Methodology of Educational Measurement and Assessment

Bernard P. Veldkamp
Cor Sluijter *Editors*

Theoretical and Practical Advances in Computer-based Educational Measurement

Springer Open

Methodology of Educational Measurement and Assessment

Series Editors

Bernard Veldkamp, Research Center for Examinations and Certification (RCEC), University of Twente, Enschede, The Netherlands
Matthias von Davier, National Board of Medical Examiners (NBME), Philadelphia, PA, USA

This book series collates key contributions to a fast-developing field of education research. It is an international forum for theoretical and empirical studies exploring new and existing methods of collecting, analyzing, and reporting data from educational measurements and assessments. Covering a high-profile topic from multiple viewpoints, it aims to foster a broader understanding of fresh developments as innovative software tools and new concepts such as competency models and skills diagnosis continue to gain traction in educational institutions around the world. Methodology of Educational Measurement and Assessment offers readers reliable critical evaluations, reviews and comparisons of existing methodologies alongside authoritative analysis and commentary on new and emerging approaches. It will showcase empirical research on applications, examine issues such as reliability, validity, and comparability, and help keep readers up to speed on developments in statistical modeling approaches. The fully peer-reviewed publications in the series cover measurement and assessment at all levels of education and feature work by academics and education professionals from around the world. Providing an authoritative central clearing-house for research in a core sector in education, the series forms a major contribution to the international literature.

More information about this series at http://www.springer.com/series/13206

Bernard P. Veldkamp · Cor Sluijter
Editors

Theoretical and Practical Advances in Computer-based Educational Measurement

Springer Open

Editors
Bernard P. Veldkamp
University of Twente
Enschede, Overijssel, The Netherlands

Cor Sluijter
Cito
Arnhem, Gelderland, The Netherlands

ISSN 2367-170X ISSN 2367-1718 (electronic)
Methodology of Educational Measurement and Assessment
ISBN 978-3-030-18479-7 ISBN 978-3-030-18480-3 (eBook)
https://doi.org/10.1007/978-3-030-18480-3

© The Editor(s) (if applicable) and The Author(s) 2019. This book is an open access publication.
Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.
The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Professor T. H. J. M. Eggen–Theo for short—has dedicated almost all of his scientific career to applying psychometrical methods to solve practical problems in educational testing and to improve test quality. Abroad, he is probably best known for his work in item response theory (IRT) in general and computerized adaptive testing in particular. Academics in the Netherlands, though, know that his scope has been much broader than that. Generically speaking, the main theme in his work is increasing the quality of educational measurement.

Having both worked long with him as close colleagues, we thought it would be fitting to publish a book in the year of his retirement that as a whole reflects the different topics that he has been working on throughout his career as an educational measurement specialist. This book therefore focuses on several themes in educational measurement and both addresses quite practical and more theoretical issues. It consists of five different parts: the quality of educational tests; psychometrics, large-scale educational assessment, computerized adaptive testing, and innovations in educational measurement. All contributors to its 20 chapters have been working with Theo, either as close or remote colleagues, as (former) Ph.D. students, or both, or all.

However, Theo would be quite cross with us, if this book would only serve as a token of our appreciation for his work. We therefore tried to bring together a set of chapters that reflect in one way or another on a practical innovation, on a solution to a problem in educational measurement, or give more insight into a topic in educational measurement that might be of interest to you. For this reason, we hope that this book will be a source of help or inspiration, for researchers, psychometricians, or practitioners.

Enschede, The Netherlands — Bernard P. Veldkamp
Arnhem, The Netherlands — Cor Sluijter

Introduction

In a time of digital innovations, educational assessment, and especially computer-based educational assessment, evolves rapidly. New methodology and new technology become available, and questions arise about how to benefit from these developments in large-scale testing and about the consequences for the quality of measurement.

In this book, a number of technical, theoretical, and other advances and innovations in different fields of educational measurement are described. They have in common that they all in one way or another contribute to solving practical problems in educational measurement. They spring from the area of quality assurance, psychometrics, large-scale assessment, computerized adaptive testing or are related to the technology or materials being used in educational assessment.

