target student population. Homebound flyers and school newsletters were used for disseminating useful information to parents.

18.3.2.6 English Storybooks as Seeds

A few hundred English storybooks donated by a Hong Kong publisher were presented to the schools to buttress up their collection. The US-based Lexile Framework was also introduced to allow at both conceptual and practical level matching reading materials to reading proficiency of the student.

18.4 The Impact of the Project

18.4.1 Physical Environment

The changes in the physical environment of the project schools are the most visible. Thanks to the project activities, English has become more integrated in the daily school life. Let us look at some examples.

18.4.1.1 The English Reading Corner in Da Guan Primary School

Using the storybooks donated from Hong Kong as seeds, the DG English teacher secured the permission from her principal to set up an English Reading Corner. Students can now stop by to browse through the books on display during their recess or lunch break.

The principal also worked with the English teacher and school librarian to launch an English Book Donation. The school then would appeal to the parents to let their children participate in the Book Swap Club by contributing used books to school. Thus, the collection of English books can be expected to grow quickly. Similar book drives were also observed at the other two project schools.

18.4.1.2 The English-Speaking Staircase at YMJX Primary School

In order to increase the students' exposure to English, the teachers and the principal at this school hit upon a brilliant idea. When it was time to repaint the stairs in the school, they engineered to have English painted on the vertical side of each step. They selected and compiled various simple short sentences, catch phrases, or slogans as extra contents for learning English. All of a sudden when the paint finally dried, the routine to climb the stairs gained an added dimension of fun and intellectual challenge. Quotes from famous figures or literary works were later added on walls and hallways. A subtle message to the students would be: Look, Kido. English is just a good friend around you.

18.4.1.3 The English Billboard Gardens at Guangya

There was a tremendous amount of information to be communicated to the school community. Photos of student performances, awards students won at various English speech/drama competitions, notices and announcement for English club activities, plus model student English Reading Log entries, etc., all needed space to present to everyone in school. Their colorful and artistic display boards were the solution.

18.4.1.4 Bilingual School Newsletters Going Home at All Three Schools

In a similar vein, the need occurred to relay messages to parents at home using both English and Chinese. Simple English learning tips, games, puzzles, or a homework assignment are now routinely sent home with proper translation in Chinese. It helps to secure the legitimacy of English for daily communication from the eye of students and teachers. This has since become a routine at all three project schools.

18.4.2 Library English Collection

At the start of the project, a cursory look at the school library would quickly reveal that there was little or nothing to speak of in terms of any English book collection. In most cases one would run into a collection of old classical novels written by Charles Dickens or Jack London together with textbooks collecting dust on bookshelves.

Thanks to the project, all three schools have increased budget to boost the English collection. Given the newly acquired expertise on what books suit their students, the school librarian now knows what to acquire and what to skip over on the basis of the student fluency levels and interests. Thanks to the initiatives from the English teachers, many child-friendly stories were quickly compiled and added.

18.4.3 The Teachers' Professional Growth and Subject Expertise

One of the most rewarding results from the project was to see the tremendous professional growth on the part of the teachers. The key contributing factor for securing this steep learning curve was a good match between the needs and challenges they faced everyday and the targeted professional development provided by the project.

The teachers were highly motivated because they regarded this project as a golden opportunity to upgrade their expertise. They worked very hard on the project devoting often extra time and care. They also took full advantage of all the valuable learning opportunities the project presented. Their learning curve was rather steep and professional growth most impressive. Their enhanced capacity to teach English reading has since been recognized by their school principals and officials from the local governments. They have been called upon to conduct model lessons, make presentations on pedagogy, or conduct training workshops for their colleagues in the field.

18.4.4 Impressive Innovation and Creativity

The researchers were very impressed by the high energy and creative spirit of the teachers. They didn't just borrow ideas wholesale. Since the bottom line in the new approach is student-centered, then they must always ask the question "Is it right for my students?"

One good example is the adaptation of the "official version" of the English Reading Log. The teachers found it to be unwieldy for implementation. The solution was a much streamlined version containing all essential ingredients printed with bilingual instructions.

18.4.5 Personal Initiatives and Commitments

Inspired by the project, the teachers went beyond the realm and scope of their normal duties. Their motivation came from their early recognition that they could access valuable research-based theories and practices which would enable them to do their job better. By working closely with the researchers, they hoped to further enhance their capacity on how to facilitate early literacy development. So they let go whatever concerns and reservations they might have and took off with the project in big strides.

18.4.5.1 At Da Guan

Ms. Gao initiated a series of activities to provide more opportunities for English reading at her school. Here is a partial list of those activities she started:

- A half-hour Noontime English Reading Salon during which she herself would lead students per grade level to check out books from the library to read

- An English Reading Growth Exhibit to showcase examples of English Reading Log entries completed by students
- Sharing sessions to encourage students to exchange ideas and experiences regarding their own development of English reading skills and strategies
- English Reading Contest to further jazz up the reading boom at school
- Various special columns in the school newsletter to feature student selected readings and their reflections

18.4.5.2 At YMJX

When Ms Zheng set up an English Reading Club, almost one third of her grade 5 students signed up. She utilized local materials to expand the exposure to English for her students through games, puzzles, and songs. She also introduced her club members to reading English storybooks for leisure and personal enjoyment. Whenever appropriate, Ms. Zheng also demonstrated for her students how and when some of the key English reading strategies could be used to boost for better understanding. She even made extra effort to get the parents more involved by sending home reading tips to boost parents' understanding on how to assist their children.

18.4.5.3 At Guangya

Ms Peng and her colleagues launched a number of extracurricular events to promote English literacy and celebrate efforts and achievements by the students. Here is look at three such events:

- Better Late Than Never, a Short English Drama
  On December 30, 2008, the Premier of Guangzhou Citywide Children's English Drama Festival was held to showcase English teaching and learning at various primary schools. Guangya became one of the 17 schools featured at this festival.

  The Guangya teachers wrote a play in English entitled "Better Late Than Never" based on the Chinese proverb with the same title. The play involved about 20 students on stage and many more off stage. Thanks to the great display of stage sets, solid preparation, original play, creative costume, and best of all the fluent English conversation onstage, the Guangya team received the "second place prize."
- English Morning Post
  Student English Morning Post was another way Guangya teachers created to boost student exposure to English on a daily basis. In anticipation of Beijing Olympics to be hosted by China in the month of August 2008, the Morning Post launch a contest in April to feature posters designed and handwritten by students in celebration of the century-old aspiration of the Chinese people to host the Olympic Games. A great many posters dripping with their creativity, enthusiasm, and desire to master the English language were submitted to the contest.

- English Speech Contest

  When an English Speech Contest was launched by a language training organization in the city, the Guangya teachers saw this as another good opportunity to expand and enrich their students' after-school English learning activities. Based on student self-interest and teachers' recommendations, ten students volunteered to join the contest. On March 29, 2008 the stage was set for all the participants.

  According to the contest rules, students were divided into groups of five. Each group then drew a lot to determine which story to study and commit to memory. The best students would be chosen based on their performance of narrating the story with right pronunciation, emotion, and without having to look at the written version. While some of the students were seriously challenged by picking the longer stories to prepare, none backed out of the contest. Under the teachers' guidance, some of the ten delegates from Guangya even made into the second run.

18.4.6 Teachers' Growth and Career Advancement in Their Own Words

As part of the project requirements, the three lead teachers were encouraged to periodically reflect on their own learning and challenges as the project took its course. The essence of this reflective thinking was captured in their reflection journals. Here are some excerpts taken from their written reflection. Since the originals were written in Chinese, here are the English versions.

18.4.6.1 A Reflection on Teaching Module #3 by Ms. Peng from Guangya

My Reflections on Teaching This Module

The topic of the current module is "plants." The class already learned some plant names in addition to the vocabulary for park activities and sentence patterns. Since this class aims at integrating the skills, I tapped into my research on English reading in coming up with the lesson plan. To target student self-directed learning needs, I reorganized the contents of this module and wrote up a new plan for the teaching procedure. I decided to put emphasis on English reading and writing.

The objective of the lesson is to train students to make suggestions in correct English based on their grasp of vocabulary and sentence patterns regarding the park. Students will work in small groups sharing how to use reading strategies to decode the reading text and drafting rules for proper behavior in public places. Consequently, students will be more cooperative and team-oriented. They can also develop their self-reflective capacity so as to assess each other and selves more accurately.

> After the class, I reflected on my teaching:
>
> I see need for further improvement in the following aspects:

- To further cultivate the group culture, I realize that more regular training is necessary for the leaders of those small groups.

- To further boost student confidence, more channels have to be used to assess student performance. Formative assessment and summative assessment should be both applied. Sometimes groups should be assessed to encourage more cooperative learning between more advanced students and those with weaknesses.
- Enriching textbook with more learning opportunities for English reading and writing. Based on student interest and current level, more reading materials need to be added to allow more practice. To further boost student knowledge base on the chosen topic, appropriate videos can also be introduced so as to expose the students to new vocabulary and language materials.

In summary, I really feel honored to participate in the project. This provides good opportunities to further elevate my learning and professional development.

18.4.6.2 Reflective Journal by Ms. Gao from Da Guan

The Current Status of Da Guan English Reading Project and Future Expectations (October 2009)

It has been more than one semester since the project started. Almost all the students have shown some quite visible progress in the following aspects:

- Visible Progress
  - Their reading strategies are applied more effectively.
  - Rapid recognition of old vocabulary.
  - Integrating new learning more effectively.
  - Ability to gather information from reading.
  - Ability to pick reading strategies for solving problems in reading.
  - Enhanced oral skill to retell stories.
  - More positive learning attitude.

- Sound Learning Habits Being Shaped
  There is a saying, “a good habit is an asset for life.” It is not too difficult to get an English language learner to sit down and read a story in English. But doing it as a habit requires cultivation. If the book becomes too difficult, the student naturally will back away. When this happens, it is the duty of us English teachers to encourage them not to give up.

  We avoid taking any formal assessment for the extra reading activities. As long as the student can identify anything, be it a new word, a sentence, or some new information from the story, they can get a prize. Gradually my students can sustain their interest and find that reading in English is not that difficult after all. The amount of reading just multiplies. Once you set up a momentum, every student wants to join. We set up English Reading Time on every Monday, Wednesday, and Friday during which not only students have sufficient time to read but also have opportunities to share with one and another. Once a week, everyone will submit one reading report for parent review. It has since become a routine to include English reading as part of everyday life for these students.

- Enriched Campus Activities

  The English Reading Project has added immensely to the campus life for our students. Using it as a platform, we launched a series of events called, “English Reading on the Raft,” bundling a selection of good storybooks like a raft passing through different classes. Along the way, more good reading logs were added. Another event organized was “the English Reading Contest.” Judging by the grades from the final exams, the average scores of my project students demonstrated a clear advantage of 3–4 points over the comparison group. Periodically students were encouraged to create English posters. They had to do everything by themselves: design, draw, and write all by hand.

  In the future, we plan to expand the project to all our grades, allowing everyone at school to benefit from this approach.

18.4.6.3 Reflective Journal by Ms. Zheng from YMJX

The English Reading Project: Current Status and Future Expectations

- Current Status of the English Reading Project

  Thanks to the collaboration with The Hong Kong Institute of Education, we have scientifically assessed the current English reading level of our students and their relevant experience. We are now trying to select reading materials to fit their ability level.

  Based on the prior experience in reading, very few students have ever been taught on how to use reading strategies. So more or less they headed into English reading blindfolded. For example, previously, when a student ran into a new word he/she didn’t know in reading, they got stuck there. Now everyone in the class has learned how to initiate their own ideas and apply their own methods to attack those new vocabularies. They would look them up in the dictionary, trying to find clues in the context, or figuring out the meaning from their lexical composition.

  When the students are reading a story, they are also at the same time completing a reading log which includes the following:

  - Learning objectives
  - Learning plan
  - Which strategies to apply
  - Contents of the book (e.g., book title, most appealing part, my learning)
  - My reflection
  - The learning strategies I have used

  After a period of experimentation, the target students involved in the Reading Project overall have shown improvement in English. Some of them and their parents have started to acquire English books meeting their own needs from the bookstores or online. In view of these changes and collaborative interactions among the teachers, students, and parents, I see more changes at a deep level.

The project has helped my students to expand their horizons. Now, they have learned more vocabulary and expressions, gained more knowledge about the Western culture, and improved their communication skills. One thing is of lasting significance: my students have become more interested in English reading and developed ability to appreciate the magic of the language. This will motivate them to develop sound learning habits which will benefit them throughout their lives.

- Ideas and Expectations for Future

  The researchers have helped to open up many windows for us to learn about the cutting-edge research and literature in the field.

  While students have made progress, I still feel that I am not well equipped to judge student ability levels and index reading materials. Much more is still to be learned to further develop student reading capacity and assess them scientifically.

18.5 Accelerated Student Learning Results

So far there has been enough evidence to indicate the tremendous learning and accomplishments on the part of the project teaching staff. Student achievement is also in plain sight. Given the primary purpose of this project being "to accelerate student English reading," the picture of the project impact will not be complete without a detailed look at the learning achievement of the target students.

18.5.1 Measures

18.5.1.1 Pretest and Posttest of Student English Reading Comprehension

The target students' English proficiency was measured using items from HKIED CARD's English Reading Comprehension Scale which is a vertical scale comprising items developed by CARD to measure the English reading proficiency of second language users between primary 2 and secondary 4 levels. These items have been calibrated using the Rasch model (Bond and Fox 2007).

18.5.1.2 Student Self-Directed English Reading Attitude Surveys

The students' competencies of self-directed learning were gauged using a questionnaire which comprised two scales each with five 4-point Likert-type items. The first scale measures students' *self-assessment* in reading and the second measures students' *self-regulation* in reading. These two scales have been validated for use with secondary students before (Mok et al. 2006).

Table 18.1 Da Guan students' performance in assessments 1 (pretest) and 3 (posttest)

	Frequency			Percentage (%)		
Class	Improved (+)	No significant change	Decreased (−)	Improved (+)	No significant change	Decreased (−)
One	11	22	0	33.3	66.7	0
Two	14	14	4	43.8	43.8	12.5
Three	7	19	4	23.3	63.3	13.3
Group average	32	55	8	33.7	57.9	8.4

18.5.2 Quantitative Data Results

18.5.2.1 Da Guan Student Assessment Data

The cohorts at Da Guan took the pretest (assessment 1) and posttest (assessment 3) in May 2008 and May 2009, respectively. In between, they also took the midcourse test (assessment 2) in December 2008. The data was collected and analyzed at CARD. Here are the results:

Table 18.1 provides a summary of students' English reading ability changes across the pre- and post-assessments for all classes. The second to fourth columns present the frequencies of "improved (+)," "no significant change," and "decreased (−)," respectively. The fifth to seventh columns present the percentages of those three categories.

As indicated in this table, although 57.9% of students did not show significant change in their ability across the two assessments, the number of "improved" students far outnumbered those who showed "decreased" ability by 4 to 1. More than one third (33.7%) of all the students have made significant improvement in the project period as measured by the pre- and posttests.

However, further examination of the change pattern has yielded some more meaningful insight. Figure 18.1 provides data on the relationship between students' ability change and students' ability baseline. In this figure, the x-axis is students' ability baseline (i.e., students' ability measures in assessment 1) and the y-axis is their ability change across the two assessments (i.e., assessment 3 measures minus assessment 1 measures). We can see a clear tread in this figure that the students with low baseline ability in assessment 1 were more likely to gain improvement in assessment 3 (positive change). The students with high baseline ability in assessment 1 were less likely to gain improvement in assessment 3.

This change pattern tells us that the English Reading Project had different impact on different groups of students. For students who had relatively low ability measures in assessment 1, the project seemed to benefit more and lead to more positive gains. For more capable students who had relatively high ability measures in assessment 1, they seemed NOT to have gained as much.

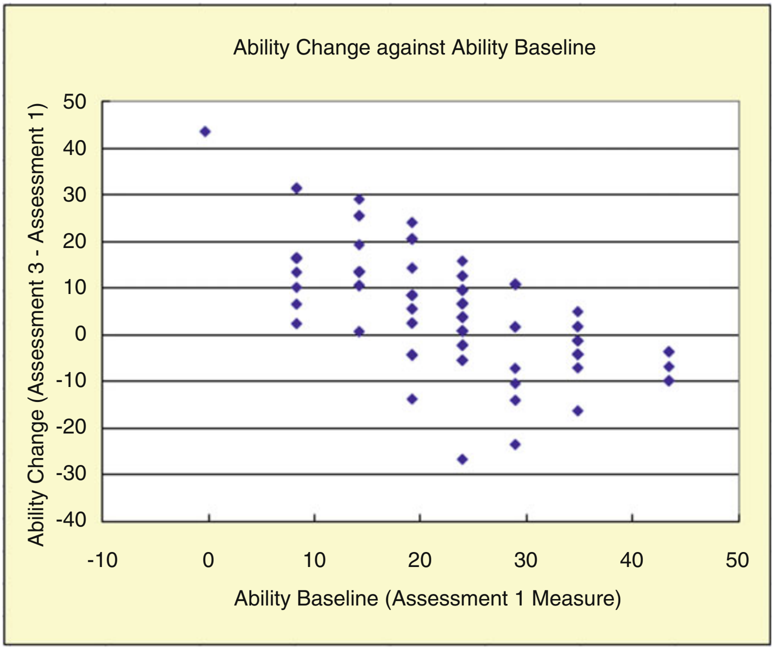


Fig. 18.1 Da Guan student ability change pattern from pretest to posttest

Table 18.2 YMJX students' performance in assessments 1 (pretest) and 3 (posttest)

	Frequency			Percentage (%)		
Class	Improved (+)	No significant change	Decreased (−)	Improved (+)	No significant change	Decreased (−)
One	7	20	2	24.1	69.0	6.9
Two	10	14	3	37.0	51.9	11.1
Three	10	15	2	37.0	55.6	7.4
Four	17	10	1	60.7	35.7	3.6
Group average	44	59	8	39.6	53.2	7.2

18.5.2.2 YMJX Student Assessment Data

The YMJX cohort followed the same schedule as Da Guan. Here are the results (Table 18.2).

Based on the data above, the number of students who showed "improved" results during the project period outnumbered those whose results "decreased" by more than 5–1. Almost 40% (i.e., 39.6%) of the total students made significant improvement in English reading ability as measured by assessment 1 and assessment 3 (Fig. 18.2).

Again, the data on the change pattern seemed to indicate that the project has seemed to benefit the less able students more than it did the more able students.

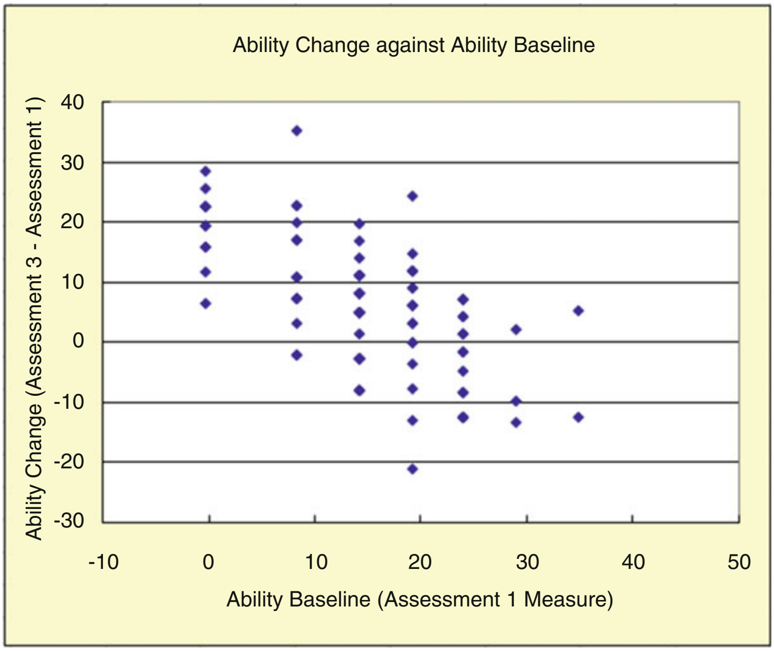


Fig. 18.2 YMJX student ability change pattern from pretest to posttest

Here are some general conclusions that can be drawn from the data above:

- Overall, English Reading Comprehension level increased significantly between pretest (assessment 1) and posttest (assessment 3) for a large segment of the target students at both Hangzhou schools.
- The data revealed that students which scored in the lowest quartile for the pretest seemed to gain most from the project activities as shown by the posttest results.
- Students who scored above average for assessment 1 seemed to gain less or make no significant progress between the two assessments.

(Due to the scheduling difficulties at Guangya Primary School, their students took the assessment only once. Therefore, no comparison data was generated.)

18.5.3 Qualitative Data Results

As part of the Self-directed Reading Project, a student reading attitude survey "My Self-directed Reading Experience" was taken by the two cohorts in Hangzhou at the beginning and end of the project. The survey aimed to uncover students' self-reflective capacity regarding reading English materials and track their metacognitive development.

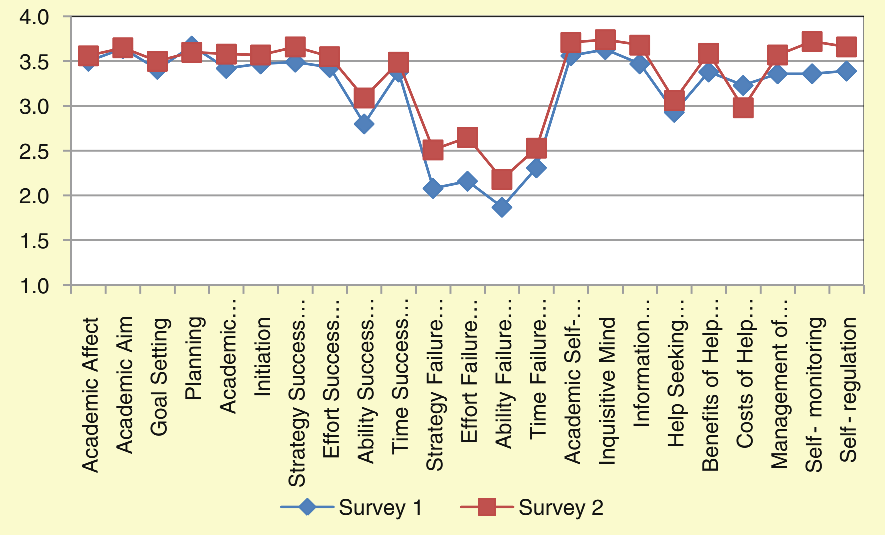


Fig. 18.3 Students' performances in survey 1 and survey 2

The questionnaire attempted to measure different components of self-directed reading. Each subscale contains four to ten Likert-type items. The response scale for the items was a 4-point Likert scale coded as 1: strongly disagree/never, 2: disagree/seldom, 3: agree/sometimes, and 4: strongly agree/often. All items are positively worded except the subscale *Costs of Help Seeking*. For positively worded items, the coding is such that higher scores represent greater inclination than lower scores; for negatively worded items, the coding is reversed.

18.5.3.1 The Da Guan Results

Comparison of Performance in Subscales Between Survey 1 and Survey 2

A total of 97 students participated in survey 1 and 96 students participated in survey 2. Data from both surveys is presented in Fig. 18.1. For each subscale, only those students who had scores in both survey 1 and survey 2 were included in the comparison because paired-samples *t*-test was carried out to investigate the differences (Fig. 18.3).

The comparison between survey 1 and survey 2 revealed that, among the 23 subscales, students had higher mean ratings in survey 2 than that in survey 1 for 21 subscales, and the differences are statistically significant for 11 subscales (academic motivation, strategy success attribution, ability success attribution, strategy failure attribution, effort failure attribution, ability failure attribution, information processing, benefits of help seeking, management of reading environment, self-monitoring, and self-regulation). Students had lower mean ratings in survey 2 than that in survey 1 for 2 subscales (planning and costs of help seeking) and the differences are not statistically significant.

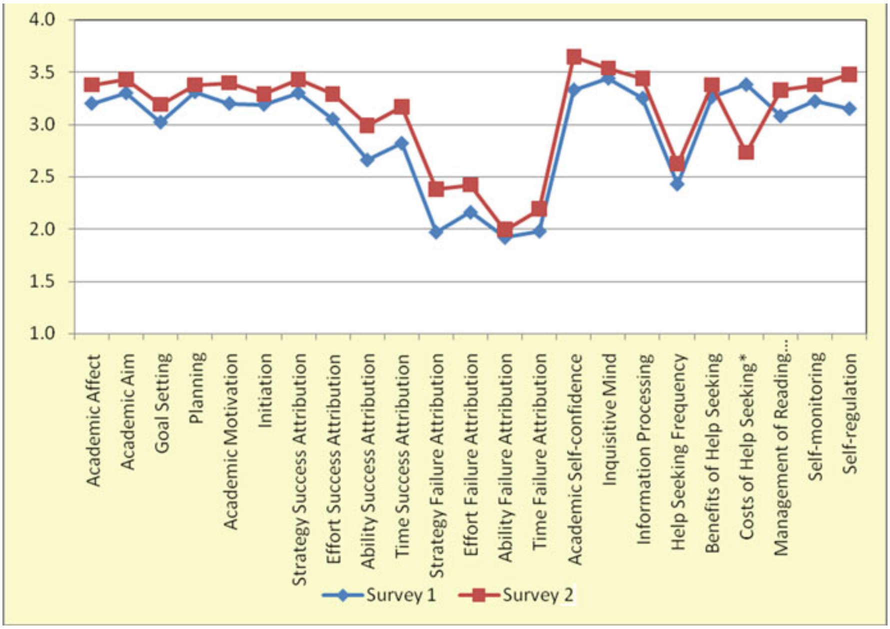


Fig. 18.4 Students' performances in survey 1 and survey 2

18.5.3.2 The YMJX Results

Very similar results were observed below for the students at YMJX Primary School (Fig. 18.4).

The comparison between survey 1 and survey 2 reveals an identical picture as captured earlier for the Da Guan group.

18.5.3.3 The Guangya Results

All 234 students from grade 5 took the survey only once in the early spring of 2009. Therefore, no comparison data was available.

18.6 Observations and Conclusions

18.6.1 The Target English Teachers

It seems obvious based on the three case studies cited above that frontline primary English teachers in China can benefit from the site-based professional development activities this project initiated. The targeted training on self-directed English reading strategies has led to salient improvement in teaching methodologies which in turn have enhanced students' English reading achievement.

Caught between the state mandate for primary English instruction and the lack of readiness at the classroom level, the English teachers would be the first one to spot any viable opportunity to advance their competency. When this project came along, it was like quenching a severe thirst the way the teachers volunteered and soaked everything in. There was no complaint about taking up extra duties or spending extra hours at school for the project.

Given the success of this project in boosting student English reading achievement, more English teachers from primary schools should be able to access similar regiment of professional enhancement. While the services under this project were underwritten by the donors in Hong Kong and provided to the project schools free of charge, this cannot be expected to continue. Education authorities at all levels must make budget available so that the sound practices and lessons proven effective under the project can be duplicated on a much larger scale to benefit more frontline teachers.

18.6.2 Primary Students Learning English

Thanks to the professional dedication and personal creativity plus the targeted training under the project, the three teachers have produced very impressive results in the learning of their students.

Consequently, student experience inside the classroom and outside has undergone a huge transformation. The amount of student-centered, inquiry-based activities has increased to mobilize and stretch the metacognitive capacity of the students. The students become more empowered to initiate and regulate their own learning activities. Using the reading log and extra reading time, they have increased their exposure to English many folds. As a result, English is no longer strange, mysterious, and fearsome.

One of the most intriguing findings under this project is that those less able students as measured by the pretest seemed to have gained most progress in comparison to those more able students. It is hypothesized that the reasons for the rapid improvement on the part of the former group are multiple:

- The child-friendly story reading mode, i.e., "reading aloud," stands in contrast to the textbook-driven teaching. It has helped to break down the psychological barrier to the target language and made the exposure more intimate and relevant to the child, thus mimicking the "natural order" (Dulay and Burt 1974; Fathman 1975; Makino 1980; Krashen 1988).
- The explicit modeling by the trained teachers using appropriate reading strategies while doing pre-reading, during reading and post-reading activities has led to better comprehension of the contents and skills improvement.
- The targeted training on the word-level decoding techniques has assisted those "less able" students to close up gaps and catch up rapidly.

However, it would take more research to account for the lack of significant improvement on the part of the better-than-average students (as measured by the pretest). The following hypotheses seem worthy of further studying:

- While word-level decoding is the primary concern for beginning readers, text/contextual comprehension are a far more complex process.
- In order to comprehend longer and more complex passages, more skills have to be mobilized, such as a more fluent reading speed backed by a larger vocabulary, required knowledge of English grammar, and relevant content knowledge in order to make sense of what is being read.
- All these skills require time to evolve and mature. Not least of which are increased language exposure and meaningful input. To sustain such a rate of increase in terms of both quantity and quality (or intensity) is no easy matter.

18.6.3 Primary School Community

It is so encouraging to see that at all three project schools, English has become such a well-trenched subject, enjoying its unique status among other subject matters. If one suspects it to be slightly more prestigious, it is only because of the attention and transformations brought along by the reading project. It authenticated its teaching practices using the most advanced research from the field. It helped to legitimize English as a real tool for communication in school. Henceforth, English posters, wall exhibits, newsletters, and public announcement, all have become avenues for learning and practicing English. After-school activities become enriched by a whole assortment of English events, including English club, English speech competitions, English dramas, and others.

Parents who used to stand on the sideline now regularly receive school information including how to support their kids in learning English more effectively. More books are donated to the school library to boost its English collection. It is indispensible to have the parents included as your partners to develop the literacy skills of the students.

References

Adler, C. R. (Ed.). (2001). *Put reading first: The research building blocks for teaching children to read* (pp. 49–54). Washington, DC: National Institute for Literacy.

Anderson, N. (2003, November). Scrolling, clicking, and reading english: Online reading strategies in a second/foreign language. *The Reading Matrix, 3*(3), 1–33.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah: Lawrence Erlbaum Associates.

Brown, K. (2006). Models, methods, and curriculum for ELT preparation. In K. Yamuna, B. B. Kachru, & C. L. Nelson (Eds.), *The handbook of world englishes*. Malden/Oxford: Blackwell Publishing.

Brown, A. L., Armbruster, B. B., & Baker, L. (1986). The role of metacognition in reading and studying. In J. Orasanu (Ed.), *Reading comprehension: From research to practice* (pp. 49–76). Hillsdale: Lawrence Erlbaum Associates.

Chen, L. (2008). *Primary school english education in China*. A keynote paper delivered at 2008 International Conference on Primary English in Beijing, China. http://www.chinabfle.org/YLP/en/newsdetails.asp?icntno=2317

Chomsky, N. A. (1977). On Wh-movement. In P. W. Culicover, T. Wasw, & A. Akmajian (Eds.), *Formal syntax* (pp. 94–98). San Francisco/London: Academic.

Clarke, M., & Silberstein, S. (1977). Toward a realization of psycholinguistic principles for the ESL reading class. *Language Learning, 27*, 134–154.

Coady, J. (1979). A psycholinguistic model of the ESL reader. In R. Mackay, B. Barkman, & R. Jordan (Eds.), *Reading in a second language* (pp. 5–12). Massachusetts: Newbury House.

Cortazzi, M., & Jin, L. (1996). English teaching and learning in China. *Language Teaching, 29*, 61–80.

Dulay, H. S., & Burt, M. K. (1974). Natural sequences in child second language acquisition. *Language Learning, 24*, 37–53.

Fathman, A. (1975). The relationship between age and second language learning productive ability. *Language Learning, 25*, 2.

Gallimore, R., & Tharp, R. (1990). Teaching mind in society: Teaching, schooling, and literate discourse. In L. C. Moll (Ed.), *Vygotsky and education: Instructional implications and applications of sociohistorical psychology* (pp. 175–205). New York: Cambridge University Press.

Hu, G. (2004). English language education in China: Policies, progress, and problems. *Language Policy* (2005) *4*, 5–24

Hu, Y. (2007). China's foreign language policy on primary english education: What's behind it? *Lang Policy, 6*, 359–376.

Huang, Y., & Xu, H. (1999). Trends in english language education in China. *ESL Magazine*, 39(6).

International Reading Association & National Association for the Education of Young Children. (1998). Learning to read and write: Developmentally appropriate practices for young children. *Young Children, 53*(4), 30–46.

Krashen, Stephen D. (1988). *Second language acquisition and second language learning*. New York: Prentice-Hall International.

Lam, A. (2002). *English in education in China: Policy changes and learners' experiences*. Oxford: Blackwell Publishers Ltd.

Makino, T. (1980). *Acquisition order of english morphemes by Japanese adolescents*. Tokyo: Shinozaki Shorin Press.

Mandel, L., Rueda, R., & Lapp, Diane. (2002). *Handbook of research on literacy and diversity* (pp. 277–291). New York: The Guilford Press.

Mok, M. M. C., Moore, P. J., & Kennedy, K. J. (2006). The development and validation of the self-learning scales (SLS). *Journal of Applied Measurement, 7*(4), 418–449.

Nunan, D. (2003). The impact of English as a global language on educational policies and practices in the Asia-pacific region. *TESOL Quarterly, 37*(4), 589–613.

Pearson, P. D., Barr, R., Kamil, M. L., & Mosenthal, P. (2000). *Handbook of reading research* (Vol. III, pp. 853–857). Mahwah: Lawrence Erlbaum Associates, Inc.

Roskos, K. A., Christie, J. F., & Richgels, D. J. (2003). The essentials of early literacy instruction. *The National Association for the Education of Young Children, 58*, 52–60.

Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time series analyses of diary data. *Contemporary Educational Psychology, 31*, 64–96.

Schunk, D. H., & Zimmerman, B. J. (2007). Influencing children's self-efficacy and self-regulation of reading and writing through modeling. *Reading & Writing Quarterly, 23*(1), 7–253.

Segers, E., Takke, L., & Verhoeven, L. (2004). Teacher-mediated versus computer-mediated storybook reading to children in native and multicultural kindergarten classrooms. *School Effectiveness and School Improvement, 15*, 215–226.

Strickland, D. S., Schickedanz, J. A., & Morrow, L. M. (Eds.). (2004). *Learning about print in preschool*. International Reading Association, Inc.

Teale, W. H. (1987). Emergent literacy: Reading and writing development in early childhood. *National Reading Conference Yearbook, 36*, 45–74.

Trelease, J. (2006). *The read-aloud handbook*. New York: Penguin Group (USA) Inc.

Chapter 19
Physical Education in Higher Education in Hong Kong: The Effects of the Intervention on Pre-service Sports Coaches' Attitudes Towards Assessment for Learning Used in Sports

Henry Kai On Lee

19.1 Background of the Study

In the last century, the Hong Kong education system was mainly for the elite (Fu 1988). Only a small percentage of students could enter universities, and assessment at that time was to select the most capable students for university education. At the turn of the century, Hong Kong introduced a series of education reforms, and one of these was to enlarge the function of assessment from only serving the purpose of selection to include also the mission of Assessment for Learning (Curriculum Development Council 2001, 2002). There was ample international research that demonstrated the positive effect of Assessment for Learning on student motivation to learn, as well as on their achievement levels. However, in spite of the Hong Kong Government having an Assessment for Learning policy, there was a lack of support for its implementation at the classroom level. Support was particularly lacking in sports education: many sports coaches, for example, did not receive professional training on using Assessment for Learning in their teaching (Hong Kong Sports Development Board 1999).

The notion of Assessment for Learning is that assessment serves learning. It can help to identify the gap between the learner's current performance and the targeted learning objectives (Hattie and Temperley 2007). The crucial point of Assessment for Learning is to provide quality feedback because this will enhance both teaching and learning. Recent research (Berry et al. 2003; Black and Wiliam 1998a; Lee 2007; Wiliam and Thompson 2007) indicates that Assessment for Learning raises both students' motivation for learning and effectiveness of their learning techniques.

H.K.O. Lee (✉)
Faculty of Arts and Sciences, The Hong Kong Institute of Education,
Tai Po, Hong Kong
e-mail: leeko@ied.edu.hk

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_19, © Springer Science+Business Media Dordrecht 2013

The researcher is currently a sports educator at The Hong Kong Institute of Education, and in this role, he can closely observe the teaching style of in-service sports coaches in Hong Kong. The researcher has noticed that many sports coaches still use the traditional skills-based teaching style in their sports teaching. Under this teaching style, students are restrained in their own individual thinking and self-reflective ability (Cheung 2002; Liu 1998), and this can inhibit student learning.

In response to this prevalence of the traditional teaching style, the researcher wishes to investigate what benefits Assessment for Learning can offer to student learning in sports. Several pieces of research (e.g. Casbon and Spackman 2005) have already demonstrated the effectiveness of Assessment for Learning for helping students learn in sports. Nevertheless, the extent to which these positive findings can be transferred to the educational context of Hong Kong still needs further investigation. This is the purpose of the current study.

Anecdotal evidence suggests that sports coaches in Hong Kong are still unaware of the significance of Assessment for Learning. In order to investigate the effectiveness of the pedagogy, the researcher considers that it would be strategically more efficient to initially promote this teaching style to pre-service sports coaches in Hong Kong. This is because pre-service sports coaches are more malleable and likely to change than in-service sports coaches, who are under the constraints of regulations and the workplace culture of the schools in which they are currently teaching. If the current study shows evidence of benefits of Assessment for Learning, then the style could then be promoted also to in-service sports coaches and physical education teachers.

This study developed the instructional package “Sport Coaches’ Assessment for Learning (SCAFL) Programme”, an intervention for pre-service sports coaches designed to aid the implementation of Assessment for Learning in sports teaching and to enhance the teaching quality of pre-service sports coaches. It is anticipated that results from this study can enrich pre-service sports coaches’ repertoire of teaching styles, as well as raise the teaching quality of pre-service sports coaches in sports.

19.2 Assessment for Learning

19.2.1 The Nature of Assessment for Learning

Assessment for Learning involves using assessment to provide feedback in order to enhance student learning achievement (Gilson 2009). Under this principle, students can improve most if they understand their learning objectives, where they are at the present time and how they can achieve the learning objectives (Fautley and Savage 2008). Black et al. (2003) and Black and Wiliam (2009) have all identified that Assessment for Learning involves sharing learning objectives with students, providing feedback to students to diagnose and improve their learning, giving students techniques in peer and self-assessment so that they can review and reflect on their own and others’ learning performance and progress, and recognising that students’ self-perception can be enhanced.

19.2.2 The Effects of Assessment for Learning

Numerous studies from around the world have demonstrated that Assessment for Learning has significant effects on student learning (Assessment Reform Group 2002; Berry 2008; Black and Wiliam 1998a; Casbon and Spackman 2005; Earl 2003; Hattie and Temperley 2007; Mok 2010; Sadler 1989). Some studies (e.g. Black and Wiliam 1998a; Stiggins 2005) showed that learning gain can be achieved in a very short time. Furthermore, Assessment for Learning has been shown to be good for long-term retention of learning, especially for weaker learners (Black and Wiliam 1998a), while students' self-reflective and assessing ability, critical thinking and learning motivation can be enhanced through peer and self-assessment (Assessment Reform Group 1999; Black and Wiliam 1998a; Sadler 1989). Assessment for Learning also helps students to learn how to analyse, assess and reflect on what they are accomplishing (Earl 2003; Stiggins 2005). In addition, through the students actively participating in recognising the next learning steps, they can become more motivated to reflect on what needs to be learnt and how to learn it (Black and Wiliam 2009). Finally, students' self-confidence in learning can be strengthened (Assessment Reform Group 1999; Casbon and Spackman 2005).

19.2.3 Strategies for Implementing Assessment for Learning

Teachers play a crucial role in the successful implementation of Assessment for Learning in Hong Kong schools (Berry 2008; Pang and Leung 2010). To begin with, teachers need to have a detailed and regular cycle of planning, teaching, assessing students' work, reviewing and revising (Casbon and Spackman 2005). Through this cycle, teachers can gain a thorough picture of how students learn, what needs to be improved and what students need to do to progress to the next objective.

In addition, frequent provision of immediate, effective and specific feedback is crucial (Hattie and Temperley 2007). To do this, teachers need to monitor student learning frequently during the learning process, including how the students apply their knowledge, their attitudes towards learning and whether they can achieve the learning goals (Berry 2005). Ongoing monitoring will enable teachers to diagnose student strengths and weaknesses (Savage 2011). Under Assessment for Learning, traditional assessment methods, such as tests or exams, projects and presentations, can still be used (Berry 2008). Students' self-assessing and self-reflective abilities can be cultivated by showing students exemplars of high- and low-quality work: when looking at exemplars of high-quality work, students can think about what they would need to do to achieve such a standard, and when faced with exemplars of low-quality work, the students can give their teachers suggestions as to how to improve the work (Berry 2008). In this way, students learn how to take more responsibility for their own learning rather than just learning passively from their teachers (Nygaard et al. 2009).

Formative assessment is a key part of Assessment for Learning. Formative assessment usually involves: (a) providing frequent feedback, (b) self- and peer assessment and (c) formative use of summative tests (Black et al. 2003; Wiliam 2000, 2007a, b). Marks on written work indicate standards achieved, but they do not give students information on how to improve their learning—this is where frequent formative feedback can help; self- and peer assessment are practices that can help students to become independent learners as their self-reflective ability and critical thinking can be enhanced significantly (Hinett and Thomas 1999; Whitfield 2000); and results from summative tests can also be used to give formative feedback (Berry 2008, 2011), which is an extension of the concept of providing frequent feedback.

Students can learn more effectively when formative assessment is used suitably during the teaching process (Berry 2008; Black and Wiliam 1998a, 2009). Through formative assessment, information is collected about the students' learning. The teachers then analyse the information and provide specific and prompt feedback to the students. If students have already achieved the targeted goals, different learning activities can be prepared so that students can keep progressing (Berry 2008; Black et al. 2003).

19.2.4 Implementing Assessment for Learning in Hong Kong

The Hong Kong Government strongly supported the adoption of Assessment for Learning as a key element of the assessment reform. It encouraged schools to switch from Assessment *of* Learning to Assessment *for* Learning (Curriculum Development Committee 2001, 2002). The Curriculum Development Committee (2002) highlights in its school policy on assessment document that all schools should review their current assessment practices and put more emphasis on Assessment for Learning. It is a process in which teachers seek to identify and diagnose students' learning problems and provide quality feedback for students to improve their work.

To support the implementation of Assessment for Learning within the Hong Kong education system, the Education Bureau (EDB) recommended using formative assessment strategies such as feedback sheets, and school-based assessment (SBA) (Curriculum Development Council 2002; Hong Kong Examinations and Assessment Authority 2007; School-based Assessment Consultancy Team 2007). Yet, despite the Hong Kong Government strongly pushing Assessment for Learning, there has been little change in the classroom. What seems to be lacking, perhaps, are concrete strategies for the implementation of Assessment for Learning in schools and a determined promotion of teacher professional development. Many schools have still kept their original teaching pattern—Assessment *of* Learning, rather than *for* Learning. So why have schools been reluctant to change? It seems that the implementation of Assessment for Learning is inhibited by several factors, including parental pressures for their children to enter universities, employers using published examination results as recruitment criteria and societal expectations of defining excellence by examination results (Carless 2002, 2005; Davison 2007).