The first part of the book is on improving the quality of educational measurement. In its first chapter, by Saskia Wools, Mark Molenaar, and Dorien Hopster-den Otter, the impact of technological advancements in educational assessment is discussed in the context of the argument-based approach to validity. In the second chapter, inaccessibility to an assessment—an important source of construct irrelevance—is discussed by Erik Roelofs. A framework to ensure the accessibility of items is presented as well as an application of said framework. Next, Arnold Brouwer, Bernard Veldkamp, and Marieke Vroom introduce the system-oriented talent management (STM) model which makes it possible for companies to achieve a sustainable match between organizations and their (potential) employees. This chapter also explores the possible application of the STM in the world of educational measurement. In the fourth and final chapter of this first part, Cor Sluijter, Bas Hemker, and Piet Sanders present the RCEC review system that is specifically tailored for evaluating the quality of educational tests and exams. To illustrate how the system works in practice, it is applied to review the quality of a computer-based entrance test for mathematics for Dutch teacher colleges.

The second part of this book contains five psychometrically oriented chapters. In the first one, Maarten Marsman, Charlotte Tanis, Timo Bechger, and Lourens Waldorp present the Curie–Weiss model from statistical physics as a suited paradigm for the analysis of educational data. They study the statistical properties of this model and discuss its estimation with complete and incomplete data. A simulated example is presented, and the analysis of fit of the model is illustrated using real data. The second chapter by Anton Béguin and Hendrik Straat provides techniques to determine the number of items and corresponding cut scores that are necessary to decide on mastery in formative educational measurement. These techniques are applied in situations with different item characteristics and the outcomes for varying test situations are provided, illustrated by a practical example. In the third chapter, Norman Verhelst explores two models for continuous responses that allow for the separation of item and person parameters. One can be seen as a Rasch model for continuous responses, and the other is a slight generalization of a model proposed by Müller. For both models, estimation procedures are proposed and a comparison of the information function between models for continuous and for discrete observations is discussed. The fourth chapter is a contribution by Matthieu Brinkhuis and Gunter Maris, in which trackers are presented as instruments with specific properties to deal with changes in model parameters, like ability and item difficulty, when measurements extend over longer periods of time. These trackers retain the strengths of both state space models and rating systems, while resolving some of their weaknesses. They have properties that make them especially suitable for tracking individual progress in education or any aggregate thereof, as in reporting or survey research. The fifth chapter is by Monika Vaheoja. In this chapter, the process of resetting performance standards, with small samples, in different versions of an exam is discussed. Performance setting methods using circle-arc equating or concurrent calibration with OPLM as an IRT model are being compared. In this comparison, attention is paid to sample size, test length, test difficulty, and respondents' abilities.

The third part of the book has large-scale assessment as its subject. The first chapter by Qiwei He, Dandan Liao, and Hong Jiao presents an exploratory study in which cluster analysis is applied to investigate the relationship between behavioral patterns and proficiency estimates, as well as employment-based background variables. For a problem-solving item in the PIAAC survey, the communality of behavioral patterns and backgrounds is addressed. Besides, the focus is on how problem-solving proficiency is influenced by respondents' behavioral patterns. The second chapter by Cees Glas is about methods for relating standards on the number-correct scale to standards on the latent IRT scale. The size of standard errors when equating older versions of a test to the current version is estimated. The local reliability of number-correct scores and the extra error variance introduced through number-correct scoring, rather than using IRT proficiency estimates, are estimated as well. In the third chapter, Remco Feskens, Jean-Paul Fox, and Robert Zwitser examine some effects of the alterations in the test administration procedure that took place in the 2015 PISA survey. The switch from paper-based assessments to computer-based assessments was studied by evaluating if the items used to assess trends across the PISA surveys 2012 versus 2015 are subjected to differential item functioning. The impact on the trend results due to the change in assessment mode of the Netherlands is also assessed. The results show that the decrease reported for mathematics in the Netherlands is smaller when results are based upon a separate national calibration. In the fourth and final chapter of this part, Hyo Jeong Shin, Matthias von Davier, and Kentaro Yamamoto investigate rater effects in international large-scale assessments. They illustrate the methods and present the findings about rater effects on the constructed-response items in the context of the fully computer-based international student skill survey PISA 2015.

The fourth part focuses on computerized adaptive testing. In the first chapter, Maaike van Groen, Theo Eggen, and Bernard Veldkamp discuss two methods in multidimensional computerized classification testing in the case of both between-item multidimensionality and within-item multidimensionality. The two methods are Wald's sequential probability ratio test and Kingsbury and Weiss' confidence interval method. Three different procedures are presented for selecting the items: random item selection, maximization at the current ability estimate, and the weighting method. Two examples illustrate the use of the classification and item selection methods. Chapter two, by Bernard Veldkamp and Angela Verschoor, is about robust test assembly as an alternative method that accounts for uncertainty in the item parameters in CAT assembly. Not taking the uncertainty into account might be a serious threat to both the validity and viability of computerized adaptive testing, due to capitalization on chance in item selection. In this chapter, various methods for dealing with uncertainty are described and compared. In the third chapter, Angela Verschoor, Stéphanie Berger, Urs Moser, and Frans Kleintjes focus on reducing the costs of CAT production. On-the-fly item calibration strategies are compared for various settings of operational CAT. Elo chess ratings, joint maximum likelihood, and marginal maximum likelihood-based strategies are considered. A combination of various strategies is proposed to be applied in practice. The fourth chapter is by Darkhan Nurakhmetov. He applies reinforcement learning for item selection in computerized adaptive classification testing. The item selection method is introduced and advantages of the new method for content balancing and item exposure control are being discussed.