19.2.5 Teachers' Role in the Success of Assessment for Learning

Teachers play an important role in student learning: effective teachers can transfer the latest knowledge to students, and this is crucial to ensuring their achievement (Choi and Tang 2009; Troman and Raggal 2008). As research has shown that Assessment for Learning is one method that can contribute greatly to student learning (Berry 2008; Black et al. 2003; Black and Wiliam 1998a; Lee 2007; Wiliam 2001), then Assessment for Learning has strong relevance to teachers.

Indeed, teachers play an important role in Assessment for Learning. However, before this methodology can be successfully implemented and student learning improved, teachers must be willing to implement several critical practices. Firstly, teachers need to clearly state the learning objectives at the beginning of the lesson and also remind the students of these objectives during the lesson; this will help students understand the learning objectives of each session and help them to achieve (Casbon and Spackman 2005). Secondly, teachers should set appropriate learning activities so that students can achieve the learning objectives progressively (Troman and Raggal 2008). Thirdly, teachers should use specific (ongoing, formative and constructive) feedback to help students correct their mistakes and improve their performance (Berry 2008; Black and Wiliam 1998a; Casbon and Spackman 2005). Fourthly, teachers should use peer and self-assessment to develop their students' critical thinking and self-reflective ability so that the students understand how to learn by themselves (Whitfield 2000). And finally, teachers should develop their students' metacognitive ability in the aspects of self-reflection and self-knowledge (Mok 2010). After teachers have adopted these practices of Assessment for Learning, it is hoped that students will experience a positive change in how they perceive themselves.

19.2.6 Barriers to Implementing Assessment for Learning in Hong Kong

Although Assessment for Learning is effective for student learning, there are several barriers to its successful implementation in Hong Kong schools. Firstly, teachers are already burdened with a huge workload (Chan et al. 2006), and so it is not easy to promote a huge education reform like Assessment for Learning in schools. Secondly, support from school principals is crucial (Henkin and Holliman 2009)—if the principals do not recognise the effectiveness of Assessment for Learning on student learning, then implementation is difficult. Thirdly, student belief in Assessment for Learning is also crucial. As Assessment for Learning involves active learning, passive students may not like this teaching style. And finally, many teachers do not receive sufficient professional development on Assessment for Learning (Berry 2008), and so, their understanding and familiarity with the use of Assessment for Learning in practice may not be adequate.

19.2.7 Domain of Assessment for Learning in This Study: Feedback

In the current study, the researcher focused on the domain of Assessment for Learning: feedback. Indeed, specific feedback is a key feature in Assessment for Learning (Black and Wiliam 1998b; Chappuis and Stiggins 2001; Hattie and Temperley 2007; Mok 2010). Feedback has been shown to be significant in enhancing student learning and performance (Black and Wiliam 1998b; Curriculum Development Council 2004; Hattie and Temperley 2007; Lee 2007). It explains to students what a good performance is, delivers high-quality information to students about their learning performance, informs students whether they have achieved the targeted learning goal or how to achieve it, and refines the existing curriculum and teaching strategies (Berry 2008; Curriculum Development Council 2006; Hattie and Temperley 2007; Wiliam 2007b).

Feedback can be received positively or negatively (Black and Wiliam 1998b; Hattie and Temperley 2007; Wiliam 2007a), and different types of feedback can influence student learning (Hattie and Temperley 2007). On the positive side, feedback about the gap between a learner's performance and their learning goal is powerful because it can provide constructive information on how the learner can improve their performance (Black and Wiliam 1998b; Curriculum Development Council 2004; Hattie and Temperley 2007; Sadler 1989). However, students who perceive feedback as negative may develop a habit of avoiding challenging tasks for fear of failure (Black and Wiliam 1998b; Curriculum Development Council 2004; Hattie and Temperley 2007).

Feedback can be summative or formative (Wiliam 2000). Summative feedback is usually in the form of marks or grades and informs the students about their learning outcomes. Because summative feedback occurs at the end of the learning process, it does little to influence the learning process (Berry 2008). Formative feedback, however, informs the learner of any weaknesses *during* the learning process and so gives them the chance to further improve or to enhance their performance (Black and Wiliam 1998b; Hattie and Temperley 2007; Wiliam 2001). Formative assessment is also effective because it focuses on the quality of a student's work or performance, and most students feel comfortable with this (Casbon and Spackman 2005).

19.3 Research Design and Methodology

This section outlines the research design and methodology of the current study, namely, (a) research questions, (b) plan of the study, (c) research methodology, (d) the participants, (e) the intervention of Assessment for Learning in sports teaching, (f) experimental design, (g) analytical methods, (h) assumptions of the study, (i) limitations of the study, (j) measuring instrument (the questionnaire) and (k) strategies to ensure the validity of the study.

The stages listed above are a step-by-step way of conducting the study. They organise the overall framework of the study and provide a suitable strategy for tackling the research questions.

19.3.1 Research Questions

1. What is the immediate effect of the intervention on pre-service sports coaches' attitudes towards Assessment for Learning in sports?
2. What is the long-term effect of the intervention on pre-service sports coaches' attitudes towards Assessment for Learning in sports?

19.3.2 Plan of the Study

The study comprised several steps. The first step was conducting a pilot study to test the reliability and validity of the questionnaire. The pilot study enabled the researcher to assess whether the questionnaire could be used in the main study and also enabled him to make immediate changes if necessary.

The second step was conducting the main study. A questionnaire was administrated to the participants at the beginning, in the middle and after the experiment so that their changes of attitudes towards Assessment for Learning in sports could be measured.

The quantitative data collected from the questionnaires was analysed. An analysis of covariance (ANCOVA), after adjusting the measurement of scores collected at mid-test using the procedure of Wolfe and Chiu (1999), was used to address the first research question, and a paired-sample *t*-test to address the second research question.

19.3.3 Research Methodology

In this study, a questionnaire was the main quantitative research tool used to collect data for analysis. Questionnaires are one of the most common tools in education research because they are cost-effective and the researcher can collect the required dataset within a short time (Gall et al. 2010; Hartas 2010).

In this study, two participants (3.2%) were not willing to reveal their attitudes in the questionnaire. Nevertheless, the questionnaire was still considered to be an appropriate tool for investigating the current study, as it was assumed that all participants expressed their views truthfully and honestly. This assumption was made on the grounds that the researcher had developed a good relationship of trust with his students.

19.3.4 The Participants

All the participants of the study were students at The Hong Kong Institute of Education. They were studying for an associate degree (AD), majoring in "Sports Coaching and Management". After the students graduate, one of their possible future careers will be as full-time sports coaches. During the study, however, they were all pre-service sports coaches.

Sixty-three pre-service sports coaches were included in the study. Forty-two were year 1 students, and 21 were year 2 students in the AD sports programme. Most (50 of the 63) of the students were male. In this study, there was no intention to compare year 1 and year 2 students nor to compare male and female students.

The researcher chose pre-service, rather than in-service, sports coaches as participants in the study for two reasons. Firstly, as the participants were students of the associate degree sports programme, it was easy for the researcher to approach them—he was teaching the participants in the academic year of 2009/2010—and invited them to join the study. Secondly, the researcher considered that pre-service sports coaches would be more open to innovation and that they would accept the new and creative Assessment for Learning pedagogy more easily because they had not been in the field and so been influenced by sports and social culture. In comparison, in-service sports coaches have already established their own teaching style, and so, the researcher believed it would not be so easy to change their philosophy and teaching practices. In-service coaches are also often constrained by other factors, including sports culture, parental expectations and government policy. Consequently, the researcher considered that it was more suitable to invite pre-service sports coaches to participate in the study.

19.3.5 The Intervention of Assessment for Learning in Sports Teaching

In the past, there has been little research on Assessment for Learning in sports teaching. In this study, the researcher designed an intervention programme to help pre-service sports coaches understand how to implement Assessment for Learning in sports teaching. The intervention was designed to equip the pre-service coaches with adequate knowledge to enable them to refine their traditional sports teaching practices. The intervention programme lasted a full semester.

The intervention comprised four components, namely, (a) preparation with a 2-day workshop that gave the students the theory behind Assessment for Learning and demonstrated use of the assessment technique in sports teaching, (b) instructions prior to every experimental teaching session on how to implement Assessment for Learning in their class, (c) experimental teaching sessions using Assessment for Learning in sports teaching and (d) debriefing on students' implementation of Assessment for Learning after each experimental teaching session in order to enhance their next implementation. These four components are elaborated in the following paragraphs.

19.3.5.1 The Two-Day Workshop

In the 2-day workshop, participants were briefed with the purpose of the research study and given an explanation of the experimental design, a description of the current climate of sports teaching in Hong Kong, a summary of the weaknesses of traditional teaching methods in sports, an elaboration on Assessment for Learning, a demonstration of both a traditional teaching method in sports and how to adopt Assessment for Learning in sports teaching, a comparison between Assessment for Learning and traditional teaching methods in sports teaching, and, finally, they had the opportunity to debrief.

This workshop was crucial for the participants because they were able to build up their competence and confidence in using Assessment for Learning in sports teaching. They learnt what Assessment for Learning was and how to implement it appropriately into a sports teaching programme, which is important for their future careers as sports coaches.

After the workshop, the participants reflected that they had learnt a lot about how to use Assessment for Learning in sports teaching. Furthermore, they thought Assessment for Learning was effective in enhancing sports performance when compared with traditional teaching methods. They generally agreed that Assessment for Learning enabled students to learn about sports to a greater depth. With their positive feedback, the researcher felt confident he would be able to conduct this study.

19.3.5.2 Instructions Prior to Every Experimental Teaching Session

Prior to each experimental teaching session, the researcher had regular tutorials with each of the experimental groups who were learning to teach with Assessment for Learning. During the tutorial, the researcher discussed with the teaching group about whether the teaching contents were suitable for their peers, whether their activities facilitated the learning progress of the participants, and how to implement suitably during the teaching process the domain of *feedback* of Assessment for Learning being used in this study. The tutorials gave the students an opportunity to discuss and share their ideas, and this helped them in their understanding of Assessment for Learning. By the end of each tutorial, the teaching group had been able to clarify all their problems, and so, their confidence in using Assessment for Learning was enhanced. Thus, the tutorial was an important part for the intervention.

19.3.5.3 Experimental Teaching Sessions

The experimental teaching sessions served an important role in this intervention because they gave all the participants the opportunity to try out and practise Assessment for Learning in sports teaching. The experiment comprised 28 sessions: 14 sessions were for associate degree year 1 participants, and another 14 sessions were for associate degree year 2 participants. The experimental teaching groups implemented Assessment for Learning in sports teaching. In general, the participants integrated Assessment for Learning well into their teaching—most of them used the Assessment for Learning domains on the right issues and at the right moment.

All participants were divided randomly into two groups, namely, experimental and control groups. In each experimental teaching session, both experimental and control groups took turns to act as teaching and learning groups to teach and learn the same sports skills by using Assessment for Learning and traditional teaching methods, respectively. Each group was formed by five participants. For instance, the experimental group adopted Assessment for Learning method to teach and learn basketball drilling in one half court; meanwhile, the control group used traditional teaching method to teach and learn basketball drilling in the other half court.

In the following session, the roles of teaching and learning groups were interchanged. The other group mates of both experimental and control groups were asked to act as sports coaches to teach another new sports skills by using Assessment for Learning and traditional teaching methods, respectively. The roles of sports coaches and learners took turn continuously in each session.

During the experimental teaching sessions, the researcher sat nearby to observe the participants' teaching performance and how their use of Assessment for Learning was influencing learning effectiveness. He did not interfere with the participants during their teaching.

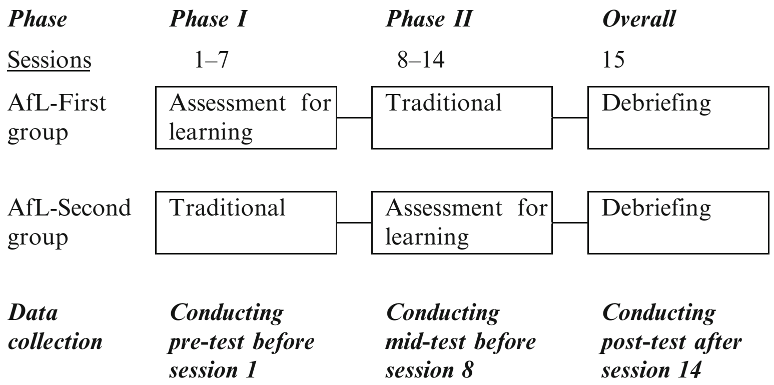
19.3.5.4 Debriefing

After each experimental teaching session, the researcher gave detailed, specific and constructive feedback to the teaching groups. The purpose of the debriefing was to help the students to perform better in their next teaching session. In particular, the researcher compared Assessment for Learning with traditional teaching methods in sports so that the participants understood more clearly how to effectively implement the new method of assessment in sports teaching.

19.3.6 Experimental Design

The experiment was conducted from January to July 2010. It comprised 28 teaching sessions, with AD year 1 participants taking part in 14 sessions and AD year 2 participants taking part in the other 14 sessions. Each session lasted for 3h. Each 14-session experiment was in two phases: Phase I comprised sessions 1–7 (from pre-test to mid-test), and Phase II comprised sessions 8–14 (from mid-test to post-test).

The participants of each AD class were divided into two groups: AfL-First and AfL-Second. In Phase I, AfL-First (the experimental group) was invited to teach and learn certain sports skills using Assessment for Learning, while AfL-Second (the control group) was invited to teach and learn the same sports skills using traditional teaching methods. In Phase II, the roles of the two groups were swapped: AfL-First became the control group (albeit a "contaminated" one, since they had already gone through some intervention already) and was invited to teach and learn certain sports



Phase	*Phase I*	*Phase II*	*Overall*
Sessions	1–7	8–14	15
AfL-First group	Assessment for learning	Traditional	Debriefing
AfL-Second group	Traditional	Assessment for learning	Debriefing
Data collection	*Conducting pre-test before session 1*	*Conducting mid-test before session 8*	*Conducting post-test after session 14*

Fig. 19.1 Experimental design

skills using traditional teaching methods, while AfL-Second became the experimental group and was invited to teach and learn the same sports skills using Assessment for Learning. After each session, the researcher debriefed with the experimental group, and this debriefing was an important part of the intervention.

Questionnaires were administered at the beginning of the intervention and after both Phase I and Phase II to examine the attitude changes of the participants towards Assessment for Learning as used in sports teaching. The participants were invited to fill in a questionnaire before session 1 (which served as the pre-test), in the middle (at the end of session 7) and at the end of session 14 (the post-test). Figure 19.1 shows the overall experimental design of this study.

The researcher used cross-over design (e.g. Gall et al. 2010) in the experimental part of this study. The cross-over design ensured that all participants, in both AfL-First and AfL-Second, had the opportunity to teach and learn sports skills in both an Assessment for Learning style and by using traditional teaching methods. As the researcher is also an educator, such an arrangement was necessary for both educational and ethical reasons.

19.3.7 Analytical Methods

During the experiment, 189 questionnaires were collected. Information from the questionnaires was entered into 2007 Microsoft Excel spreadsheet files. The data was then analysed using Winsteps® (Linacre 2010), a Rasch analysis software program, in order to check for unexpected scores and fit statistics for further analysis.

Several indicators were used to check the quality of the data. In particular, the benchmark for acceptable range of the infit and outfit for item and person was set between 0.6 and 1.4 (Bond and Fox 2007), and the acceptable separation index was set at 2.0 or over, following guidelines given by Linacre (2010). Dimensionality of scales was also checked to ensure that each scale in the analysis was unidimensional.

To measure the attitude change of pre-service sports coaches towards Assessment for Learning in Phases I and II, an analysis of covariance (ANCOVA), using the procedures described by Wolfe and Chiu (1999) which was used to address the first research question (about the immediate effect of the intervention), and paired-sample *t*-tests were used to address the second research question (about long-term effects).

Wolfe and Chiu's (1999) five-step procedure and how change was measured for the two research questions are discussed in the next two subsections.

19.3.7.1 The Five-Step Wolfe and Chiu (1999) Procedure

When applying a standard Rasch Rating Scale analysis to evaluate the changes from one occasion to another occasion, there is always the risk of unstable rating scale calibrations. To minimise this risk, Wright (1999b) wrote an algorithm to measure changes across different occasions. In the same year, Wright's algorithm was applied by another two scholars (Wolfe and Chiu 1999) to measure the change across two occasions (for instance, from pre-test to post-test or from pre-test to mid-test). After that, Wolfe and Chiu further extended Wright's algorithm to measure the changes across three occasions (for instance, from pre-test to mid-test and from mid-test to post-test). Wolfe and Chiu's (1999) five-step procedure is now well recognised as a method that can be applied to measure changes across occasions while reducing the misfit that can occur with the Rasch Rating Scale Model.

In the current study, Wolfe and Chiu's (1999) procedure was applied to measure pre-service sports coaches' change in attitudes towards Assessment for Learning from pre-test to mid-test. During his analysis, the researcher began to appreciate the importance and usefulness of the five-step Wolfe and Chiu (1999) procedure. For instance, by removing the instability of rating scale calibrations, the number of participants in this study ($n = 63$) who could fit the rating scale model increased, and an increased sample size increases the reliability of the study.

19.3.7.2 Measurement of Change in the Two Research Questions

An analysis of covariance (ANCOVA) was used to address the first research question about the immediate effect of the intervention on sports coaches' attitudes towards Assessment for Learning. The ANCOVA was used on the Phase I data to compare the attitudes towards Assessment for Learning between the two experimental groups, AfL-First and AfL-Second, on the mid-test, after controlling for any differences found at pre-test. If the intervention had an immediate effect on attitude, then the averaged attitude of the AfL-First group would be more positive than the averaged attitude of the AfL-Second group on the mid-test, after controlling for their initial differences at pre-test.

The second research question, about the long-term effect of the intervention, was addressed by comparing the attitudes of participants in the AfL-First group at post-test against the attitudes of this same group at mid-test. A paired-sample *t*-test was used for this analysis. During Phase II, there was no intervention for AfL-First. Consequently, if the averaged attitude of participants in the AfL-First group at the end of Phase II were greater than or equal to the same group's averaged attitude at the beginning of Phase II, then there would be evidence that the intervention effect at Phase I has been sustained long term.

19.3.8 Assumptions of the Study

In this study, several assumptions had to be made to support the validity of the study. Firstly, as the participants of the study are sports students at The Hong Kong Institute of Education, it was assumed that they would be confident in teaching sports by using both Assessment for Learning after the 2-day workshop and traditional teaching methods. Secondly, it was assumed that the questionnaire was reliable because it had been tested with a pilot study. And thirdly, it is assumed that the participants would report their responses in both the questionnaire and the in-depth focus-group interviews truthfully and honestly.

19.3.9 Limitations of the Study

Several factors limit the accuracy of the results of this study. Firstly, although the questionnaire was regarded as a valid and appropriate quantitative instrument for this study, there are nevertheless limitations to interpreting results from a questionnaire. Responses to Likert-type scale questions, for example, are very much just a numerical summary—they do not reveal participants' underlying thoughts.

A second limitation of the study is that the sample size (63 participants) was not high enough. Initially, the targeted participants of the study were both physical education students of the bachelor programme and sports students of the associate degree programme at The Hong Kong Institute of Education. Unfortunately, none of the physical education students were available for the entire period of the experiment, and only associate degree sports students were able to participate. As a result, only pre-service sports coaches, and not pre-service physical education teachers, were used in the study. Nevertheless, after successful completion of this study, the intervention could be easily extended to pre-service physical education teachers.

The study was further limited because of the paucity of previous research in this area. Although the effect of Assessment for Learning has been well investigated, there has been no specific study focusing on Assessment for Learning in a sports teaching context. This made it difficult for the researcher to find reference material that could be helpful when designing this study. However, the lack of existing research into Assessment for Learning within a sports context increases the value of the current study. This is even more so because Assessment for Learning has been a huge innovation in the education reform in Hong Kong.

Because the study was constructed for pre-service sports coaches in Hong Kong, it may not be suitable to apply its findings to other countries. Nevertheless, results of this study will be disseminated at international conferences and through refereed journals.

The study may also be limited because it relies on self-report measures. Although this investigation assumed that all 63 participants were honest in their responses, there are no guarantees. Nevertheless, the researcher had a good relationship, built on trust, with his students, which he hopes means that his students responded truthfully to the experiment.

The experimental result of the current study might also have been affected by the Hawthorne effect. The Hawthorne effect says that if an experiment involves people, minor influences have already been enough to twist the experimental results (Richard 1991; Stephen 1992). In the current study, the participants knew that they were participating in an experiment and learning some useful knowledge (the method of Assessment for Learning in sports teaching) which is favourable and crucial to their sports teaching career; thus, their attitudes towards the research focus will already be affected by their participation. When an experimental group is aware that they are participating in an experiment and receiving particular care, this can help to enhance or improve their performance or quality of work significantly (Richard 1991; Stephen 1992).

In the current study, students in AfL-Second taught and were taught with traditional sports-teaching methods in Phase I. Meanwhile, those in AfL-First were trained in using a new teaching method in sports (Assessment for Learning) and received considerable help and attention in implementing the new method. Such attention will usually result in changes in sports coaches' performance and student achievement favourable to the new teaching method.

In the process of trying out new teaching methods on participants in the experiment, the experimental result is almost certainly influenced by the Hawthorne effect (Ormrod 2011; Richard 1991; Stephen 1992; Tuckman and Monetti 2011). It may be explained that teachers usually approach a new method with some enthusiasm, and the students, aware that they are being taught by a new and different method, are also likely to display more interest and motivation than usual. Nevertheless, the influence of the Hawthorne effect can be expected to decrease as the novelty of the new method is reduced (Ormrod 2011; Stephen 1992; Tuckman and Monetti 2011).

In the current study, after the intervention in Phase I, students in AfL-First were taught with and about traditional teaching methods in Phase II. However, to maintain the intervention effect of Assessment for Learning on the experimental group long term, some follow-up was provided after the completion of the intervention. Workshops, group discussions and group self-reflection can be used to keep the enthusiasm of the experimental group fresh, so that they will continue to want to learn more about Assessment for Learning, and their attitudes will be sustained or even increased.

Given these limitations, the findings of this study are viewed as merely a starting point towards understanding the attitudes of pre-service sports coaches towards Assessment for Learning in sports teaching. In the future, a modified intervention programme will be promoted to both in-service sports coaches and physical education teachers in Hong Kong.

19.3.10 Measuring Instrument: The Questionnaire

19.3.10.1 Background of the Questionnaire for This Study

Over the years, different instruments have been developed to measure student attitudes towards Assessment for Learning. However, none of these instruments has focused on measuring Assessment for Learning in sports teaching, and so, a new questionnaire was constructed specifically for the current study.

The questionnaire comprised two sections. In section 1, participants were invited to fill in their personal information, including their name, student number, date, class, class year and gender. Section 2 comprised the participants' attitudes towards Assessment for Learning (feedback). The questionnaire had 19 Likert-type items.

The responses in section 2 were scored on a six-point Likert-type scale, with response options ranging from "strongly disagree" (coded as 0) to "strongly agree" (coded as 5). No time limit was set for the completion of the questionnaire. All participants were invited to complete the questionnaires at the pre-test, mid-test and post-test stages of the experiment. The completed questionnaires were collected by the researcher and then carefully checked for errors (e.g., whether there were multiple responses to a single item). Where there was an error, the response was scored as missing and omitted from the analysis.

Two steps were required to construct the questionnaire: collecting suitable scales to develop the items of the questionnaire and checking the reliability and validity of the questionnaire. For step one, relevant literature was reviewed in the areas of Assessment for Learning, assessment reform and feedback, with attention specifically focusing on available measurement instruments. A questionnaire that focused on measuring pre-service sports coaches' attitude towards Assessment for Learning in sports was then constructed.

The questionnaire consists of 19 items which are categorised into three domains, including overall feedback, corrective feedback and enhanced feedback. All these items were developed by the researcher in reading the literature about Assessment for Learning, teaching and learning effectiveness, assessment in education, educational psychology, sports and physical education. The design of the questionnaire is based on the rationale that the items are highly related to measure the attitude of the participants towards Assessment for Learning in sports. In order to investigate research questions (1) and (2), the items of the questionnaire are used three times (pre-test, mid-test and post-test) to measure the immediate and long-term effects of the intervention on the participants' attitudes towards Assessment for Learning in sports.

Table 19.1 Scales of the questionnaire

Domain of assessment for learning	Scale used in the study	Focus of the scale	Example item
Feedback	Overall feedback	Students receive overall feedback from peer sports coaches	We can receive feedback on how to correct mistakes
	Corrective feedback	Students receive feedback on how to correct mistakes from peer sports coaches	When we have done wrong, sports coaches will say this to help us do better: "Do you think what you have done is wrong?"
	Enhanced feedback	Students receive feedback from their peer sports coaches on how to improve their sports performance	When we have done well, sports coaches will say this to help us do even better: "You have a great talent!"

Table 19.2 Reliability of the scales

Scales	*CFI*	*TLI*	*RMSEA*	No. of items	Cronbach's alpha	Item separation reliability
Overall feedback	0.991	0.981	0.078	5	0.823	0.71
Corrective feedback	0.804	0.726	0.274	8	0.790	0.81
Enhanced feedback	0.988	0.981	0.105	6	0.889	0.56

Step two was to make sure that the questionnaire was reliable, useable and valid. To test the questionnaire, a pilot study was held in January 2010. Forty-two participants were invited to complete the questionnaire. They were bachelor students who studied physical education at The Hong Kong Institute of Education. The purpose of the pilot study was to eliminate any potential problems regarding possible ambiguities or problematic statements of items and to test the time required to complete the questionnaire.

As the questionnaire was presented in Chinese, the researcher had invited a Chinese editor to edit each item to clarify meaning and remove ambiguities prior to the pilot study. Based on the editor's advice, those problematic statements were modified accordingly before the pilot study.

After the participants completed the questionnaire, they were invited to give feedback to the researcher concerning difficulties in answering the questionnaire. No problems were reported, and all participants were able to complete the questionnaire within 20 min. In general, the overall structure of the questionnaire was acceptable.

In addition to feedback from participants in the pilot study, both Cronbach alpha and Rasch measurements were used to check the reliability of the questionnaire, in particular, to determine which scales should be retained and which deleted. The results are presented in Tables 19.1 and 19.2.

Careful examination of the validity of the questionnaire was also necessary. Validity refers to the extent to which the instrument measures what it is planned to measure (Fraenkel et al. 2011; Lissitz 2009). To address the question of validity, students were interviewed after their completion of the pilot study questionnaire. This interview was to check whether the participants had misunderstood any items. The questionnaire was also reviewed by an expert in the sports field, Dr Chen Shihui (the associate supervisor of this study), to ensure that the items were measuring the participants' attitudes towards Assessment for Learning in sports teaching.

Administration of the questionnaires (e.g. printing, distributing and collecting all questionnaires) was the sole responsibility of the researcher. This was done to further reduce potential error variance and so ensure the validation of the product.

19.3.10.2 Domain and Scales of the Questionnaire

The questionnaire comprised three scales, namely overall feedback, corrective feedback and enhanced feedback. Each part of the questionnaire, i.e. each scale, was analysed separately. The items making up these 3 scales are presented in Table 19.1.

19.3.11 Strategies to Establish Instrument Validity and Reliability

The reliability and validity of the scales are developed in this section. Confirmatory factor analyses (CFA) were undertaken for each scale in order to confirm the construct measured by the scales. CFA was used instead of EFA (exploratory factor analysis) because the scales were constructed according to the literature, and the items were compiled to reflect the specific constructs relevant to Assessment for Learning. The MPlus (Version 6) software was used to confirm the unidimensionality of each scale. The MPlus software reports three statistics: confirmatory factor index (CFI), Tucker-Lewis Index (TLI) and root mean square error of approximation (RMSEA); these statistics indicate the goodness of fit to a single-factor model of the items making up the scale. Bryne (2011) states that CFI and TLI values greater than 0.9 indicate a "good" fit, and values greater than 0.7 indicate an "adequate" fit. The third statistic, RMSEA, is a "badness of fit" index: it reflects any discrepancy between the model and the data. The minimum value of RMSEA is zero, which indicates a "perfect" fit. Bryne (2011) states that RMSEA values less than 0.05 indicate a "good" fit, values between 0.05 and 0.1 indicate a "moderate fit" and values which are greater than 0.1 indicate that the fit is "poor". However, RMSEA values are affected by sample size, with a larger sample size tending to be related to smaller RMSEA values. Noting this constraint, the researcher used Bryne's (2011) benchmark values as the criteria in this study.

Cronbach's alpha coefficient is an indication of the internal consistency of a scale (Bond and Fox 2007) and so was a useful statistic for this study. A Cronbach's alpha value greater than 0.7 represents an "adequate" fit, and a value greater than 0.9 represents a "good" fit (Bond and Fox 2007). Bryne (2011), however, added this cautionary note about the interpretation of Cronbach's alpha: the statistic is affected by the length of the scale, with scales with more items tending to get a higher Cronbach's alpha value.

Finally, the Rasch model was used to establish validity of the scales. The Rasch model converts raw ordinal-level data (Bond and Fox 2007) into interval-level data by means of a logistic regression transformation. The rating scale model, the algorithm for this transformation, is given in the following equation:

The probability of a person n with ability β_n selecting option x with difficulty level δ_i in responding to item i is given by

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} \left[\beta_n - (\delta_i + \tau_j) \right]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} \left[\beta_n - (\delta_i + \tau_j) \right]}, x = 0, 1, \ldots, m,$$

where $\tau_o \equiv 0$ so that $\sum_{j=0}^{0} \left[\beta_n - (\delta_i + \tau_j) \right] = 1 \cdot$

(Wright and Masters 1982)

The Rasch Rating Scale Model (Wright and Masters 1982) was fitted to the 6-point Likert-type response scales used in this study. Note that "ability" was used here in line with the Rasch terminology, to represent the person's attitude, inclination or willingness to agree or disagree with an attitude statement. Similarly, the word "difficulty" was used here to indicate the difficulty of an item to be agreed upon by the respondents (Bond and Fox 2007). The Winsteps® computer software version 3.71.0 (Linacre 2010) was used for the analyses.

The Winsteps® software uses several statistics to indicate the adequacy of fit to the Rasch Rating Scale Model, including item infit, item outfit, point-measure correlation, item separation reliability and dimensionality of the scale. The infit and outfit measures are used to reflect the extent to which the data adheres to the Rasch model. Scales with infit and outfit item values between 0.6 and 1.4 are considered to have a "good" fit (Bond and Fox 2007). The point-measure correlation is the correlation between each item and the rest of the items in the scale. Values between 0.4 and 0.8 are considered to reflect internal coherence of items making up the scale (Linacre 2002). Item separation reliability is a statistic which can theoretically range from 0 to 1, with values closer to 1 (greater than 0.9) being an indication of higher scale reliability (Bond and Fox 2007). Unidimensionality is crucial to scale construction (Linacre 2002) in that it indicates that only a single construct is being measured by the items making up the scale. The Winsteps® program calculates the eigenvalues of the main dimension, as well as eigenvalues of residual dimensions (in the name of "contrasts" to the first dimension). Eigenvalues less than 2 of the residual dimensions are indications that essentially only one dimension is being measured by the scale (Bond and Fox 2007).

19.3.11.1 Dimensionality and Reliability (Cronbach's Alpha) of the Scales

Results of the confirmatory factor analysis are presented in Table 19.2. It can be seen from Table 19.2 that the majority of the scales (2 out of 3) used in this study

Table 19.3 Infit and outfit statistics

		Infit MNSQ			Outfit MNSQ		
Scale	No. of items	Mean	Range	Item underfit	Mean	Range	Item underfit
Overall feedback	5	0.99	0.83–1.18	0	0.98	0.82–1.19	0
Corrective feedback	8	0.99	0.72–1.61	1	0.97	0.71–1.54	1
Enhanced feedback	6	1.00	0.49–1.21	0	0.98	0.54–1.15	0

Note: An item with MNSQ greater than 1.40 is called an "underfit item"

Table 19.4 Point-measure correlation range, item separation reliability and eigenvalues

			Point-measure correlation				Eigen-value	
Domain	Scale	No. of items	Range	No. of items lower than 0.4	Item separation reliability	Gi	Measure	1st contrast
Feedback	Overall feedback	5	0.75–0.78	0	0.71	1.224	5.6	1.6
	Corrective feedback	8	0.44–0.77	0	0.81	2.132	6.2	2.6
	Enhanced feedback	6	0.72–0.85	0	0.56	0.636	9.0	2.0

had *CFI* and TLI greater than 0.9; only the scale, corrective feedback (*CFI* = 0.804, *TLI* = 0.726), had scores less than 0.9. On the other hand, the RMSEA values ranged from 0.078 (overall feedback) to 0.247 (corrective feedback). However, only 1 scale had RMSEA values below 0.1: overall feedback. Given that only 63 participants were involved in this study, the large RMSEA might be related to the small sample size.

19.3.11.2 Item Statistics of the Scales Using Rasch Rating Scale Model

Item statistics of the scales using the Rasch Rating Scale Model (Wright and Masters 1982) are presented in Table 19.3. It can be seen from Table 19.3 that the scale corrective feedback has the largest range of infit (0.72–1.61) and outfit (0.71–1.54) values. In addition, the mean values of infit and outfit MNSQ of all 3 scales are around 1.00.

Table 19.3 also shows that 1 of the 3 scales had underfit items, including SSQ26 ("When we have done wrong, sports coaches will say this to help us do better: 'You have done wrong.'"). In addition, none of the scales had more than one item with underfit.

Table 19.4 shows point-measure correlation ranges, item separation reliability and eigenvalues. The results show that no scales with a point-measure correlation lower than 0.4. The item separation reliability ranged from 0.56 (enhanced feedback) to 0.81 (corrective feedback). Two scales (out of 3) had separation reliability above 0.7, which means they are considered to be of acceptable reliability (Bond and Fox 2007). Of the 3 scales, not one had a separation reliability below 0.5.

Table 19.4 also shows that 1 of the 3 scales had *G* values over 1.5, which means that this scales can be assumed to be reliable. The good scale was corrective feedback ($G = 2.132$). One of the 3 scales (overall feedback) had eigenvalues of 1st contrast below 2.0, which is also considered reliable (Linacre 2010, p. 318). The other two scales had eigenvalues of 1st contrast above 2.0, which indicated that some residual factors exist in these two scales which could be influencing the results.

19.4 Results of the Study

19.4.1 The Immediate Effect of the Intervention on Attitudes Towards Assessment for Learning in Sports

This section presents the means and adjusted means of the scales used in the study. The means represent the attitudes of participants in AfL-First (the experimental group, which studied with and about Assessment for Learning) and AfL-Second (the control group, which studied using traditional teaching methods). These mean values will give an indication of the differences between the two groups at the pre- and mid-test stages of the experiment. The mean values are also presented graphically. Graphs of the means will clearly show whether the mean attitudes of each group have remained constant or, if not, the extent of any change between the pre- and mid-test stages of the experiment. A large difference between the two groups would imply possible intervention effect. This section then presents the results of the analysis of covariance (ANCOVA) of the 3 scales used in the study. The ANCOVA tests the statistical significance of any difference in attitudes' change of each group for each of the 3 scales by using the scores from the pre-test and from the mid-test as the two covariates. An ANCOVA is the traditional approach that would be used in such an analysis and enables the attitudes of participants in the experimental and control groups (AfL-First and AfL-Second, respectively) to be compared after the intervention and after controlling for any differences that might have been present before the intervention. Finally, the Rasch model (Wright and Masters 1982) is used to identify changes for individual sports coaches, after accounting for possible changes in items of the measurement instrument before and after the intervention. This is done using the set of procedures developed by Wolfe and Chiu (1999). Results of the analysis using the Wolfe and Chiu procedures are presented following the ANCOVA results.

Table 19.5 The adjusted means of three feedback scales from the mid-test for both AfL-First and AfL-Second groups

			Pre-test		Mid-test		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Scale	Group	*N*	Mean	SE	Adjusted mean	SE	(95% Confidence interval of adjusted mean)
Overall feedback	AfL-First	32	0.818	0.418	3.183	0.361	(2.462, 3.905)
	AfL-Second	31	1.884	0.437	2.668	0.367	(1.935, 3.401)
Corrective feedback	AfL-First	32	0.117	0.133	1.576	0.145	(1.287, 1.865)
	AfL-Second	31	1.392	0.164	0.418	0.147	(0.123, 0.713)
Enhanced feedback	AfL-First	32	0.586	0.251	1.764	0.265	(1.234, 2.294)
	AfL-Second	31	1.917	0.420	1.518	0.270	(0.978, 2.057)

19.4.1.1 ANCOVA to Compare the Difference Between AfL-First and AfL-Second Groups on Feedback Scales at Mid-test

It can be seen from Table 19.5 that at pre-test, participants from AfL-Second had higher mean attitude scores than those from AfL-First, and that this was consistent across all three scales relating to feedback. At the mid-test stage of the experiment, however, all scale means for AfL-First were higher than those for AfL-Second.

Plots of the mean attitude scores of AfL-First (the experimental group using and learning about Assessment for Learning) and AfL-Second (the control group using traditional teaching methods) at pre-test and mid-test for the three feedback scales are presented in Figs. 19.2, 19.3, and 19.4. The figures show that the attitudes of participants in the experimental group (AfL-First) improved from pre- to mid-test for all three feedback scales. This means AfL-First group participants rated feedback more importantly at the mid-test stage of the experiment, i.e. after the intervention when feedback had been provided by their peer sports coaches, than they had at the pre-test stage, prior to the intervention. This was true for all three feedback scales: overall feedback (Fig. 19.2), corrective feedback (Fig. 19.3) and enhanced feedback (Fig. 19.4).

Similarly, AfL-Second's mean for the overall feedback scale also improved from pre- to mid-test, even though these participants had been taught using traditional teaching methods during Phase I of the experiment. This result is shown in Fig. 19.2. Conversely, AfL-Second's means for the corrective feedback (Fig. 19.3) and enhanced feedback (Fig. 19.4) scales regressed from pre- to mid-test. This means that while participants from AfL-Second rated overall feedback more importantly at the mid-test stage of the experiment than they had at the pre-test stage, they rated specific feedback, either to correct mistakes or to further enhance performance, less importantly.

Analysis of covariance (ANCOVA) showed that there were statistically significant differences in the sports coaches' attitudes towards corrective feedback between the

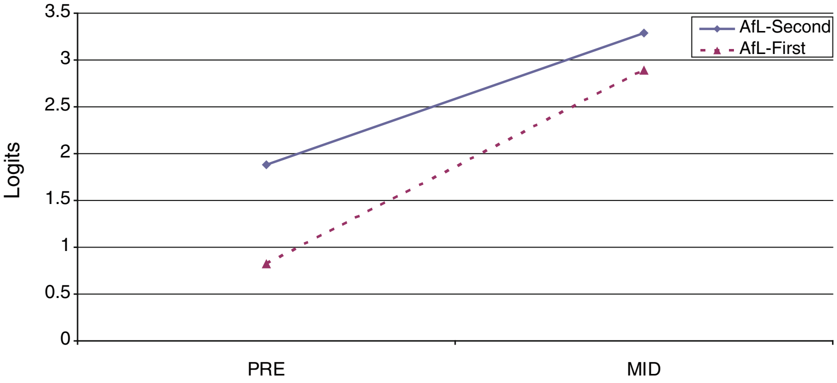


Fig. 19.2 The means of attitude scores towards overall feedback of AfL-First and AfL-Second groups at pre-test and mid-test

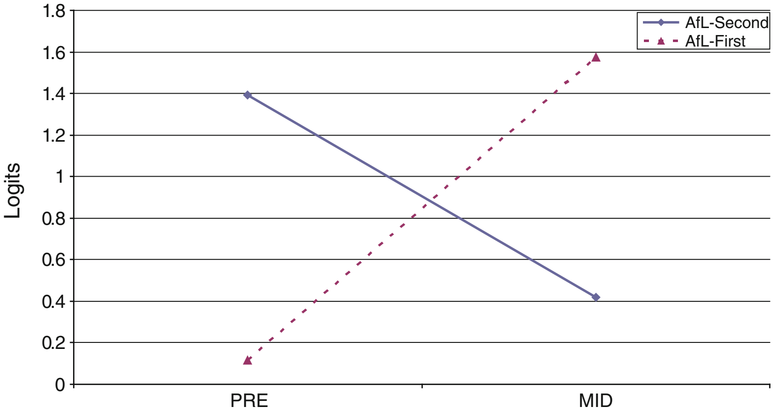


Fig. 19.3 The means of attitude scores towards corrected feedback of AfL-First and AfL-Second groups at pre-test and mid-test

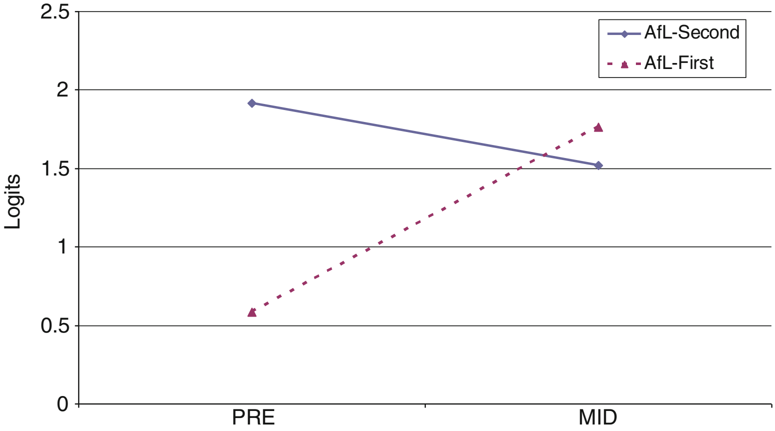


Fig. 19.4 The means of attitude scores towards enhanced feedback of AfL-First and AfL-Second groups at pre-test and mid-test

Table 19.6 Comparison by ANCOVA of the mean attitude scores towards three feedback scales between AfL-First and AfL-Second groups

Domain	Scales	df	*F*	*p*-value	Effect size
Feedback	Overall feedback	1, 60	0.980	0.326	0.016
	Corrective feedback	1, 60	25.552	0.000	0.299
	Enhanced feedback	1, 60	0.403	0.000	0.007

Table 19.7 Comparing individuals' changes in attitude towards three feedback scales between participants in AfL-First and AfL-Second groups on corrected measures from pre-test to mid-test

Scales	Groups	Substantial improvement	Improvement	No change	Deterioration	Substantial deterioration
Overall feedback	AfL-First	4/32	20/32	7/32	2/32	0/32
		12.90%	62.50%	21.88%	6.25%	0.00%
	AfL-Second	4/31	8/31	12/31	7/31	0/31
		12.90%	25.81%	38.71%	22.58%	0.00%
Corrective feedback	AfL-First	6/32	16/32	10/32	0/32	0/32
		18.75%	50.00%	31.25%	0.00%	0.00%
	AfL-Second	0/31	2/31	12/31	15/31	2/31
		0.00%	6.45%	38.71%	48.39%	6.45%
Enhanced feedback	AfL-First	7/32	9/32	14/32	2/32	0/32
		21.88%	28.13%	43.75%	6.25%	0.00%
	AfL-Second	2/31	6/31	12/31	8/31	3/31
		6.45%	19.35%	38.71%	25.81%	9.68%

AfL-First and AfL-Second groups at mid-test ($F(1,60)=25.552$, $p=0.000$) and also in their attitudes towards enhanced feedback ($F(1,60)=0.403$, $p=0.000$), after controlling for their initial differences at pre-test. Results of the ANCOVA are presented in Table 19.6. These results showed that the intervention had made the sports coaches feel more positive about both corrective and enhanced feedback in the 7 weeks from pre-test to mid-test. However, it can also be seen from Table 19.6 that although corrective feedback had a partial eta-squared value of 0.299, the effect sizes for overall and enhanced feedback were small.