The fifth part of this book is on technological developments. In the first chapter, Fabienne van der Kleij, Joy Cumming, and Lenore Adie investigate the feasibility and value of using easily accessible equipment for practitioner use and research purposes. They study the use of a GoPro camera and an Apple iPad in capturing one-to-one teacher–student feedback interactions and subsequent individual video-stimulated recall (VSR) for self-reflection. Findings suggest that such technology has potential for use by teachers and students to improve reflection and feedback interaction and thus to enhance student learning. In the second chapter, Jos Keuning, Sanneke Schouwstra, Femke Scheltinga, and Marleen van der Lubbe explore the possibilities of using a board game to collect high-quality data about children's spoken interaction skills in special education. The quality of the conversations between the children was evaluated with a specially designed observation form. Video recordings showed that almost all children were willing to participate in the game, even the children who usually barely speak in class. Moreover, the game provided more than sufficient data to assess different dimensions of spoken interaction skills. In the third chapter, Sebastiaan de Klerk, Sanette van Noord, and Christiaan van Ommering discuss two models that are used in educational data forensics: the Guttman error model and the log-normal response time model. Next, they report the results of an empirical study on the functioning of the Guttman and response time model. In addition to this, the design, development, and validation of a protocol on the use of educational data forensics are presented.

Bernard P. Veldkamp
Cor Sluijter

Contents

Part I Improving Test Quality

Part II Psychometrics

Part III Large Scale Assessments

Part IV Computerized Adaptive Testing in Educational Measurement

Part V Technological Developments in Educational Measurement

Part I
Improving Test Quality

Chapter 1
The Validity of Technology Enhanced Assessments—Threats and Opportunities

Saskia Wools, Mark Molenaar and Dorien Hopster-den Otter

Abstract Increasing technological possibilities encourage test developers to modernize and improve computer-based assessments. However, from a validity perspective, these innovations might both strengthen and weaken the validity of test scores. In this theoretical chapter, the impact of technological advancements is discussed in the context of the argument-based approach to validity. It is concluded that the scoring and generalization inference are of major concern when using these innovative techniques. Also, the use of innovative assessment tasks, such as simulations, multi-media enhanced tasks or hybrid assessment tasks is quite double-edged from a validity point of view: it strengthens the extrapolation inference, but weakens the scoring, generalization and decision inference.

1.1 Introduction

Increasing technological possibilities encourage test developers to improve computer-based assessment in multiple ways. One example is the use of authentic items that have the potential to improve construct representation, making it possible to assess complex constructs like skills or competences (Sireci and Zenisky 2006). Furthermore, complex scoring methods make it possible to include both students' responses and decision making processes (e.g. Hao et al. 2016). In addition, recent new insights in adaptive algorithms could help to develop personalized learning and assessment systems. Thus meeting the increased need for personalization in both learning and assessment. All these innovations are promising in a sense that they can improve the quality of assessments significantly. Caution is required, however,

S. Wools (✉)
Cito, Arnhem, The Netherlands
e-mail: saskia.wools@cito.nl

M. Molenaar
Open Assessment Technologies, Luxemburg, Luxemburg

D. Hopster-den Otter
Universiteit Twente, Enschede, The Netherlands

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_1

since these innovations can also negatively impact important values of testing such as comparability and transparency.

Innovations in computer-based assessment can be described in a context of validity. Validity is one of the most important criteria for the evaluation of assessments (AERA, APA and NCME 2014) and is often defined as the extent to which test scores are suitable for their intended interpretation and use (Kane 2006). This chapter aims to address general aspects of computer-based assessment that can guide future validation efforts of individual computer-based assessment for a particular purpose and interpretation.

Validation efforts can be structured according to the argument-based approach to validation (Kane 2006, 2009, 2013), which is a general approach that can be used as a framework to structure validity evidence. The argument-based approach to validation aims to guide validation efforts by proposing a procedure that consists of two stages: a developmental stage and an appraisal stage. In the developmental stage, an interpretation and use argument (IUA) is constructed by specifying all inferences and assumptions underlying the interpretation and use of a test score. In the appraisal stage, a critical evaluation of these inferences and assumptions is given within a validity argument.