19.4.1.2 Inspection of Changes of Individual Sports Coaches in AfL-First and AfL-Second Groups on Feedback from Pre-test to Mid-test

Table 19.7 compares the changes of the attitudes of individual coaches in both AfL-First and AfL-Second groups. It can be seen that those in AfL-Second (the control group) had more cases of "no change", "deterioration" or "substantial deterioration" than did those in AfL-First (the experimental group). Furthermore, Table 19.7 shows that there were many more coaches from the experimental group, AfL-First, who showed "substantial improvement" or "improvement" than from the control group, AfL-Second. These results mean that the changes tended to be more positive for those who had participated in the Assessment for Learning intervention.

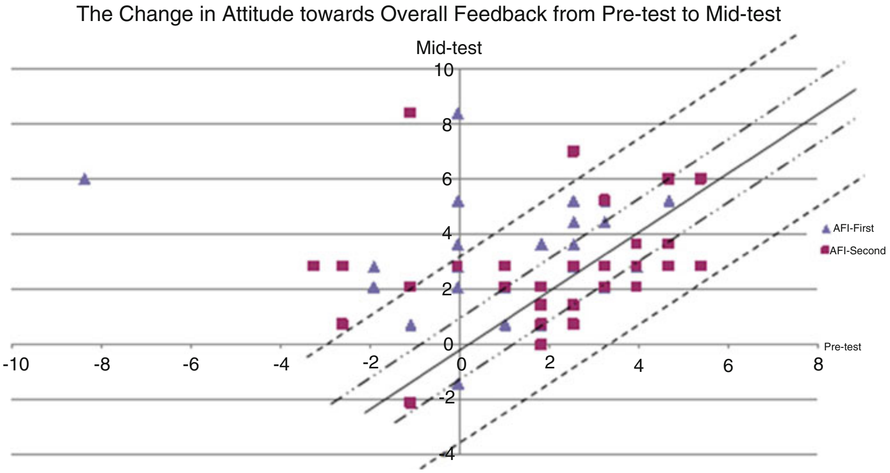


Fig. 19.5 The change in attitude towards overall feedback from pre-test to mid-test

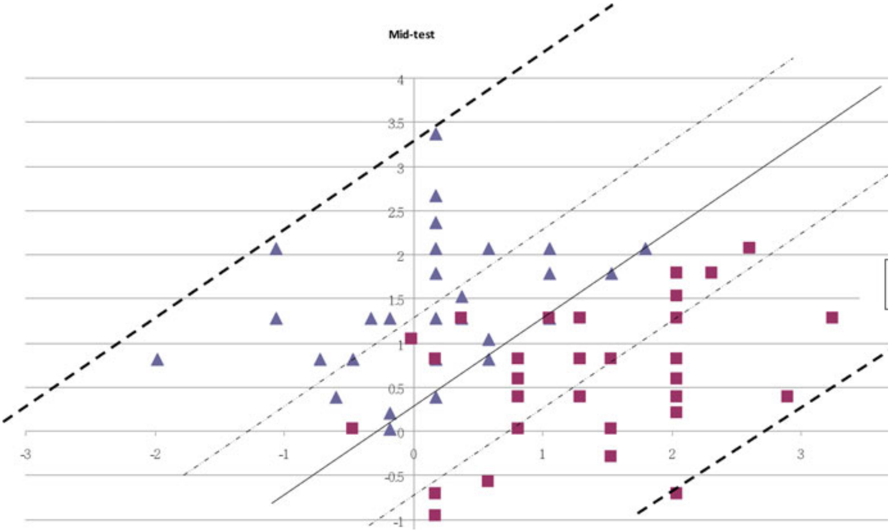


Fig. 19.6 The change in attitude towards corrective feedback from pre-test to mid-test

The individual sports coaches' changes in attitude are presented graphically in Figs. 19.5, 19.6 and 19.7. The figures reveal that some participants from the experimental group (AfL-First) showed large gains in their score, while a large proportion of those from the same group showed at least some gains. A few participants from AfL-First and many from AfL-Second fell in the "no change" region, while some participants from AfL-Second were in the "regressed" or "strongly regressed" regions.

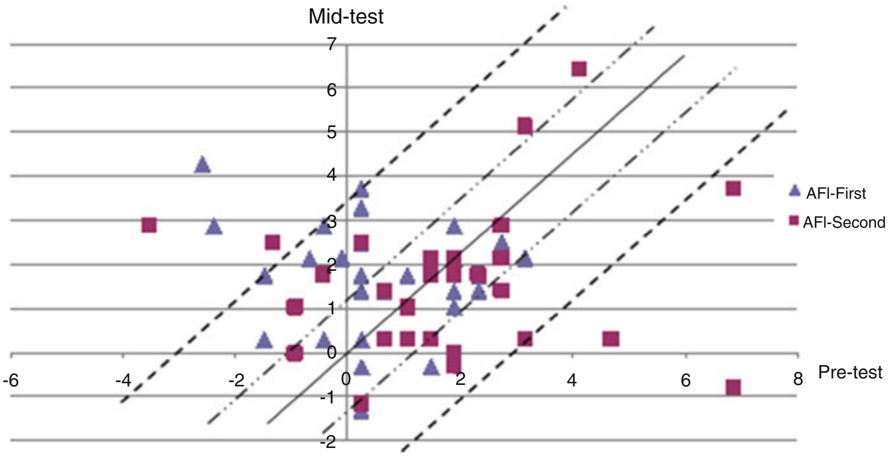


Fig. 19.7 The change in attitude towards enhanced feedback from pre-test to mid-test

Table 19.8 Paired differences of mean attitude values towards three feedback scales from mid-test to post-test of participants in AfL-First group

		Paired differences (post-test–mid-test)						
Domain	Scales	Mean	SD	SE mean	95% confidence interval	*t*	df	*p*-value
Feedback	Overall feedback	0.146	3.281	0.580	(−1.037, 1.329)	0.252	31	0.803
	Corrective feedback	0.132	1.684	0.298	(−0.475, 0.739)	0.444	31	0.660
	Enhanced feedback	0.083	2.297	0.406	(−0.745, 0.911)	0.204	31	0.840

These results show that the intervention had a greater effect on sports coaches in AfL-First (who were taught about and with Assessment for Learning) towards the three feedback scales than was felt by those in AfL-Second (who were taught using traditional teaching methods).

19.4.2 The Long-Term Effect of the Intervention on Attitudes Towards Assessment for Learning in Sports

The results in Table 19.8 show that the means for all three feedback scales increased from mid-test to post-test for participants from AfL-First (overall feedback increased by 0.1, corrective feedback by 0.132 and enhanced feedback by 0.083). However, once again, these changes were not statistically significant (overall feedback: paired-*t* (31) = 0.252, *p* = 0.803; corrective feedback: paired-*t* (31) = 0.444, *p* = 0.660; enhanced feedback: paired-t (31) = 0.204, p = 0.840). This means that the long-term effect of the intervention on attitudes towards feedback was sustained after 7 weeks without intervention for the sports coaches in AfL-First.

19.5 Discussion

19.5.1 Discussion of the Immediate Intervention Effects on Pre-service Sports Coaches' Attitudes Towards Assessment for Learning in Sports

According to the results of this study, the intervention had a more significant immediate effect on the attitudes of the experimental group towards corrective and enhanced feedback than it did on the attitudes of the control group. On the other hand, the intervention had no significant immediate effect on the attitudes of either the experimental or control group towards overall feedback.

19.5.1.1 Corrective Feedback and Enhanced Feedback

The results of this study showed that the intervention helped the experimental group to understand both corrective and enhanced feedback: the participants in the experimental group were able to correct their mistakes and improve any weaknesses in their sports skills immediately, as well as consolidate their sports skills effectively. This is in line with the literature (Black and Wiliam 1998a, 2009; Hattie and Temperley 2007; Tan et al. 2011; Wiliam 2011) which highlights that specific feedback helps students to better understand and internalise learning. In addition, the intervention helped the experimental group to understand that corrective feedback and enhanced feedback could significantly improve their sports performance. This is also in line with the research (Brookhart 2008; Hattie and Temperley 2007; Irons 2008) which suggests that specific feedback is beneficial in significantly enhancing student performance.

Corrective feedback and enhanced feedback are key elements in Assessment for Learning because they help students to recognise their own learning needs (Black and Wiliam 1998a, 2009; Tan et al. 2011; Wiliam 2011). Specific feedback can help a student correct their mistakes and enhance their performance in the learning process (Black and Wiliam 2009; Rowe 2005; Stiggins 2005; Wiliam 2011). In line with the literature (Black and Wiliam 2009; Stiggins 2005; Wiliam 2011), sports coaches should use more corrective feedback and enhanced feedback to instruct students to correct their mistakes and this will directly enhance sports performance.

Corrective feedback and enhanced feedback act as a form of "scaffolding" for student learning (Black and Wiliam 1998a, 2009). In providing specific feedback, the literature (Black and Wiliam 2009; Curriculum Development Council 2004;

Hattie and Temperley 2007; Lee 2007; Tan et al. 2011) advises that students should not be immediately given the complete solution when they have problems; instead, the students need to learn to think things through for themselves (Hewitt 2008; Reid and Green 2009). Furthermore, students should be helped to find alternative solutions rather than just simply repeating an explanation (Wiliam 2001, 2011; Wiliam and Thompson 2007). The quality of dialogue in specific feedback is important (Black and Wiliam 1998b, 2009; Hattie and Temperley 2007): students should always keep asking questions, and sports coaches should encourage them to do so (Hardman and Jones 2011; Robinson 2010). Finally, it is crucial for sports coaches to establish their students' trust because if the coach has a good relationship with their students, then the students will be more willing to receive specific feedback to improve their weaknesses (Hardman and Jones 2011; Robinson 2010). The experimental group in this study highlighted this point.

Hattie and Temperley (2007) showed that good quality corrective feedback and enhanced feedback can facilitate student learning effectively. Participants in this study highlighted the importance of high-quality specific feedback in improving sports performance significantly. Hattie and Temperley (2007) suggested that high-quality specific feedback should be timely and presented as concretely as possible; it should indicate the level of the learners, show the expected level of the learners, indicate student progress and development and give ways for improvement.

The experimental group in this study endorsed that positive and specific feedback can have a crucial impact on learning. Positive and specific feedback can stimulate students' motivation to learn (Black and Wiliam 1998b, 2009; Chappuis and Stiggins 2001; Hattie and Temperley 2007; Tan et al. 2011; Wiliam 2011). In Hong Kong, high-skilled students are always invited by sports coaches to do demonstrations for their classmates. To maintain their coaches' attention, they need to keep practising hard. By the same token, if sports coaches give negative feedback to students, students may feel uncomfortable and lose confidence, even though the feedback is very specific and may include suggestions for improvement (Kidman and Hanrahan 2011; Robinson 2010). Students may think that they do not have enough ability or talent. If sports coaches humiliate their students, the students may give up on sports completely (Comer and Gould 2011; Ormrod 2011). In the long term, negative feedback can affect a student's health and personal character (Jones et al. 2004), and so needs to be avoided.

19.5.1.2 Overall Feedback

Much research (e.g. Black and Wiliam 1998b, 2009; Curriculum Development Council 2004; Evers and Spencer 2011; Hattie and Temperley 2007; Lee 2007; Miller and Nendel 2011; Tan et al. 2011; Wiliam 2011) has shown that overall feedback can help student learning. In this study, both the experimental and control groups also showed positive attitudes towards overall feedback used in sports teaching and learning, but no statistically significant difference between the two groups was found.

Inklings for possible reasons underpinning the lack of statistical difference between the experimental and control groups could be found in the interviews. Students indicated at the interviews that the overall feedback was not helpful to them to improve their learning techniques. They felt that the overall feedback was not useful enough in helping them to correct their weaknesses and enhance their performance in sports. Furthermore, they thought that overall feedback was not easily internalised, and internalisation is crucial for consolidating knowledge and understanding in sports skills (Black and Wiliam 1998b; Casbon and Spackman 2005; Hattie and Temperley 2007). It was possible that overall feedback was not specific enough, both for the experimental and control group, for any impact on learning to take place.

In Hong Kong, sports coaches do provide feedback to students during the sports class. Nevertheless, many coaches only give overall feedback to the whole class. Why does the situation happen? In Hong Kong, students usually trained with sports skills under traditional teaching methods, which mainly focus on mechanical drilling. Students usually receive few specific feedbacks in the traditional approach. Even though feedbacks are provided to students, they tend to be limited to overall comments at the end of the learning activity or training session, rather than during the process of learning, immediate or about specific issues. In particular, explanations as to *why* a sports action (e.g. shooting a basketball) has to be done this way or that way and *how* an action can be done better (e.g. how to shoot at a certain angle) are rarities. Students are told in the traditional approach whether or not they are performing well at the end of the learning, but not why or how to make the learning better. Overall feedback at the end of the learning session is summative and gives students no chance to practise and remedy their weaknesses.

A possible reason why formative feedback is not common in sports is that sports coaches might have a misconception that specific feedbacks could take up precious teaching time. In Hong Kong, the normal physical education class is about 70 min per class (Li 2004). In this situation, some sports coaches and physical education teachers might question the feasibility of providing specific feedback for each student or group of students.

In recent years, the researcher has observed that many sports coaches like to provide overall feedback to students during class. For example, “You have done a good job!”, “You have a great talent!”, “Your class practises well today!”, “This motion is not correct!” or “If you practise harder, you must perform better!” are common feedbacks used by sports coaches. However, such overall feedback is not constructive or concrete enough to help students correct their mistakes, proceed from “good” to “excellent” or to overcome their learning difficulties.

As can be seen from this study, effective specific feedback can be really brief (Black and Wiliam 2009; Downey et al. 2009; Evers and Spencer 2011; Hattie and Temperley 2007; Lee 2007; Miller and Nendel 2011; Tan et al. 2011; Wiliam 2011). For instance, the participants in this study found such specific feedback as, “You should perform like this” (The sports coach demonstrated with explanations), or “You have not done it correctly because…(explanations)” to be helpful in correcting their mistakes when learning sports skills. In addition, they also found specific feedback such as, “Your strategy is right because… (explanations)”, or “Please explain what you have done right to other classmates" to be helpful in enhancing their sports performance. One implication is that sports coaches need to learn to provide specific feedback efficiently (Hardman and Jones 2011; Robinson 2010), and the participants could achieve it through the study.

19.5.2 Discussion of the Long-Term Intervention Effects on Pre-service Sports Coaches' Attitudes Towards Assessment for Learning in Sports

The long-term effect of the intervention on pre-service sports coaches' attitudes towards Assessment for Learning in sports was tested using paired-sample *t*-tests (George and Mallery 2010; Green and Salkind 2011; Muijs 2011). The tests were applied to the pairs of scores from the AfL-First group's questionnaires administered at the mid-test and post-test stages of the experiment. Students in AfL-First received a 7-week intervention between administration of the pre-test and mid-test questionnaires, and then the intervention was withdrawn between the mid-test and post-test stages of the experiment. Three outcomes were possible by comparing the mid-test and post-test scores of the AfL-First group: (a) a drop in their attitudes from mid-test to post-test would indicate that the intervention had no long-term effect; (b) no change in their attitudes would indicate that the intervention effect was being sustained even after withdrawal of the intervention; and (c) an increase in their attitudes would indicate that the intervention effect was not only sustained but also has a long-term positive effect. The paired-sample *t*-tests showed that there was no significant difference between the mid-test and post-test scores of the AfL-First group in all of the 13 scales across the five domains. This means that the effect of the intervention had been sustained long term for all the scales (Damiani 2011; Echevarría and Vogt 2011; Goodwin 2010; Hoodin 2011; Hoy 2010; Norrie 2011; Richards and Leafstedt 2010; Sah 2008; Sonnet 2010; Russell 2002; Wilkinson 2010): the AfL-First group was at least as positive towards Assessment for Learning at the post-test stage of the experiment as they had been at the mid-test stage.

The results of this study showed that all three feedback scales where the attitude means increased from mid-test to post-test. In addition, students in AfL-First (the experimental group) had a more positive attitude than those in AfL-Second (the control group) after 14 weeks in overall. It is noteworthy that even though the attitudes of the experimental group had decreased in the majority of scales after returning from the Assessment for Learning intervention to traditional methods of teaching, the reduction did not result in the experimental group going back all the way to the original (pre-test) level.

The reasons for a long-term effect on the AfL-First group could perhaps be explained by the timing of the intervention (Norrie 2011; Russell 2002; Wilkinson 2010). The AfL-First group had gone through the intervention for 7 weeks, and then no intervention for the next 7 weeks. If the intervention effect was not sustainable, then we should have seen AfL-First group at the end of the 14 weeks to return to its starting point of pre-test. In this study, this had not happened in any of the 3 scales, implying that the long-term intervention effect could be sustained, even though some reduction in attitudes was observed in the majority of scales, after the intervention stopped at mid-test (i.e. after 7 weeks of intervention).

Another reason to explain the phenomenon of such long-term effect might be the John Henry effect (Heppner et al. 2008) and testing effect (Creswell and Clark 2010). The John Henry effect refers to the phenomenon that the control group (in this case, AfL-First group) that had no intervention exerted extra effort when members knew that they were in the control group and hence showed a performance above average expectation. This might have happened to the AfL-First group during the 7 weeks from mid-test to post-test where no intervention was provided. During this period, the AfL-First group might have exerted extra effort due to members' recognition that they were in the control group, or in other words, they operated under the John Henry effect.

Another possible explanation of the long-term effect was due to the testing effect. In this case, the testing effect suggests that it might be possible that participants of AfL-First group became more positive or sustained their positive attitudes when they were asked to fill in the same questionnaire a third time at post-test, a point at which group members were already very familiar with the ideas of Assessment for Learning.

19.6 Conclusion

This study has shown the implication of the long-term effect: the hands-on approach like an intervention was very effective in keeping the sports coaches' recognition, enthusiasm and positive attitude on Assessment for Learning in the long term. This is consistent with findings from other research (e.g., Craig and Deretchin 2010), which show that a hands-on approach is more effective than training only through lectures, workshops or seminars.

This study implicates that the intervention is powerful in providing the users with more opportunities to learn and practise Assessment for Learning in real teaching contexts rather than just sitting in the workshops; this way, the sports coaches can recognise its significance on student learning and keep applying Assessment for Learning on sports teaching in the long term.

References

Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box*. Cambridge: University of Cambridge School of Education.

Assessment Reform Group. (2002). *Assessment for learning: 10 principles*. Available on the Assessment Reform Group website: www.assessment-reform-group.org.uk

Berry, R. (2005). Entwining feedback, self- and peer-assessment. *Academic Exchange Quarterly, 9*(3), 225–229.

Berry, R. (2008). *Assessment for learning*. Hong Kong: Hong Kong University Press.

Berry, R. (2011). Assessment trends in Hong Kong: Seeking to establish formative assessment in an examination culture. *Assessment in Education: Principles, Policy and Practice, 18*(2), 199–211.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*, 7–74.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: GL Assessment.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5–31.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham: Open University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah: Erlbaum.

Brookhart, S. M. (2008). *How to give effective feedback to your students*. Alexandria: Association for Supervision and Curriculum Development.

Bryne, B. M. (2011). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge Academic.

Carless, D. (2002). The 'mini-viva' as a tool to enhance assessment for learning. *Assessment and Evaluation in Higher Education, 27*(4), 353–363.

Carless, D. (2005). Prospects for the implementation of assessment for learning. *Assessment in Education, 12*(1), 39–54.

Casbon, C., & Spackman, L. (2005). *Assessment for learning in physical education*. Leeds: Coachwise Business Solutions.

Chan, W. K., Sum, K. W., & Lau, K. O. (2006). Barriers to the implementation of physical education (PE) assessment in Hong Kong. *The International Journal Learning, 13*(4), 165–170.

Chappuis, S., & Stiggins, R. J. (2001). Classroom assessment for learning. *Educational Leadership, 60*(1), 40–43.

Cheung, H. P. R. (2002). *Factors influencing attitudes of Hong Kong secondary school students toward physical education*. Manchester: University of Manchester.

Choi, P. L., & Tang, S. Y. F. (2009). Teacher commitment trends: Cases of Hong Kong teachers from 1997 to 2007. *Teaching and Teacher Education, 25*(5), 767–777.

Comer, R., & Gould, E. (2011). *Psychology around us*. Hoboken: Wiley.

Craig, C. J., & Deretchin, L. F. (2010). *Cultivating curious and creative minds: The role of teachers and teacher educators*. Lanham: Rowman & Littlefield Education.

Creswell, J. W., & Clark, V. L. P. (2010). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks: SAGE Publications.

Curriculum Development Council. (2001). *Learning to learn – The way forward in curriculum development*. Hong Kong: Government Printer.