The IUA is structured according to several, predefined inferences (Wools et al. 2010): scoring, generalization, extrapolation, and decision. Every inference holds its own assumptions and underlying claims to argue valid use of test scores. When computer-based assessments are used, several additional claims are made—and subsequently, must be validated. At the same time, innovations in computer-based assessments can provide us with additional data or evidence that can support the validity of these assessments.

This chapter aims to describe and discuss several innovations in computer-based assessment from a validity perspective. The central question is: what are the threats to, and opportunities for, innovative computer-based assessments in the context of validity? In the first section, we describe the trends and innovations regarding computer-based assessments. These can be categorized into three categories: innovations in items or tasks; innovations in test construction, assembly and delivery; and innovations that accommodate students personal needs and preferences. The second section introduces the concept of validity and validation more thoroughly and describes the inferences underlying validity arguments. The two sections come together in the third, where the impact of technological innovation is discussed to establish the effect on the inferences from the argument-based approach to validation. In this section, we argue that these technological advancements are both improving as well as threatening the validity of assessments. And we propose some research questions that should be posed during the validation of innovative computer-based assessment.

1.2 Innovations in Technology-Enhanced Assessments

The use of technology in education is increasing significantly, access to the internet is ubiquitous, schools adopt new digital tools and students bring their own devices to the classroom. These technological advancements are not only limited to learning materials, also assessment can benefit. For example, when audio and video are used to create a rich and authentic assessment context that is appealing to modern-day students (Schoech 2001). Or, when process data and response times are gathered to improve insights in the behaviour on individual items (Molenaar 2015). These techniques can be used to further improve computer-based assessment of skills and competences. New technology can also help measure skills that were hard to measure by traditional CBA's. For example, previously, speaking ability could be measured through recording of speech, but scoring was done manually. Nowadays, cloud computing allows for AI-based automated scoring that was not possible before (Zupanc and Bosnic 2015). Technology can also be used to measure "new" competences like 21st century skills (Mayrath et al. 2012). As an example, assessing "collaborative problem solving" requires new types of items that include inter-agent interaction (OECD 2017). Digital technology makes it possible to create these types of items that go beyond the limits of what can be tested on paper with traditional multiple choice and constructed response interactions.

New (types of) devices and peripherals are being introduced at a rapid pace. The first Apple iPhone was introduced in 2007 and revolutionized mobile, personal computing and touch-based input. In 2009, Fitbit introduced the concept of wearable computing or "wearables", which has since evolved and branched out into head mounted displays (Google Glass 2013, Oculus Rift 2016) and smart watches (Google Watch 2014, Apple iWatch 2015). The iPad popularized the tablet in 2010 and received instant appeal from educators based on its friendly form factor and ease of use. In 2012, Microsoft's 2-in-1 Surface bridged the gap between tablets and laptops, appealing to audiences in higher education. And most recently smart speakers like Amazon Alexa (2014) and Google Home (2016) truly introduced us to the age of the assistant.

These devices have introduced new form factors, new types of input (interactions) and new output (sensor data). Natural touch/gesture based input has made technology more usable, allowing even infants to use it. Mobile phones have made audio (speech) and video recording accessible to all. And geographic, accelerometer and gyroscope data allow for more natural interactions with devices, localization and improved ease-of-use.

At the same time, the ubiquity of the internet allows for access to tools and information anywhere and at any time. Cloud computing has propelled machine learning. Providing us with endless possibilities, like on-the—fly video analysis to detect suspicious behaviour in airports or which groceries are put in shopping baskets (Johnston 2018). Voice assistants can send data to cloud-based algorithms to process natural language and follow-up on the requests of the user, including making reservations at a restaurant by a bot indistinguishable from a real-life person (Velazco

2018). Even in the area of creativity, AI has demonstrated being capable of creating artworks and composing music (Kaleagasi 2017).

This chapter discusses several practical implementations in the area of educational assessment today. Although there are many more technological innovations that impact assessment practices, examples are chosen within three distinct categories or levels:

1. Items and tasks: innovations in individual item and task design, ranging from simulations to multi-media enhanced items to hybrid tasks, (partially) performed in the real world
2. Test construction, assembly and delivery: innovations in automated item generation, adaptive testing and test delivery conditions by use of (online) proctoring
3. Personal needs and preferences: innovations to adapt to the personal needs and preferences of the individual student, ranging from accessibility tools to Bring Your Own Device (BYOD) to personalized feedback and recommendations in the context of learning.

1.2.1 Innovations in Items and Tasks

In educational measurement, technological innovations seem promising to support the assessment of "new" competences such as collaborative problem solving, creativity or critical thinking. These constructs are typically assessed through complex and authentic tasks. When these tasks are developed leveraging the possibilities of new technologies, new types of items emerge. These items have become increasingly more complex and are often referred to as Technology Enhanced Items (TEIs) (Measured Progress/ETS Collaborative 2012, p. 1):

> Technology-enhanced items (TEI) are computer-delivered items that include specialized interactions for collecting response data. These include interactions and responses beyond traditional selected-response or constructed-response.

By using these item types, it is possible to include sound and video and animations within the assessment. At the same time, performances are measured more direct and authentic. Finally, these items provide us with the opportunity to gather data beyond a correct or incorrect response. This additional data includes for example log-files, time stamps and chat histories. In general, challenges of TEIs are typically that they might favor digital natives, are harder to make accessible, are more expensive to develop in comparison with traditional items and that the resulting data are harder to analyze than with classical approaches.

As an example, we distinguish three types of TEI. The first TEI is a *simulation.* This is a digital situation where a student can roam a virtual environment and complete relevant tasks (Levy 2013). These simulations could be simple apps rendered within the context of a classic test to full immersive environments enriched by use of head-mounted VR headsets. Simulations are developed to simulate a specific situation that invites students to respond to in a particular way: often these simulations are used to simulate an authentic situation.

The second type of TEI is one that can be used for video and audio recording (OAT 2018), using the student's device to record speech or capture video: *multi-media enhanced items*. These items are used to gather data that goes beyond constructed responses or multiple choice items. The speech and video fragments that are collected collected could be routed to either manual scorers or automated scoring algorithms (Shermis and Hammer 2012).

The last TEI is referred to as a *hybrid tasks*. These tasks are (in part) performed in the real world and can be directly (automatically) evaluated, allowing for greater interactivity. An example of a hybrid task is solving a table-top Tangram puzzle, which is recorded by a tablet-webcam and instant feedback is provided. (e.g., Osmo Play 2017).

1.2.2 Innovations in Test Construction, Assembly and Delivery

Regarding test construction activities, such as assembly and delivery, digital technology allows for automated item generation, adaptive testing and online (remote) proctoring for enhanced test security.

Automated item generation is the process of generating items based on predefined item models (Gierl 2013). These models can be very complex, taking into account the required knowledge and skills, but also personal preferences of students to make items more appealing, e.g. by posing questions in the context of topics personally relating to students like football or animals. This process can take place a priori to generate thousands of items to bootstrap an itembank, but also in real-time to adapt to personal preferences, apply digital watermarking for detecting origins of itembank-leaks (Foster 2017) and/or take into account real-time test-delivery data (responses, scores, process data).

Computer Adaptive Testing (CAT) has been around for a long time and is described as the process where test construction and test administration are computerized and individualized (Eggen 2007). There are many different types of CAT, each with their own objectives and technological implementations, but all with an adaptive engine that is used for real-time adaptive test assembly. From a technology standpoint, CAT can benefit from advancements in cloud computing, allowing real-time (re)calibration and even more complex computations and constraints to be evaluated. CAT-engines can also be designed to take into account prior data, demographics and personal needs and preferences. Prior data could be anything from previously selected items to the result of a previous test to an ability estimate by a teacher, in order to select a better first set of items to be administered, as that can increase efficiency significantly (van der Linden 1999). Demographics and personal needs and preferences are specific to the individual and provide instructions to the adaptive algorithm to balance content (e.g. exclude a certain topic) and take into account accessibility needs (e.g. exclude items with images unsuitable for people suffering from color-blindness) or even device preferences (e.g. do not select items with drag & drop as this user is using a mouse).

On the test-delivery level, *online proctoring* allows for secure testing on any location, providing greater flexibility on where and when tests can be delivered. An adaptive test could be delivered in the comfort of a student's own home, while online proctors would proctor the test remotely by webcam surveillance. Leveraging cloud computing, real life proctors could be assisted by AI to process the video data detecting aberrant behavior (e.g. sudden movements or voices in the background). And real-time data forensics engines could be used to spot anomalies during the actual test taking, e.g. answering items correctly at a very high speed or suspicious answers patterns indicating possible collusion with other students.