Curriculum Development Council. (2002). *Physical education key learning area, curriculum guide (Primary 1 - Secondary 3)*. Hong Kong: Government Printer.

Curriculum Development Council. (2004). *English language education key learning area: English language curriculum guide (Primary 1–6)*. Hong Kong: Government Printer.

Curriculum Development Council. (2006). *English language education key learning area: New senior secondary curriculum and assessment guide (secondary 4–6)*. Hong Kong: Government Printer.

Damiani, V. B. (2011). *Crisis prevention and intervention in the classroom: What teachers should know* (2nd ed.). Lanham: Rowman & Littlefield Publishers.

Davison, C. (2007). Views from the chalkface: School-based assessment in Hong Kong. *Language Assessment Quarterly, 4*(1), 37–68.

Downey, C. J., et al. (2009). *50 ways to close the achievement gap* (3rd ed.). Thousand Oaks: Corwin Press.

Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks: Corwin Press Inc.

Echevarría, J., & Vogt, M. E. (2011). *Response to intervention (RTI) and English learners: Making it happen*. Boston: Pearson.

Evers, R. B., & Spencer, S. S. (2011). *Planning effective instruction for students with learning and behavior problems*. Boston: Prentice Hall.

Fautley, M., & Savage, J. (2008). *Assessment for learning and teaching in secondary schools*. Exeter: Learning Matters.

Fraenkel, J., Wallen, N., & Hyun, H. (2011). *How to design and evaluate research in education* (8th ed.). New York: McGraw-Hill Humanities/Social Sciences/Languages.

Fu, F. H. (1988). School physical education in Hong Kong. *Physical Education Review, 11*(2), 147–152.

Gall, M. D., Gall, J. P., & Borg, W. R. (2010). *Applying educational research: How to read, do, and use research to solve problems of practice* (6th ed.). Boston/Hong Kong: Pearson.

George, D., & Mallery, P. (2010). *SPSS for Windows step by step: A simple guide and reference, 17.0 update* (10th ed.). Boston/Hong Kong: Allyn & Bacon.

Gilson, R. (2009). *Professional development in assessment for learning*. United States: Arizona State University.

Goodwin, C. J. (2010). *Research in psychology: Methods and design* (6th ed.). Hoboken: Wiley.

Green, S. B., & Salkind, N. J. (2011). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (6th ed.). Boston: Prentice Hall.

Hardman, A., & Jones, C. (2011). *The ethics of sports coaching*. Milton Park/Abingdon/Oxon/New York: Routledge.

Hartas, D. (2010). *Educational research and inquiry: Qualitative and quantitative approaches*. New York: Continuum International Publishing.

Hattie, J., & Temperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Henkin, A. B., & Holliman, S. L. (2009). Urban teacher commitment: Exploring associations with organizational conflict, support for innovation, and participation. *Urban Education, 44*(2), 160–180.

Heppner, P. P., Wampold, B. E., & Kivlighan, D. M. (2008). *Research design in counseling* (3rd ed.). Belmont, Calif: Thomson/Brooks/Cole.

Hewitt, D. (2008). *Understanding effective learning: Strategies for the classroom*. Maidenhead: McGraw Hill/Open University Press.

Hinett, K., & Thomas, J. (1999). *Staff guide to self and peer assessment*. Oxford: Oxford Centre for Staff and Learning Development.

Hong Kong Examinations and Assessment Authority. (2007). *Longitudinal study on the implementation of the school-based assessment component of the 2007 HKCE English Language Examination. (Final Report).* Hong Kong: Faculty of Education, The University of Hong Kong.

Hong Kong Sports Development Board. (1999). *Local educational courses in sport & physical education*. Hong Kong: The Board.

Hoodin, R. B. (2011). *Intervention in child language disorders: A comprehensive handbook*. Sudbury: Jones and Bartlett Publishers.

Hoy, W. K. (2010). *Quantitative research in education: A primer*. Thousand Oaks: SAGE Publications.

Irons, A. (2008). *Enhancing learning through formative assessment and feedback*. London: Routledge.

Jones, R., Armour, K., & Potrac, P. (2004). *Sports coaching cultures: From practice to theory*. London: Routledge.

Kidman, L., & Hanrahan, S. (2011). *The coaching process: A practical guide to becoming an effective sports coach* (3rd ed.). Abingdon/New York: Routledge.

Lee, I. (2007). Feedback in Hong Kong secondary writing classrooms: Assessment for learning or assessment of learning? *Assessing Writing, 12*, 180–198.

Li, C. (2004). *From students to teachers – A longitudinal study of occupational socialization of pre-service physical education teachers in Hong Kong*. England: University of London.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85–106.

Linacre, J. M. (2010). *WINSTEPS Rasch measurement computer program*. Chicago: Winsteps.com.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte: Information Age Publishing.

Liu, Y. K. R. (1998). *The implementation of a cognitive teaching approach to games in Hong Kong*. England: Loughborough University.

Miller, M. P., & Nendel, J. D. (2011). *Service-learning in physical education and related professions: A global perspective*. Sudbury: Jones and Bartlett Publishers.

Mok, M. C. M. (2010). *Self-directed learning oriented assessment: Assessment that informs learning & empowers the learner*. Hong Kong: PACE.

Muijs, D. (2011). *Doing quantitative research in education with SPSS* (2nd ed.). London/Thousand Oaks: Sage.

Norrie, M. (2011). *Humanitarian intervention and the United Nations*. Edinburgh: Edinburgh University Press.

Nygaard, C., Holtham, C., & Courtney, N. (2009). *Improving students' learning outcomes*. Koge: Copenhagen Business School Press.

Ormrod, J. E. (2011). *Educational psychology: Developing learners* (7th ed.). Boston/Hong Kong: Pearson/Allyn & Bacon.

Pang, N. S. K., & Leung, Z. L. M. (2010). *Teachers' competence in assessment for learning in early childhood and primary education*. Hong Kong: Faculty of Education, Hong Kong Institute of Educational Research, Chinese University of Hong Kong.

Reid, G., & Green, S. (2009). *Effective learning*. London: Continuum International Publishing Group.

Richard, G. (1991). *Manufacturing knowledge: A history of the Hawthorne experiments*. Cambridge: Cambridge University Press.

Richards, C., & Leafstedt, J. M. (2010). *Early reading intervention: Strategies and methods for struggling readers*. Boston/Hong Kong: Allyn & Bacon.

Robinson, P. E. (2010). *Foundations of sports coaching*. Abingdon: Routledge.

Russell, E. D. (2002). *The California School of Professional Psychology handbook of multicultural education, research, intervention, and training* (1st ed.). San Francisco: Jossey-Bass.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Sah, B. (2008). *Effect of intervention on the development of preschool children: An experimental study*. New Delhi: Dominant Publishers and Distributors.

Savage, J. (2011). *Cross-curricular teaching and learning in the secondary school*. New York: Routledge.

School-based Assessment Consultancy Team. (2007). *Professional development for the school-based assessment component of the 2007 HKCE English Language Examination [DVD]*. Hong Kong: Hong Kong Examinations and Assessment Authority and The Faculty of Education, The University of Hong Kong.

Sonnet, H. (2010). *Positive intervention for pupils who struggle at school: Creating a modified primary curriculum*. New York: Routledge.

Stephen, J. R. G. (1992). Was there a hawthorne effect? *American Journal of Sociology, 98*(3), 451–468.

Stiggins, R. J. (2005). *Student-involved assessment for learning* (4th ed.). Upper Saddle River: Merrill/Prentice Hall.

Tan, O. S., et al. (2011). *Educational psychology: A practitioner-researcher approach* (2nd ed.). Singapore: Cengage Learning Asia Pte Ltd.

Troman, G., & Raggal, A. (2008). Primary teacher commitment and the attractions of teaching. *Pedagogy, Culture and Society, 16*(1), 85–99.

Tuckman, B. W., & Monetti, D. M. (2011). *Educational psychology* (1st ed.). Belmont: Wadsworth/Cengage Learning.

Whitfield, A. H. (2000). *Student teacher self-assessment: A proposed method of professional development*. New Orleans: UMI Dissertation Services.

Wiliam, D. (2000). Formative assessment in mathematics part 3: The learner's role. *Equals: Mathematics and Special Educational Needs, 6*(1), 19–22.

Wiliam, D. (2001). An overview of the relationship between assessment and the curriculum. In D. Scoot (Ed.), *Curriculum and assessment* (pp. 165–181). Westport: Alex.

Wiliam, D. (2007a). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053–1098). Greenwich: Information Age.

Wiliam, D. (2007b). Content then process: Teacher learning communities in the service of formative assessment. In D. Reeves (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 182–204). Bloomington: Solution Tree.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3–14.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah: Erlbaum.

Wilkinson, L. A. (2010). *A best practice guide to assessment and intervention for autism and Asperger syndrome in schools*. London: Jessica Kingsley Publishers.

Wolfe, E. W., & Chiu, C. W. T. (1999). Measuring change across multiple occasions using the Rasch rating scale model. *Journal of Outcome Measurement, 3*(4), 360–381.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analyses*. Chicago: MESA Press.

Chapter 20
The Case of St Margaret's Girls' College: How SLOA Promotes Self-Assessment and Peer Assessment to Enhance Secondary School Student English Learning

George C. Yu

In meeting the challenges of globalization, education reform often targets assessment because it is "key to improving learning achievement" (UNESCO EFA Monitoring Report 2005, p. 158). The 3-year Self-directed Learning-Oriented Assessment Project (SLOA) aimed at changing the assessment culture in Hong Kong by enhancing assessment literacy and practices. In recognizing the key role of teachers in quality education (Leu et al. 2004), site-based professional development serves as the cornerstone to engineer teachers' behavior change and to impact on student achievement. This chapter takes a good look at how the quality of English teaching and learning at one of the project's partner schools was improved. Of particular interest is to examine how the participating teachers were empowered to take a leap of faith in adopting both the rationale and practices of student self-assessment and peer assessment, in order to enhance student motivation and learning results for English language acquisition. Both quantitative and qualitative data were collected and analyzed. Discussions are offered at the end of the chapter on student achievement, professional growth of the teachers and their implications for other schools.

The St Margaret's Girls' College (SMGC) is an English medium secondary school located at the Mid-level on the Hong Kong Island. It has an enrollment of about 480 female students and 29 teachers. Throughout 2006–2008, it participated in the Self-directed Learning-Oriented Assessment Project (SLOA) in a partnership with The Hong Kong Institute of Education Centre for Assessment Research and Development (CARD) (Mok 2010). The team of five English teachers at SMGC worked diligently and closely with the designated Assessment Development Officer from CARD in implementing the new assessment initiatives for the project.

G.C. Yu (✉)
Director and Senior Consultant of the Hong Kong Language and Culture Institute,
Centre for Assessment Research and Development, The Hong Kong Institute
of Education, 23D Tower 1, Harbour Place, 8 Oi King Street, Kowloon
e-mail: georgeyu@ied.edu.hk or gyu5279@gmail.com

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_20, © Springer Science+Business Media Dordrecht 2013

20.1 The Aims of the SLOA Project

In meeting the challenges of globalization, Hong Kong's education reform seeks to bring about profound changes in conceptions of learning, curriculum design, and content standards. Changes in the assessment of student learning must go hand in hand with other education reform initiatives, for "regular, reliable and timely assessment is key to improving learning achievement" (UNESCO EFA Monitoring Report 2005, p. 158).

The assessment project aims at changing the assessment culture in Hong Kong by developing assessment literacy and practices in three dimensions of assessment, namely, assessment *of* learning, assessment *for* learning, and assessment *as* learning (Mok 2010). The project focus is to integrate more ongoing, formative assessment within teaching and learning so that the examination, as the typical form of summative assessment used mainly for student selection or certification (Somerset 1996), ceases to be the only means to gauge learning achievement. The following table puts the two forms of assessment in vivid contrast:

	Summative assessment	Formative assessment
Purpose	To evaluate and record a learner's achievement	To diagnose how a learner learns and to improve learning and teaching
Judgment	Criterion-referenced or norm-referenced; progression in learning against public criteria	Criterion-referenced and student-referenced
Method	Externally devised tasks or tests; reviewing written work and other products (portfolio) against criteria applied uniformly for all learners	Observing learning activities, discussing with learners, reviewing written work and other products (portfolio), learner self-assessment and peer assessment

Sources: Harlen and James (1997), Black et al. (2003)

The project is to achieve its goal through applied research, site-based professional development, and workshops and seminars for teachers and parents (Mok 2010). What happened at SMGC highlights how the assessment project has impacted on the quality of education at this school.

20.2 Site-Based Professional Development

Among all the activities conducted under the project, the site-based teacher professional development is the cornerstone effecting changes in school. Therefore, a research-based professional development model was adapted to secure high-quality interface between CARD and the school team. In recognizing the key role of teachers in quality education (Leu et al. 2004), much energy was directed toward ongoing professional development consistent with the switch of priority from initial teacher training at universities to continuing, in-service education at the school site (OECD 2004a).

In order for teacher education to be effective, the following strategies have been long recognized by experts in the field:

- Most training both formal and nonformal takes place in schools, where trainees observe, assist, and teach.
- Training occurs throughout the teacher's career.
- Training emphasizes actual classroom teaching behaviors.
- Self-study and self-learning are critical.
- Groups or cohorts of teachers are trained together.
- The inspection system supports good teaching practice.
- Training begins with teachers identifying needs and demands.
- Reform of teacher education is an integral part of curriculum and other reforms (Adapted from Craig et al. 1998).

The assessment project adopted some of these strategies to enhance the effectiveness of its on-site professional development work with the partner schools (Mok 2010). In the case of SMGC, the following parameters were defined in consultation with the school:

- The bulk of the consultations took place at the school in the form of face-to-face meetings on a monthly basis. The meetings were scheduled by mutual agreement. The time and place remained constant throughout the project.
- The meeting agenda consistently focused on developing practical concrete classroom assessment tools to guide teaching and learning. Problems and challenges were brought to the table, and solutions were strategized. Meeting minutes were shared with all stakeholders on a timely basis.
- The practice of self-reflection and self-assessment at regular intervals allowed time for the teachers to consolidate learning and refocus. Each term, the teachers completed three reflective journals to document their own learning.
- The teachers worked in a team environment where sharing and brainstorming further strengthened the bond and trust among the team members.
- The school team had opportunities to select topics for discussion and exploration based on their own professional needs.

However, no good professional development strategy would function well without the support from the school leadership (Cummings and Brocklesby 1997; Williams 1997). Schools have to become learning organizations where their leadership prioritizes learning and harnesses the different capacities of teachers to address common learning difficulties (Sayed and Jansen 2001). The SMGC made good of their commitment to the project by delivering the following:

- Allocation of time for the members on the team to meet and coplan lessons
- The presence of a vice principal at all the meetings while the principal paid close attention to progress of the project
- Accommodation within the school's established assessment structure for more formative assessment to allow for innovation
- Decision to participate in all the semi-annual English assessment and student attitude surveys administered by CARD

The team focused its effort on further developing English reading and writing skills for the targeted secondary 3-student group. The two teachers for the target grade were Mr. Shane W. Early and Ms. Mee Ling Lam; however, since all the English teachers in the school participated in the meetings and discussions, the impact of the project reached the entire student group. The following are the key discussion topics for the on-site meetings on how to enhance teaching and learning via formative assessment:

- Applying scoring rubrics with consistency and accuracy
- Applying student self-reflective exercises to nurture self-directed learning capacity among students (Mok and Lung 2005)
- Applying prewriting strategies to better prepare students for writing tasks
- Developing student capacity for self-monitoring, self-regulation and self-directing in their learning process (Garrison 1997)

20.3 Student Self-Assessment Practices

The ultimate target of the SLOA Project remained the improvement of student learning results, while the ongoing professional development and support aimed to nourish the teachers and get them ready to try out new assessment techniques in their classrooms. Good understanding of the SLOA concepts and practices on the part of the teachers led to effective integration of the new practices with the existing curriculum and school conventions. Among the key, SLOA practices are those of student self-assessment and student peer assessment (Mok 2010). In the context of SMGC, the following instruments of self-assessment were implemented effectively and henceforth highlighted here:

- Scoring rubric for English writing
- Reading strategies self-assessment checklist
- Student self-reflective journal

20.3.1 Scoring Rubric for English Writing

Student self-assessment exercises are those intended to enhance students' self-awareness of their relative strengths and weaknesses under a noncompetitive and low stake frame of mind. In a typical scenario, under more traditional summative assessment, student learning is assessed at the end of a learning period, i.e., a term, a course, or a program. An arbitrary score would be given to reflect the ranking of the assessed student among his or her peer group. Very little useful information is provided to guide the student for further improvement, or the feedback comes too late as judgment is already rendered. This dilemma presented challenges to the English teachers at SMGC as they strove to deliver effective English writing instruction.

Based on their training under the SLOA Project, the scoring rubrics gained the teachers' buy-in and they decided to trial them in teaching English writing.

For an English learner, writing presents multiple challenges as the student has to tackle tasks related to different areas of competency (i.e., content, organization, language, etc.). Often each student seems to have only a vague sense of what constitutes a good piece of writing. A writing rubric can help students to articulate the specific performance standards/criteria for each level of writing competence. It empowers the student to know how he or she can improve and achieve certain desired results throughout the writing process. The steps of implementing the writing rubric successfully at SMGC are illustrated below.

20.3.1.1 Student Buy-in

The teachers realized from their own experience that if they simply designed the world's best writing rubric and posted in the classroom, the impact would be minuscule as the students would treat it as another set of top-down arbitrary performance standards. For students to buy in, the teachers introduced a discussion topic "How to write a good piece of writing in English?" to the class. Everyone brainstormed and contributed their own ideas while the teacher facilitated the discussion. Based on the discussion notes, a draft scoring rubric for English writing gradually took shape.

The scoring rubric for English writing developed by the SMGC teachers and used by their students uses a simple format. Vertically it lists five levels of performance per target student group. Horizontally it illustrates what the specific criteria are, in terms of performance categories, such as content, language, and organization. The number of categories and levels can all be adjusted to reflect different purposes and student proficiency levels. Generally speaking, if the students are more advanced in their English studies, the performance descriptors/criteria also become more elaborate and sophisticated, in order to document the wider range of proficiency levels or writing skills.

20.3.1.2 Student Practice Runs

At the beginning, in order to help the students to familiarize with the rubric, teachers intentionally made frequent references to the criteria listed. Model writing samples were also introduced to give concrete examples for the various standards.

Next, students were asked to score writing samples and discuss their scores in their small group. They had to achieve group consensus on scoring the same writing sample by articulating their reasoning, negotiating their understanding, listening to feedback, and reaching consensus. This exercise allowed the students to practice using the rubric to judge a piece of writing, based on its merits or lack thereof. This set the stage for the next step, self-assessment via the scoring rubric.

20.3.1.3 Student Self-Assessing/Scoring

Based on their involvement in and contribution to the development of the scoring rubric, student interest got further boosted by their growing expertise in using the rubric to render judgment on writing samples. This experience allowed next step, to guide and measure their own writing, to be easily taken since the students had become far more articulate and familiar with the standards that they could judge themselves by the same standards.

20.3.1.4 Postwriting Reflection

As one of the SMGC teachers insisted, "Reflection for students after writing is needed." This enables large amounts of learning to be captured by the student and sorted out, recognized, owned, valued, consolidated, and kept for later use. Opportunity must be provided to students to reflect upon their specific learning experience whether it is, for example, progress in articulation, a relative weakness in vocabulary, or a lack of background knowledge. At this point, it is important for teachers and other stakeholders (such as parents) to convince the students that their voice, opinion, and judgment are all important and valued by the adult world. This helps to reduce the chances of an adverse relationship developing between the student, on one hand, and the teacher and parents, on the other. In concrete terms, the students were invited to write in their own words their reflections and their future goals for improvement.

The reflection could be facilitated by a simple questionnaire in which the student is asked to answer guiding questions such as:

- Out of a score 1–10, how much effort did you make for this assignment?
- How satisfied were you with your finished product?
- Where have you made most improvement?
- What areas will you focus your effort on for improvement?
- What is the most important thing you have learned from this assignment?

20.3.1.5 Teachers' Feedback on Using the Scoring Rubric

The teachers at SMGC were overwhelmingly positive about their experience using the scoring rubric in their English writing class. The following summary of their feedback was based on their on-site meeting minutes, e-mail exchanges, and their own publications.

The implementation of the scoring rubric must be in line with student-centered pedagogy in order to be effective in empowering the learner. Patience must be exercised and mini-steps taken to give students time and space to buy in. Educators must bear in mind that the majority of secondary school students have already developed coping mechanisms to work with the more traditional assessment regime, which tend to leave out students' own learning experience.

The scoring rubric lends the same language to both the students and the teacher so that they can work closely together to facilitate more learning. The same language and criteria allow the students and the teacher to interact more meaningfully. Their verbal or written exchange tends to be more in-depth, precise and focused.

Due to the increased amount of positive involvement in the teaching and learning process, most of the students have become more active learners. They tend to develop better self awareness, become more confident and self-motivated.

The development of scoring rubrics is very labor-intensive. While a generic version provides the comprehensive criteria for English writing in general, writing for a particular purpose (i.e. a business letter vs. a marketing brochure, an expository article or commentary) calls for customizing it to fit to the task. Emphasis on a specific function needs to be stressed at a certain juncture in teaching and learning so that a particular skill and competency can be targeted. This is of particular importance for the ELL students who must marshal their limited language resources to "zero in" on an area of further growth. However, writing a different scoring rubric for each special function would prove to be too cumbersome and unwieldy. Undeterred, the teachers at SMGC quickly came up with a creative solution. Depending the particular emphasis desired, the teacher would choose from Content, Language and Organization one area to be further elaborated. By re-defining the criteria for what constitutes excellence in Content, Language or Organization, the students, led by the teacher, would come to new understandings of writing excellence. The resulting effect facilitates the development of student language skills, mimicking the process of language acquisition whereby a student's knowledge and skills increasingly become more concrete, specialized and sophisticated. The creative teachers at SMGC had a good name for this customization, namely, "Targeted Competency."

20.3.2 Reading Strategies Self-Assessment Checklist

In the past, emphasis in an English reading class tended to focus only on developing language competency, be it literacy skills, vocabulary, or decoding techniques. Students tended to be rushed into reading with barely enough time to familiarize themselves with the new vocabulary and expressions. The typical mode of operation seemed to be "jump first and ask questions later."

However, in a more progressive reading classroom, *frontloading* becomes an important part in preparing students for the approaching task of reading in English. According to Dr. Wilhelm (2002), frontloading activities can be used before reading to measure what conceptual, linguistic, or genre knowledge a student may require to succeed on subsequent reading tasks. Thus, a quick survey of students' levels of preparedness would necessitate the monitoring or revision of instructional activities and materials to respond to student needs. Henceforth, the use of reading strategies self-assessment checklist fits in the role of frontloading quite nicely.

This checklist contains some of the most basic reading strategies that can benefit students already in secondary school. As research (e.g., Rosenshine and Stevens 1986;

Almasi 2003) indicates, explicit instruction on teaching students how to use various reading strategies has led to improved learning results for both first language speakers and second language learners alike. As part of the frontloading exercise, the teacher can introduce these reading strategies at the beginning of a reading assignment, a reading course, or any English class which has a reading component.

The reading strategies self-assessment checklist was introduced to the teachers at the SMGC as part of the on-site professional development. The teachers provided input for further modifications, and extensive discussion was conducted to explore ways to apply this checklist effectively in their classes.

One of the biggest ironies in the second language classroom is that students are often assumed to have very little to contribute to the learning process. Regardless of their age, they are treated like children knowing nothing about the target language. However, each student has accumulated a sizable collection of linguistic and cultural knowledge and developed skills for acquiring new knowledge and information as they move through grades at school. Some of this knowledge is universal and true, crossing language and cultural barriers. This universal knowledge and its associated skills can be utilized to provide a tremendous boost and insight in aiding the acquisition process of the target language. It is in this context that the introduction of the checklist to the students at SMGC turned out to be very effective; it uncovered and released this powerful wealth of knowledge and skills to boost the students' English reading skill development.

All the reading strategies in the checklist aim to facilitate the student to fulfill one of the four roles/functions of the reader, first proposed by Freebody and Luke (1990).

20.3.2.1 Text Participant

Here the student uses the relevant strategies to comprehend written text. In concrete terms, he or she uses their prior knowledge to make meaning throughout reading. In bringing meaning forward, he or she makes predictions and monitors/modifies those predictions pending new information.

The student sets purposes for reading, making reading a powerful act of inquiry and intellectual pursuit. They then make increasingly complex inferences based on simple and complex relationships. They can also construct and understand characters and their evolving relationships.

The relevant strategies from the checklist are:

- *Activate prior knowledge*: I ask myself what I already know about the topic when I approach a new text.
- *Set a purpose/reason/goal for reading*: I ask why I am reading this. Is it for pleasure, an assignment, or just gathering information?
- *Make personal connections:* I compare and contrast my knowledge and experience with what is presented and revealed in the text.
- *Make predictions:* I look for clues about a new story from the title, table of contents, dedication, number of pages, font size, photographs, commentary, etc. I check and revise my initial reactions and predictions as my reading progresses.

In teaching these strategies, the teacher must take time to solicit relevant experiences and stories from the students. The students invariably need teacher affirmation and recognition of their relevant skills and prior knowledge, developed through their first language literacy, as being "given permission" to actively participate in the reading process by engaging their personal experience. In other words, it empowers students.

20.3.2.2 Text Code Breaker

In this role, the student strives to decode various codes and conventions of written English. To do so, they must recognize the rules governing word formation (e.g., phones, phonemes, roots, suffixes, prefixes), sentence order (e.g., parts of speech, word order, clauses), and narrative conventions (e.g., prose, poetry, business writing, legalese).

In the checklist, there is only one relevant strategy in this area listed:

- *Decode text into words and meanings*: I still work hard to define new vocabulary by using contextual and lexical clues (e.g., prefixes, suffixes, word roots)

Essentially, all the grammatical rules and stylistic conventions could be listed here, but, since they make up the bulk of second language acquisition, this would overwhelm a young reader. Although the reader will have developed the strategies in tackling his or her mother tongue (Chinese) – in terms of its mechanics and technicalities – these are language-specific. Therefore, students need to find a way to strategize using the new conventions for the target language. For example, in Chinese all the modifications or descriptions of something must be loaded in front of the noun. In English, however, a single word or phrasal modifier can also be loaded in front, but anything longer than this must be placed *after* the noun.

20.3.2.3 Text User

In this role, the student learns to understand the purpose of different written texts for various cultural and social functions. As well as learning to identify the various features of a particular text type – its style, structure, or genre – students also work to differentiate the personal purpose of writing a diary versus the possible social purposes of publishing a personal journal.

There two relevant strategies from the checklist that relate to this:

- *Apply what has been learned*: I always ask myself, "How can I use this information?" "Is my reading useful just for my class or is it applicable to my life?"
- *Ask questions*: I have the habit of asking questions about the text, the writer, and even my own responses. I work through my confusions and get a clear understanding by rereading difficult parts or get help from others.

20.3.2.4 Text Analyzer

Here, the student aims to understand how the text interacts with readers by interpreting characters and their perspectives. They will also try to gain insight into the particular cultural and social context in order to fully grasp the significance of the writing.

There are also two relevant strategies from the checklist:

- *Visualize*: I try to create a mental picture of the setting and imagine what the characters would look like while reading a story. I also use visual symbols, concept webs, or mind maps to keep track of the information and organize it if the text is abstract.
- *Summarize and clarify understanding*: I collect and store key pieces of information along the way to help myself make sense of what I read. I review these collected items regularly in order to understand the main ideas/plot of the story and evaluate the text properly and accurately.

It must be pointed out that both text user and text analyzer represent the status of a more sophisticated reader. As the fluency level of the target students at SMGC continues to improve, their teachers are fully aware that relevant strategies will be added for emphasis.

20.3.3 Implementation of the Reading Strategies Checklist

The initial buy-in process asked the students guiding questions to solicit personal experiences and anecdotal accounts of how they had applied similar strategies in other fields like Chinese or science. After a question-and-answer session, during which students became quite familiar with the strategies, the teacher modeled one of two strategies using the “think aloud” technique. In this, the teacher verbalized what was going through his or her head as he or she tackled a particular reading passage. This modeling or demonstration set an example for students to try out these strategies on their own.

In addition, the checklist can be used as a self-assessment instrument. At the beginning, midpoint, and end of the reading course, students can be surveyed to indicate how effectively they are applying these strategies. Class discussions can be conducted to maximize the sharing of good applications and success stories.

20.3.4 Student Self-Reflective Journal for Drama Class

One of the English teachers who taught English drama at SMGC took the initiative to apply the concept of using written reflection as a way of assessing her class on a regular basis. She reported that her students "all felt empowered to reflect on their own learning instead of taking a paper-pencil test." She further stated, that "the level of participation in classroom and articulation of drama theories and performance techniques went through the roof." A large collection of student reflective journal entries and corresponding comments by the teacher indicated an exceptionally high level of interactivity of the class.

20.3.4.1 Background

In response to the implementation of New Senior Secondary (NSS) Curriculum, many schools in Hong Kong were preparing for their new English curriculums and electives. A number of schools included drama as a component. However, much focus was put on the final outcome – a stage performance or show performed by a few students, instead of the actual learning process, varied experiences, and achievement of ALL students throughout the learning process. Students from secondary 1 (S1) to secondary 3 (S3) have to take one 35 min long drama lesson per week, while students from secondary 4 (S4) have to take one 40 min long drama lesson per week.

Under the SLOA Project, the English teachers at the school explored new innovative ideas and practices to enrich student experiences with school assessment and improve their learning results. The use of the student reflective journal for the drama class was one such example.

The purpose of the reflective journal is to encourage students, teachers, and schools to reflect on the learning and teaching of drama lessons and see what students learn and how they learn, as well as to draw attention to the elements students are fond of and cherish, which otherwise might have been left hidden or overlooked. In this sense, it sends a powerful positive message to the school community that:

- Students' feedback is valued for further enhancing teaching.
- Knowing how to learn is just as important as what to learn.
- Different opinions and creativity should be acknowledged and celebrated.
- Learning happens in various forms and formats, and its results may not be measured by a test.

This action research project using the reflective journal to promote self-directed learning was conducted in the school year 2007–2008. Over 1,000 journal entries have been completed and turned in by the students from S1-S3, and the results compiled and analyzed. The depth and width of the students' reflections is very impressive. The success of this application can shed light on the importance of self-reflection and self-directed learning (SDL) as a powerful motivator for learning, as well as further explore and enrich the ways that drama education can be conducted in the secondary curriculum.

20.3.4.2 Methodology/Procedure

Students in the project have drama education as a compulsory subject at school. Students from secondary 1 (S1) to secondary 3 (S3) have to take one 35 min long drama lesson per week.

Students are required to hand in one drama entry after each drama lesson. A template is given to students in advance. Students are required to write 8 to 12 entries for the course, depending on the scheme of work and the learning progress of the class.

Before students write their first entry, a marking scheme for the journal entries is fully explained and a template is given to them in advanced. Students are expected to respond to four questions in the journals. The questions include:

1. What did we do today?
2. What have I learnt today?
3. What did I enjoy the most in the class today? Why?
4. I could have done better on….

Using these questions as open-ended stimuli, students are instructed to write at least 8 lines for each entry (minimum 2 lines in response to each question). The teacher then collects the journals from the students and gives feedback or personal comments on issues students have brought up. At the end of a term, the journals will contribute to 40% of the final subject grades for the students.

20.3.4.3 Impact

One of the main purposes of the drama reflective journal is to serve as a diagnostic assessment on students' learning. Being formative in nature, this diagnostic assessment can help a teacher to measure a student's current level of knowledge and skills, gauge the changes/progress being made, and ascertain interest and value of the students. Based on this ongoing stream of feedback, the teacher can constantly update his or her knowledge on how effective the lesson has been, how students are learning, and what they need to have for further progress. Therefore, the well-informed teacher can then modify his or her teaching and develop future lesson plans accordingly. Students, on the other hand, can also benefit from the assessment, since constructive feedback would be given by the teacher and exchanges of personal experience and insight between the students, the teacher, and even parents can be made possible.

Another effective function of the drama reflective journal is to encourage self-directed learning (SDL) among the students. The students and their voices become the center of attention. Drama education is a personal thing. It provides different students with varied learning experiences. Therefore their achievements are also colorful and take different shapes and sizes. Thus, the self-reflective journal can truly cater to different levels of students. Students can choose what to express, how to express it, and how much they want to express.

20.3.5 How to Put It All Together: A Lesson Plan in the Teacher's Own Words

Subject: S2 English

Class: 2Q (with 32 students)

Unit: # 6: A Mystery

Teacher: Mr. Shane Early

The plan for Unit 6 is to have students write a mystery story. I will cover a short mystery story in class first, then give the writing assignment and finally use a specially designed scoring rubric to assess the story writing. The best mystery stories written by the students are to be placed on the English Club Board.

A short story written by Rohl Dahl will be introduced as a sample reader. It will be covered in three days. Students will review the story, actions, characters, and how the writer uses written language to achieve suspense and lead the reader into the story. Students are given a basic outline of a mystery story with a few printouts from Longman to help them write a story over the Chinese New Year holiday.

During the unit on short story, many students have to be instructed on how to read a story without focusing word by word. On their own the students would not have really read the story. Grammar topics are covered in the unit and references back to the short story and police questioning techniques are to be introduced to connect the mystery stories and the grammar items.

The students are very weak in reading and don't understand how to read for enjoyment. With grammar they need direction to see how it relates to the reading we do in each unit. The teacher has to be creative in meeting the student needs.

The composition corrections and self reflection paragraph are ok. Some students are taking it seriously and thinking about how to improve next time. Others will just write a quick short paragraph with what they think the teacher wants to hear.

Many students are more involved than usual for this unit. The guidance for mystery stories and how to write them seems to be more in depth.

Peer review, self assessment, and teacher's comments/feedback on the rubrics are used to help the students to get a better understanding of their own writing. The students are able to follow the story with my help. The compositions are above average for this class. More effort has been put into this writing exercise by many students.

The peer assessment was difficult as the students did not really understand what to do first time. The peer assessment needs to be rewritten and reviewed better next time. The assessment is intended to focus on the writing and allowing the students to review their writing, correct it and reflect on it.

The writing of a mystery story appears to be a hit with the students. Many are eager to hand in their stories after the holiday. The scoring rubric which is distributed to each student for composition corrections requires the students to write a reflection of how they can improve the story. This will be written on the bottom of the scoring rubric.

Instruct, explain, review, write, review, explain, guide and review is the basic plan. I should be following this format with each unit. I need to give my students more time and guidance in the beginning of each unit and constantly ask them to review and reflect towards the end.

20.3.6 Student Peer Assessment Practices

Parallel to the self-assessment being practiced at SMGC, many good efforts were also made to pilot instruments of peer assessment. Generally speaking, both peer and self-assessments are two variations under the Self-directed Learning-Oriented Assessment. Both aim to enhance the capacity of the student to monitor, regulate, and direct his or her own learning. Many self-assessment exercises or formats can easily be turned into peer assessment practices with minimum modifications. A few practical examples for implementing the peer assessment will be examined and highlighted here.

In teaching Unit 3, the goal is to train students to develop skills to formulate and flesh out a formal business letter for lodging a complaint. After introducing a few sample letters of the same nature, the teacher will facilitate a discussion among the students to crystallize the main ingredients for what constitutes a letter of complaint. Soon a scoring rubric takes shape and it is distributed to the students. Similar to any rubric for writing, it defines the performance criteria for contents, language, and organization. The particulars regarding the letter's purpose (e.g., nature of complaint, whereabouts of the event, resolution demands), language (e.g., business-like tone, simple and direct approach), and format (e.g., strict formula for business letter writing) are emphasized. Once a draft is produced, the element of peer assessment can be implemented.

In order to further enhance student capacity to engage in self-directed learning activities, the peer assessment approach can give them opportunities to practice using the given criteria to measure a work in progress by someone else. If organized and supervised properly, the students tend to gain in the following aspects:

- They become more articulate about the new assessment language.
- They become more objective in measuring themselves as well as others.
- They also become more proficient in spotting common mistakes.
- They become more skillful in communicating with their peers about learning.

So at this juncture, the teacher hands out copies for "Feedback Sheet – A Letter of Complaint" (Unit 3). The students then work in pairs with their partner. Each will closely read through their partner's draft letter and then document their observation regarding each criterion. After completing this form, the assessing student will verbally provide feedback to his or her partner being assessed. The three levels of performance (e.g., needs improvement, satisfactory, and well done) are self-explanatory and easy to use.

The peer proofreading checklist allows the student an extra pair of eyes to spot careless or gaping errors at a quick glance. Similar instruments can also help the student to see/hear what elements make an oral presentation powerful and effective. Due to the peer pressure and rivalry, it is believed that this exercise will provide the average student an added motivation to produce a piece of work up to their current level of capability.

Generally speaking, peer assessment requires a certain level of maturity among the students. Close supervision is required to keep the students on task. Sometimes the pairs can switch partners resulting in them getting a second or even third opinion. Peer assessment often takes a lot of time in class; hence it should be carefully planned and properly executed.

20.3.6.1 Student Attitude Survey

In order to gauge the impact of the project, partner schools under the SLOA Project were also invited to participate in a specially designed survey to measure any changes or progress during the three years of the project. Limited by its sample size and other statistical constraints, the findings will be interpreted carefully. They nevertheless enrich our discussion regarding the effectiveness of the new assessment thinking and practice.

Thanks to the full support from the leadership and willingness of the teachers to take up additional workload, the cohort of students at SMGC participated in the Student Attitude Survey twice (survey 1 in 2006 and survey 2 in 2008, respectively). The results of the data analyses are presented below.

Developing student capacity to be independent, self-directed in their learning processes will empower them to take initiatives and prepare for the journey of lifelong learning (Gibbons 2002). However, students can develop into independent learners only if they have a clear picture of the strengths and weaknesses of their learning processes and know what strategies can help them improve. That means that knowing *how* to learn is just as important as *what* to learn. Metacognition, therefore, is crucial for self-directed learning. The "My Self-directed Learning Experience" is a student learning attitude survey intended to investigate students' self-reflective capacity regarding learning and to track metacognitive development among Hong Kong students.

The survey questionnaire contains six major sections, printed in Chinese for easy comprehension and input. The sections are:

1. My Academic Monitoring (10 items)
2. My Strategies of Learning (10 items)
3. Academic Self-Concept (5 items)
4. Attribution – Failure (4 items)
5. Attribution – Success (4 items)
6. An open-ended question about assessment for learning

Table 20.1 Means of all subscales in survey 1 and survey 2

Subscale	Survey 1 (S2) (N = 29)	Survey 2 (S4) (N = 35)	Difference (survey 2–1)	t	p
Academic monitoring	2.60	2.63	0.03	–0.315	0.754
Strategies of learning	2.75	2.86	0.11	–1.329	0.189
Academic self-concept	3.06	3.04	–0.02	0.161	0.872
Attribution – failure	2.69	2.76	0.07	–0.517	0.607
Attribution – success	2.75	2.69	–0.06	0.538	0.593

Note: The response scale comprises a 4-point Likert scale with categories: *SD* strongly disagree, *D* disagree, *A* agree, and *SA* strongly agree. These categories are coded as 1, 2, 3, and 4, respectively

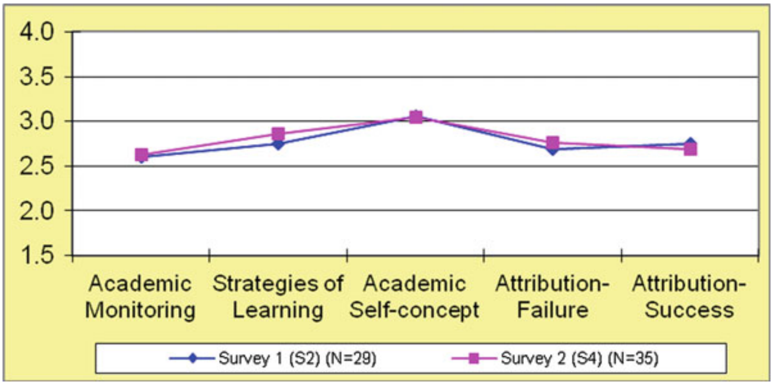


Fig. 20.1 Means of all subscales in survey 1 and survey 2

A summary of the results based on Section 1 to Section 5 is presented here. The response scale for all items comprises a 4-point Likert scale coded as 1, strongly disagree; 2, disagree; 3, agree; and 4, strongly agree. The coding is such that a higher mean rating represents a more positive attitude toward self-directed learning than a lower mean rating.

For SMGC, a total of 29 students (secondary 2) participated in survey 1 in February 2006, and the same group plus a few new comers (totaling 35 in secondary 4) participated in survey 2 in March 2008.

20.3.6.2 Results on Subscales

There are five subscales in the questionnaire. The mean ratings of the five subscales for the same cohort of students are presented in Table 20.1.

As indicated in Table 20.1 and Fig. 20.1, the mean ratings of all subscales in survey 1 and survey 2 are higher than 2.5. Overall, the results suggest that students in this school have positive attitudes toward self-directed learning and healthy attribution (strategy) to their academic failure and success.

The comparison between two surveys suggests that students have higher mean ratings in survey 2 than in survey 1 regarding three of the subscales, namely, Academic Monitoring, Strategies of Learning, and Academic Self-Concept. While this may seems to be rather insignificant statistically speaking, some researchers (Mok et al. 2006) have, however, found a trend of steady decline in academic motivation among junior secondary school students. Their study, involving over 14,000 secondary students from 23 schools, indicated that students were progressively less motivated as they moved to higher grade levels.

It appears then, that the students who took the surveys at SMGC show a remarkable resilience in maintaining their positive attitude toward their study through their secondary school years. It might not be too far-fetched to conclude that all the activities that aimed to boost their self-directed learning capacity via various practices of self-assessment and peer assessment have had a positive impact on these students.

20.4 Overall Teachers' Perspective and Reflection

The following findings are gathered from meeting minutes and teachers' reflective journal entries, submitted by the teachers:

20.4.1 On-Site Professional Development Opportunities Due to the Project

All five members of the school participated regularly in the on-site meetings. They brought their lesson and assessment plans to the table for discussion and feedback. Without the project, the teachers would not have had the "luxury" to meet and brainstorm ideas for their classroom assessment. The support and reflective environment enabled the teachers to open up and try new ideas and concepts. Some of them went as far as suggesting mini action research projects to implement self-directed learning among their students, even though these initiatives fell outside the project target student population.

20.4.2 Participation in Additional Professional Development Outside School

The teachers also took turns to participate in other professional development activities away from the school, such as the English Day Camps, seminars, and the end-of-project assessment conference organized by CARD.

This is a typical reflection: "*I attended one seminar and found the group discussions with teachers from other secondary schools helpful. In the future I hope more sharing for secondary schools will be offere*d."

20.4.3 Changes in Assessment Practice at the School

The shift from the traditional summative assessment toward the new formative/learning-oriented assessment takes both time and tremendous effort and coordination among all stakeholders in the learning process. The following is a list of features of the *assessment for learning* rationale for this paradigm shift:

- Provide teachers with information about student progress.
- Provide teachers with feedback to enhance their teaching effectiveness.
- Provide opportunity for teachers to give feedback to students.
- Provide feedback to students for their self-monitoring and subsequent learning.
- Assessment is an integral part of self-directed learning.