1.2.3 Innovations Regarding Personal Needs and Preferences

Lastly, digital technology allows assessments to be adapted to personal needs and preferences. Personal needs and preferences are typically related to (legal) accessibility requirements, but can also be personal preferences of any kind, e.g. a preferred type of device (tablet, laptop) or screen size. The latter is closely related to the phenomenon of Bring Your Own Device (BYOD), where students bring their own devices into the classroom and want to perform their tasks using the device and configuration they are familiar with. Also, when personal needs and preferences are saved, it becomes possible to present students with personalized feedback.

Tools for Accessibility are products, devices, services, or environments for people with disabilities and are becoming increasingly important in the area of assessment, to ensure all students have equal opportunities when taking a test. The foundations and the extent of accommodations may vary but in many countries it is simply required by law (e.g., American Disability Act). Also the types of accommodations can vary, ranging from always-on accessibility features, to extended features based on personal profiles and to the use of specialized Assistive Technologies like screen readers or refreshable braille devices.

Another type of personalization is the use of the preferred *device* type or form factor trough *BYOD* (*bring your own device*). Typical web applications employ the principle of responsive design: an approach to web design that makes web pages render well on a variety of devices and window or screen sizes. Based on device capability and available screen estate, content is reformatted dynamically to provide the best possible user experience. Apart from screen estate, also available input types can play an important role, e.g. a mobile phone with an on-screen keyboard, a tablet with a type cover or a laptop with a physical keyboard can yield different results (Laughlin Davis et al. 2015). Some students may be very proficient with a certain input type, whereas others might struggle. To accommodate for this, it is important for students to either be able to use the (type of) device of their preference or are allowed sufficient time to practice and get acquainted with the compulsory/recommended device and mode.

Personalization transcends the process of assessment delivery; the type and mode of feedback can also be individualized, taking into account personal preferences, prior data and learning styles. This *personalized feedback* constitutes to personalized or adaptive learning, where an AI-based recommendation engine can take into account all factors and data to provide the appropriate type of the feedback and content, tailored to the exact needs of the individual: e.g. the next chapter to read, video to watch or exercise to complete.

1.3 Validity and Validation

Extensive research caused the concept of validity to change significantly over time (Lissitz 2009). Kane (2006) summarized this by citing three general principles of validation that emerged from the widely accepted model of construct validity (Cronbach and Meehl 1955). The first principle concerns the need to specify the proposed interpretation of test scores. The second principle refers to the need for conceptual and empirical evaluation of the proposed interpretation. The third principle states the need to challenge proposed and competing interpretations. All these principles are reflected in widely known theories on validity and approaches to validation. For example, in Messick's (1989, p. 13) definition of validity:

> …an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment [italics in original].

Messick's conceptualization of validity has resulted in a validation practice that aimed to present as much validity evidence as possible. From this practice, the validity of test scores has been supported by combining countless sources of validity evidence that are either content-related, criterion-related, or construct-related. To propose a more pragmatic practice, Kane suggested the argument-based approach to validation (2004, 2006). This approach guides validation efforts through selecting the most relevant sources of evidence and therefore lessens the burden on practitioners. According to Kane (2013, pp. 8–9):

> The argument-based approach was intended to avoid the need for a fully developed, formal theory required by the strong program of construct validity, and at the same time to avoid the open-endedness and ambiguity of the weak form of construct validity in which any data on any relationship involving the attribute being assessed can be considered grist for the mill (Bachman 2005; Cronbach 1988; Haertel 1999; Kane 1992).

The argument-based approach to validation consists of two arguments: an interpretation and use argument (IUA) and a validity argument. The IUA states which inferences and assumptions underlie the intended interpretation and use of test scores. Whereas the validity argument evaluates the evidence that is presented to support or reject the inferences from the IUA and draws a conclusion on the adequacy of the validated instrument for the intended interpretation and use.

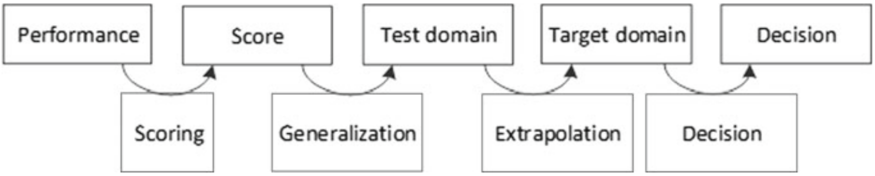


Fig. 1.1 Inferences within an IUA (Kane 2006)

When looked at in more detail, the IUA helps us to specify our reasoning from an observed performance in an assessment situation towards a decision and the use of this decision for a particular purpose (e.g., selection, classification, a didactical intervention), this is represented in Fig. 1.1. The first inference (scoring) describes how a students' performance on tasks is translated into an observed score. These scores are usually interpreted as a generalizable instance of a test domain score (generalization). A test domain represents all possible tasks that could be presented to students within the chosen operationalization of the construct. The test domain scores are subsequently extrapolated (extrapolation) to scores on a broader domain. This domain can be a theoretical competence domain, which entails an operationalization of the competence or construct that is being measured. It can also entail the practice domain, that is, a real-life situation that students can encounter in their future (professional) lives. Whether the test domain scores are extrapolated into a theoretical competence domain or a practice domain depends on the particular testing situation, in general one could say that we extrapolate to a target domain. Either way, building on this extrapolation, the final inference (decision) can lead to a decision on the students' level on the competence of interest.