A few quotes from the teachers' reflective journals concur:

> I developed many tools for classroom use. My students have gotten used to the new methods. When I don't use them, I can see where a lack of assistance hurts their performance. The scoring rubrics and other materials give them a clearer picture of what is expected of them.
>
> Our English Department has taken the lessons learned from this project and made a few changes to how we do things. One example is the Paper I Writing exam. In the past we had two writings for the exam. Now we use one. We will also include scoring rubrics for all compositions in the following year.

Finally, feedback from the school principal:

> The scoring rubrics and reference materials given are helpful to us. As the teachers reported, their students have shown much improvement in writing and they are also trying out peer assessment during oral classes. Not only have our S2 students benefited from the project, but also the English teachers and students of other forms. If possible, I would like to request the Assessment Development Officer from CARD to conduct regular lesson observation in the coming year.

References

Almasi, J. F. (2003). *Teaching strategic processes in reading: Solving problems in the teaching of literacy*. New York: Guilford Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead: Open University Press.

Craig, H., Kraft, R. J., & du Plessis, J. (1998). *Teacher development: Making an impact*. Washington DC: USAID/The World Bank.

Cummings, S., & Brocklesby, J. (1997). OR beyond the sandbox. *Journal of the Operations Research Society, 48*(4), 454–457.

Freebody, P., & Luke, A. (1990). Literacy programs: Debates and demands in cultural context. *Prospect: Australian Journal of TESOL, 5*(7), 7–16.

Garrison, D. R. (1997). Self-directed learning: Toward a comprehensive model. *Adult Education Quarterly, 48*(1), 18–33.

Gibbons, M. (2002). *The self-directed learning handbook: Challenging adolescent students to excel*. San Francisco: Wiley.

Harlen, W., & James, M. J. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education, 4*(3), 365–380.

Leu, D. J., Kinzer, C. K., Coiro, J. L., & Cammack, D. W. (2004). Toward a theory of new literacies emerging from the Internet and other information and communication technologies. In R. B. Ruddlel & N. J. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 1570–1613). Newark: International Reading Association.

Mok, M. M. C. (2010). *Self-directed learning oriented assessment: Assessment that informs learning & empowers the learner*. Hong Kong: Pace Publications Ltd.

Mok, M. M. C., & Lung, C. L. (2005). Developing self-directed learning in student teachers. *International Journal of Self-Directed Learning, 2*(1), 18–43.

Mok, M. M. C., Moore, P. J., & Kennedy, K. J. (2006). The development and validation of the self-learning scales (SLS). *Journal of Applied Measurement, 7*(4), 418–449.

OECD. (2004a, November). Executive summary. In *Teachers matter: Attracting, developing and retaining effective teachers* (p. 3). Paris: OECD.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. Wittock (Ed.), *Handbook of research on teaching* (pp. 376–391). New York: Macmillan.

Sayed, Y., & Jansen, J. D. (2001). *Implementing education policies: The South African experience* (p. 221). Lansdowne: Juta and Company Ltd.

Somerset, A. (1996). Examinations and educational quality. In A. Little & A. Wolf (Eds.), *Assessment in transition: Learning, monitoring, and selection in international perspective*. Oxford: Pergamon-Elsevier.

UNESCO EFA Monitoring Report (2005). *The quality imperative* (p. 158). UNESCO: Paris.

Wilhelm, D. J., & Wilhelm, J. (2002). *Action strategies for deepening comprehension*. New York: Scholastic.

Williams, R. B. (1997). *The relationship between personal characteristics and situation complexity and decision-making style flexibility in New Brunswick school principals.* Dissertation Abstracts International, 58(04), 1174. (UMI No. 9729635)

About the Authors

Editor

Magdalena Mo Ching Mok is chair professor of Assessment and Evaluation at Department of Psychological Studies and codirector of the Assessment Research Centre, The Hong Kong Institute of Education. Prior to joining the institute in 1999, Professor Mok was senior lecturer at Macquarie University, New South Wales, Australia, where she had served for over 10 years. She obtained her Ph.D. degree in education from the University of Hong Kong and M.Sc. Degree in statistics from the University of Glasgow. She completed a B.Sc. Degree in mathematics from the Chinese University of Hong Kong. Professor Mok is strongly committed to excellence in teaching and was recipient of Distinguished Teacher Award 2003–2004 of the institute. Her research focuses attention on the integration of assessment and self-directed learning to enhance instruction and learning.

Other Contributors

Carl Martin Allwood (Ph.D.) is professor of psychology at the University of Gothenburg, Sweden. His current research is mainly on judgment and decision making, including metacognition with respect to confidence judgments of semantic and episodic memory performance, for example, in forensic contexts. He also writes on issues in culture-oriented psychology, science studies, and theory of science.

William John Boone holds the rank of professor of Educational Psychology at Miami University (Oxford, OH, USA). From 1991 to 2005, he served as a faculty member at Indiana University-Bloomington. Dr. Boone currently serves as the director of a graduate certificate program which incorporates Rasch measurement coursework. Boone is also the director of a psychometric laboratory (test/survey design and analysis laboratory) which he founded. He has presented Rasch

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0, © Springer Science+Business Media Dordrecht 2013

workshops throughout the world. He can be contacted at boonewj@muohio.edu and at wjdboone@hotmail.com. Boone earned his Ph.D. from the University of Chicago's Measurement, Evaluation, and Statistical Analysis Program. Dr. Ben Wright directed Dr. Boone's thesis. Boone earned an M.S. in geophysics from the University of Wisconsin-Madison and a B.S. in geology from Indiana University-Bloomington.

Paisley Tsz Mei Cheung is currently education assessment services manager of Hong Kong Examinations and Assessment Authority. She has previously worked as assessment development officer at the Centre for Assessment Research and Development (CARD) of The Hong Kong Institute of Education. She obtained her bachelor of arts and master of arts (Chinese language and literature) from the Chinese University of Hong Kong. Ms. Cheung has extensive experience in teaching and has worked in schools, education bodies, and university in Hong Kong. She has been involved in funded education projects and publications in the area of Chinese language teaching and assessment for learning. Her research interests include self-directed learning oriented assessment (SLOA), assessment for learning/teaching, and school-based assessment.

Hye-Jeong Choi is an assistant professor at the University of Nebraska-Lincoln. Her research focuses on methodological issues in educational and psychological measurement, including item response theory, latent class analysis, and diagnostic classification models. In some cases, these models can be combined as finite mixture models to better understand individual differences in response patterns in testing settings. Dr. Choi has also worked on latent growth modeling with categorical data sets and Bayesian estimation as an algorithm for analyzing complex data using prior knowledge.

Michelle Davidson is director of Teaching and Learning (Assessment), Trinity Grammar School, Sydney, Australia. In her current position, Ms. Davidson is responsible for the design, implementation, and ongoing management of assessment systems and learning processes across the school. From 2006 to 2010, Ms. Davidson was director of Assessment of Pearson Research and Assessment (2006–2010). In this role, she was responsible for managing the key Pearson portfolios associated with item/test design, development, and implementation; large-scale assessment and testing; development of assessment support materials; preparation of standards; and data management, data literacy, reporting, and diagnostics. From 2002 to 2006, she was manager of Test Development (System and School Testing) and senior research fellow at the Australian Council for Educational Research (ACER). In this role, Ms. Davidson was responsible for the test development associated with the state-based testing programs of most state and territories across Australia. She has also worked on projects in the Middle East, New Zealand, and China.

Jimmy de la Torre obtained his Ph.D. in Quantitative Psychology from the University of Illinois at Urbana-Champaign. He is currently an associate professor at the Department of Educational Psychology at Rutgers University. His primary research interests include psychological and educational testing and measurement, with a specific emphasis on item response theory and latent-variable modeling, and designing diagnostic assessments that can be used to inform classroom instruction and learning.

Lorna Earl, Ph.D., is director of Aporia Consulting Ltd. She is a recently retired associate professor and head of the International Centre for Educational Change at OISE, University of Toronto. Assessment has been her career passion, and she has written many articles, books, and monographs on this topic.

Pauline Swee Choo Goh received her doctorate from the University of Adelaide and is currently a senior lecturer at the Sultan Idris Education University, Malaysia. Besides teacher education, her research and teaching interests also include the development of beginning teachers with reference to their competency, pedagogical knowledge, and actual use in classroom situations. Pauline is a lead researcher to a national grant concerning the use of standards for teacher competency and is coresearcher for two other national grants looking at improving teaching and learning in schools. She is currently involved in teaching in the undergraduate and postgraduate degree programs and supervises a range of masters and doctoral students in fields of teaching and learning. Pauline actively contributes to journals and books and is currently an editorial board member for the newly established *Journal of Research, Policy and Practice of Teachers and Teacher Education* of Sultan Idris Education University.

Chi Ming Ho was a former assessment development officer of Centre in the Assessment Research and Development (CARD), The Hong Kong Institute of Education. He obtained his master's degree of education with distinction from the University of Hong Kong. He has extensive experience in primary school teaching and professional training for teachers in Hong Kong. He has been invited as one of the associate editors of the book named *Reform, Inclusion, and Teacher Education: Towards a New Era of Special and Inclusive Education in Asia-Pacific Regions*. His research interest includes assessment for learning, pedagogy, and higher-order thinking.

Connie Chia-Ling Hsu is a Ph.D. candidate in psychology in the National Chung Cheng University in Taiwan and a senior research assistant in the Assessment Research Centre at The Hong Kong Institute of Education. She earned her doctoral candidate in 2009 and master's degree of psychology in 2006 both from the National Chung Cheng University. Her research interests include computerized adaptive testing, computerized classification testing, and item response theory. She worked as a research assistant during 2006–2009 and was responsible for projects related to quantitative psychology, psychometrics, and developing an online test of computerized adaptive testing.

Slava Kalyuga is associate professor at the School of Education, the University of New South Wales, where he received a Ph.D. and has worked since 1995. His research interests are in cognitive processes in learning, cognitive load theory, evidence-based instructional design principles, and cognitive diagnostic assessment.

His specific contributions include detailed experimental studies of the role of learner prior knowledge in learning (expertise reversal effect), the redundancy effect in multimedia learning, the development of rapid online diagnostic assessment methods, and studies of the effectiveness of different adaptive procedures for tailoring instruction to levels of learner expertise. He is the author of three books and more than 60 refereed research articles and chapters.

Steven Katz is senior lecturer in Human Development and Applied Psychology at OISE/UT, where he is also the coordinator of the Psychology of Learning and Development initial teacher education program component and a director of Aporia Consulting Ltd. He has received the governor general's medal for excellence in his field and is coauthor of *Leading Schools in a Data-Rich World*.

Sabina Kleitman (Ph.D.) is a senior lecturer at the School of Psychology, University of Sydney, Australia. Her research interests include metacognitive processes, their measurement, biases, predictors, and their role in educational settings (schools and tertiary education), as well as their impact on psychological well-being. Her other streams of research focus on decision-making and educational psychology.

Sze Ming Lam has a B.A. degree in translation from the Chinese University of Hong Kong, 2011. She is currently executive assistant at Assessment Research Centre, The Hong Kong Institute of Education.

Doris Ching Heung Lau is a Ph.D. student at the University of Hong Kong, and she was formerly an assessment development officer (mathematics) at the Centre for Assessment Research and Development (CARD), The Hong Kong Institute of Education. Her research interests relate to the learning and teaching of mathematics and assessment in the mathematics classroom.

Henry Kai On Lee obtained his doctoral degree from The Hong Kong Institute of Education. His research interests centered around educational measurement and assessment. His doctoral research focuses attention on the intervention effects of assessment for learning on preservice sports coaches' attitudes toward assessment for learning in sports. He is currently lecturer of sports studies at the School of Continuing and Professional Education, The Hong Kong Institute of Education. He has great interest in research on effective sports coaching methods, and he aspires to contributing to the future career of his students through his teaching and learning.

Anthony Wai Chi Leung is a certificate master (teacher) at a primary school in Hong Kong. He graduated from the Chinese University of Hong Kong, major in government and public administration. He obtained his postgraduate diploma in education from The Hong Kong Institute of Education. His research interest includes assessment for learning.

Karina Kar Lee Mak completed her B.Ed./B.A. (Psych) Hons at the University of Sydney, Australia, in 2009. Her honors research examined children's metacognitive regulation in the physical domain. In 2011, Karina will complete her masters of organizational psychology from Macquarie University, Australia, and become a registered psychologist with the Psychology Board of Australia. Her current research interests include proactivity in the workplace and specifically proactive career behaviors.

The late **Bobbie Matthews** was an adjunct senior lecturer in the School of Nursing and Midwifery at Flinders University, Australia. She investigated changes in values and approaches to learning over time in changing cultural milieus. She taught English as a second language, anatomy and physiology, as well as research skill development to nursing students, and critical thinking skills to overseas tertiary students. Her educational background in the sciences, multilingual ability, and her abiding interest in research in the Asia-Pacific region had enabled her to engage with a broad range of students.

Ming-Yan Ngan (Teacher Cert., B.Ed., M.Ed., Ed.D.) is assistant professor at Department of Curriculum and Instruction at The Hong Kong Institute of Education. His research focuses attention on the educational assessment and school culture for the enhancement of teaching and learning. His three latest books include *Theory and Application of the Contemporary Educational Assessment for Learning*; *Contemporary Educational Assessment: Practices, Principles and Policies*; and *Postwar Hong Kong Education*. Those books are currently main references by students in educational programs for understanding the Hong Kong education landscapes.

Min Pan, at the time of the writing of this chapter, was a graduate student in the HDQM department at the University of Maryland, which was formerly the Department of Measurement, Statistics, and Evaluation (EDMS).

Somwung Pitiyanuwat (B.Ed. (Hons), M.Ed., Ph.D.) is currently the chairman of the Executive Board for the National Institute of Educational Testing Service (NIETS), the chairman of Rajabhat Rajanagarindra University Council, and associate fellow of the Royal Institute of Thailand, Academy of Moral and Political Sciences. Previously, he was dean of the Faculty of Education, vice president for Research Affairs of Chulalongkorn University, and the first director of the Office of National Education Standards and Quality Assessment (ONESQA). He is an acclaimed expert in quality assurance, educational assessment, and teacher education and development in SE Asia.

Tan Pitiyanuwat completed a bachelor's degree of architecture in industrial design from King Mongkut's Institute of Technology Ladkrabang and a master of arts in industrial design from the Central Saint Martins College of Art and Design of the University of the Arts London, England. He is currently a deputy dean for Quality Assurance of the Kasem Bundit University Faculty of Architecture. Previously, he was a product designer at Volksmobel Co, Ltd. (furniture export company). His research interests include sustainability in furniture design, furniture design in mass production, quality assurance in higher education, and alternative assessment using teaching and learning portfolios.

André A. Rupp is an associate professor in the Department of Human Development and Quantitative Methods (HDQM) at the University of Maryland. His research interests center around the design, implementation, and data analysis for embedded assessments in digital learning environments for the development of diagnostic score reports; he collaborates with researchers at the University of Wisconsin at Madison and Cisco in these areas. Dr. Rupp's primary statistical focus is on cognitive diagnosis/diagnostic classification models as well as related multidimensional latent-variable models. In 2010, he published the first comprehensive didactic volume on diagnostic classification models with two colleagues in the field, which is available for purchase through Guilford Press in the USA.

Lazar Stankov is visiting professor at the National Institute of Education, Singapore. He taught individual differences and assessment courses at the University of Sydney for 30 years and worked at the Educational Testing Service at Princeton, NJ, for 5 years. His current research interests include the study of cognitive abilities, confidence/metacognition, cross-cultural differences, and the development of measures of militant extremist mindset.

John R. Staver holds the rank of professor of science education and chemistry and is codirector of the Center for Research and Engagement in Science and Mathematics Education (CRESME) at Purdue University. His current research focuses on constructivist epistemology, its implications for improving science teaching and learning, and the interface between science and religion, emphasizing each discipline's nature and the conflicts between them. He is a fellow in the American Association for the Advancement of Science (AAAS) and the American Educational Research Association (AERA). Staver earned an Ed.D. in science education from Indiana University, an M.S. in chemistry from Purdue University, and a B.S. in education (chemistry) from Indiana University.

Hak Ping Tam is an associate professor in the National Taiwan Normal University, where he has served for 3 years as the chairperson of the Graduate Institute of Science Education. He graduated from the University of California at Berkley with a bachelor's degree in mathematics. Later, he obtained both his master's and Ph.D. degrees from the Ohio State University, focusing on research methodology, evaluation, and applied statistics. Upon graduation, he taught for 4 years as an assistant professor in the Department of Measurement, Statistics and Evaluation at the University of Maryland at College Park. His research interests include mathematics education, statistics education, research methodology, and assessment in mathematics and science. He is now working on projects related to assessment for learning, assessment report system, large-scale assessment, and automated Chinese essay scoring system.

Stephen Yin Chuen Ting has interests in web software, software engineering, and iPhone application development. He received his B.A. degree in computing in 2008 from Hong Kong Polytechnic University and M.S. degree in information technology management in 2011 from Hong Kong Baptist University. During his work as research assistant at Assessment Research Centre, The Hong Kong Institute of Education, he assumed a key developer role on the development and implementation of the SP Xpress 2.2 program.

Jim Tognolini is a senior research fellow at the Oxford University Centre for Educational Assessment, senior vice president research and assessment for Pearson Global Strategies and Business Development, professorial fellow at Wollongong University (Australia), and adjunct professor of education at the University of Western Australia. His research focuses primarily on assessment, measurement, and psychometrics, specifically Rasch measurement. He has given numerous keynote addresses and seminars in countries all over the world, and most recently, his research activity has been focused upon developing models and procedures to support the international monitoring of educational outcomes through system-wide, school-level programs.

David Tzuriel, Ph.D., is a clinical and educational psychologist and an expert on dynamic assessment (DA) of learning potential. Currently he works as a full professor at Bar-Ilan University (Israel) and serves as the editor of the *Journal of Cognitive Education and Psychology (JCEP).* In the past he was the president (1999–2001) of the *International Association for Cognitive Education and Psychology (IACEP) and* chaired the School of Education (2003–2007).

Wen-Chung Wang is the director of Assessment Research Center and chair professor of The Hong Kong Institute of Education (HKIEd). Prior to joining HKIEd in August 2008, he was a distinguished professor and head, department of psychology at National Chung Cheng University in Taiwan. He obtained a Ph.D. degree from University of California at Berkeley in 1994. His research interests include Rasch measurement, item response theory, and psychometrics. He received Research Excellence Awards from National Science Council and National Chung Cheng University and the Mu-Dou Award for his contribution to education.

Michael Ying Wah Wong is a senior research assistant in Assessment Research Centre at The Hong Kong Institute of Education. He is undertaking his doctoral studies at The Hong Kong Institute of Education with emphases in educational measurement and assessment.

Margaret Wu has a background in statistics and educational measurement. Her main interests include item response modeling, mathematics education, and large-scale assessments. She has been closely involved in international comparative studies such as the OECD PISA project and IEA TIMSS study. Margaret is the first author of an item response modeling software, ConQuest, which fits a family of item response models.

Jacob Kun Xu is currently Ph.D. student at Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong. His area of interest is multidimensional Rasch modeling.

Melissa Seward Yale is a doctoral candidate in Research Methods, Measurement, and Evaluation at Purdue University. She earned an M.S. in applied statistics from Purdue University and a B.A. in sociology from Cedarville University. Her research interests include shared decision analysis in health care contexts, advanced statistical techniques, and quantitative measurement theory.

Zi Yan completed his Ph.D. degree at James Cook University in the field of educational measurement, and at present he is working as an assistant professor at The Hong Kong Institute of Education. His research interests include educational assessment, application of Rasch model in education and psychology, and quantitative research methods.

Jing Jing Yao is a lecturer in the Department of Psychology of Zhejiang Normal University and also a Ph.D. student at The Hong Kong Institute of Education. Her current research focuses on cognitive diagnostic assessment in education.

Sarah Young completed her B.Ed./B.A. (Psych) Hons at the University of Sydney in 2009, Australia. She is currently undertaking a Doctorate of Clinical Psychology/ Masters of Science, at the University of Sydney. Her research interests include metacognitive processes and environmental influences for self-confidence in school-aged children and eating disorders, specifically the role of exercise in recovery from anorexia nervosa.

George Yu has lived, studied, and worked in China, the USA, and Hong Kong. He has a B.A. in English language arts from Zhejiang University and a M.Ed. from Boston University in applied linguistics. Throughout his career, he has been a language teacher, program administrator, researcher, and education consultant working with many public and private schools in places like Beijing, Hangzhou, Boston, Hong Kong, Macau, and Guangzhou, interfacing with teachers, administrators, students, and their parents. Bilingual education and language development are among his main academic interests.

Yue Zhao earned her Ed.D. of psychometrics in 2008 and M.S. of statistics in 2007 both from University of Massachusetts. Her research interests include psychometrics, research methods, assessment of learning, and applied statistics. Dr. Zhao is currently working as project manager affiliated to the Centre for the Enhancement of Teaching and Learning at the University of Hong Kong. Prior to joining academic settings, she had worked as a psychometrician for Educational Testing Service from 2008 to 2010. She has extensive practical and research experience in working with various types of assessments, including admission tests, K-12 assessments, language assessments, licensure and credentialing exams, personality assessments, and classroom exams.

Author Index

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*, Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18, DOI 10.1007/978-94-007-4507-0, © Springer Science+Business Media Dordrecht 2013

Subject Index

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*, Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0, © Springer Science+Business Media Dordrecht 2013