After developing the IUA, analytical and empirical evidence are gathered to enable an evaluation of the claims stated in the IUA. The analytical evidence could entail, for example, conceptual analyses and judgments of the content of the test domain and competence domain. Most of the analytical evidence could already have been generated during development of the assessment. The empirical evidence consists of data relating to, for example, the reliability of an assessment, the structure of the construct, or relations with other measures of the construct of interest. This kind of evidence is gathered in so-called validation studies, which are designed to answer specific research questions derived from the need for specific empirical evidence. The evidence is used for the validity argument, that includes a critical evaluation of the claims in the IUA. Note that this consists of both appraising currently defined inferences and assumptions and rejecting competing interpretations.

One might think that the argument-based approach encourages to gather all possible analytical and empirical evidence. However, according to Kane (2009, p. 49):

> …some statements in the literature can be interpreted as saying that adequate validation requires that every possible kind of validity evidence be developed for validation to be complete.…

This shotgun approach is clearly unwieldy, and in its extreme form, it makes validation impossible.

Therefore, within the argument-based approach, it is argued that inferences that seem weak or that are of great interest to the intended interpretation and use of tests require more evidence than others. Although evidence is needed for every inference, the weight placed on different inferences depends on the assessment that is being validated.

The argument-based approach to validation is applied to several assessments and, when necessary, adapted to fit different perspectives or uses of tests. The most prominent shift in the approach was proposed by Kane (2013) when he argued that the use of assessment results should play a more prominent role in the approach. This resulted in a change in terminology: the theory moved from using an interpretive argument into using an interpretation and use argument. Others, also published specific applications of the argument-based approach or proposed extensions to parts of the theory: for language assessments (Llossa 2007), for assessment programs (Wools et al. 2016), for classroom assessment (Kane and Wools 2019) and for formative assessment (Hopster-den Otter et al., submitted).

The current chapter applies the argument-based approach to innovative computer-based assessments. Coming back to the subject of this chapter, when developing or validating innovative computer-based assessments, one might need to rethink which inferences seem weak or are of great interest. Also, when validation practice is not moving along with digital innovations, we might target our validation efforts at the wrong inferences. Therefore, in the remainder of the chapter we will give an overview of the impact of innovations in computer-based assessments and where they might impact our claims and assumptions underlying the inferences in the IUA.

1.4 Validity of Innovative Technology-Enhanced Assessments

The argument-based approach to validation starts with specifying inferences, assumptions and claims that are made in the assessment process. Since innovations in computer-based assessments have impact on all aspects of the assessment process, the impact on validity is large. In this section we will discuss the inferences distinguished in an IUA. From there, we discuss specific claims, assumptions and threats underlying the inferences when technological enhanced innovations are used in assessment. This provides an overview of possibilities and threats within a validity argument that should play a central role when gathering and evaluating validity evidence.

1.4.1 Inferences Within the IUA

As previously mentioned, an IUA consists of a set of inferences and accompanying claims and assumptions. It depends on the particular assessment and the intended interpretation and use of the assessment scores what claims and assumptions are relevant within an IUA. We exemplify the inferences in general with claims and inferences that are commonly used (Wools et al. 2016).

Scoring inference

When students perform an assessment task, such as answering items or solving a complex problem, data are collected to transform the students' behavior into an interpretable unit. Usually this is a score that indicates whether the answer was correct, or if possible, partially correct. This inference implies that it is possible to make statements about a task being performed correctly or not. Another assumption is that the score is a true translation of students' ability to perform on the task. The final assumption underlying the scoring inference is that students are able to show their skills or competences without barriers. In practice, this means that students know what is expected of them, that the tools work intuitively, and that they are able to perform the task without technological difficulties.

Generalization inference

Generalizing a score from an assessment to a test domain means that the responses on that particular assessment can be interpreted as representative for all possible tasks or test forms that could have been presented. It also means that the performance must be more or less the same when a student takes the test twice with different items. This implies that the tasks in one assessment must be representative for the full test domain and that this is comparable for different versions or instances of the assessment.

Extrapolation inference

When we extrapolate a score on the test domain to a target domain we assume that the tasks within the test domain are derived from this target domain. This inference relies very heavily on one claim: the task is as authentic as possible. When assessment tasks are very authentic, the extrapolation of what we observed is not far from what we would like to make decisions about.

Decision inference

The main question for this inference is: are we able to make a decision about students that is in concurrence with the intended interpretation and use of the assessment? This implies that we have meaningful cut scores or norms that can be applied to students performances. Furthermore, it is implied that the results are meaningful to students and that they can be used for the intended purpose.

1.4.1.1 Innovations in Items and Tasks

When TEI's are used, it is possible that defining the correct response becomes more complex. Different processes, responses or answers could all be considered effective behavior and therefore assumed to be 'correct'. Furthermore, translating behavior into a single score does not always reflect the effort that has been put into the tasks. Therefore, to quantify behavior on these new types of tasks, data that describe the followed process (log-files) are often collected and used for analysis and reporting. This means the use of complex algorithms to score behavior or the use of, for example, automated essay scoring to evaluate the quality of an answer. The risk with these algorithms is that scoring becomes less transparent and hard to verify, especially when machine learning is used and so called black-boxes are created. This threatens the *scoring inference* in a way that it becomes harder to evaluate whether a score is given correct.

The *generalization inference* assumes that tasks are selected to cover all relevant aspects of a construct. A risk for innovative technology enhanced assessments is construct underrepresentation. Construct underrepresentation occurs when only a small aspect of a construct is assessed. For example, we assess the ability to converse about the weather in another language while the intend was to make a decision about someone's full ability of conversing in another language. In technology enhanced assessment, TEIs are often used. However, developing these tasks is a time consuming and costly effort that leads to a limited set of contexts or tasks. Moreover, time constraints during the administration of the test, or other practical limitations in the administration, often prevent us from presenting a large number of tasks, contexts and skills. When limited items are presented, this will threaten the generalizability of the obtained scores to the full test domain.

At the same time, these TEI's provide us with more opportunities to build rich and authentic tasks. Simulations include open digital environments where a student can virtually roam and complete relevant tasks. Items that include multi-media provide the possibility to grasp performance on video or record speech. And finally, hybrid tasks invite students to perform offline (for example solve a puzzle) and provides them with online feedback. This last examples makes sure that the computer is not 'in between' the student and his or her performance anymore. All in all, all these tasks are developed to provide the candidate with an authentic experience. This way the behavior that is called for in the assessment situation is as identical as possible as the behavior that requested in the competence domain. Therefore, through these authentic items the *extrapolation inference* is strengthened.

The *decision inference* includes assumptions about cut scores and norms. Cut scores and norms are usually the result of statistical analysis, equating techniques or standard setting procedures. Unfortunately, the commonly used methods are not always suitable for the rich data that are produced through TEI's. This means that even when these tasks are scored in a transparent, reliable and comparable way, it might still be a problem to decide 'what behavior do we consider good enough to pass the assessment?'.

1.4.1.2 Innovations in Test Construction, Assembly and Delivery

Within the *scoring inference*, it is assumed that a score assigned to a student's performance is a translation of a student's ability. More specifically, that the score is only influenced by the performance of a student on the task at hand and not, for example, by other students who could help. Therefore, this assumption does not hold when students discuss their answer with others. Fortunately, cheating becomes harder to do with new security measures like (online) proctoring. Furthermore, adaptive algorithms and live test assembly allow for individual test forms, making copying of answers more difficult.

The *generalization inference* is concerned with reliability and comparability between test forms in terms of content representation. In terms of comparability between test versions, adaptive engines can be used to make sure different test versions are comparable in terms of content. These engines use sophisticated rules to sample items within certain content restrictions. To be able to do this, the item pool must be large enough. If this would be the case, then content comparability can be ensured over different versions and therefore, these engines strengthen our generalization inference.

As mentioned previously, authenticity is an important aspect of the *extrapolation inference*. One of the threats to authenticity is the availability of tasks that speak to a candidates personal interests. For example, an authentic situation for a student to read texts, is often to read a text that holds information that a student is interested in. Advances in automated item generation support test developers in constructing items that can speak to different personal interests of students. Therefore, it is possible that AIG can positively influence authenticity and therefore extrapolation.

The decision inference is concerned with cut scores, norms and score reports. The innovations mentioned regarding test construction, assembly and delivery are not related to these aspects and this inference.

1.4.1.3 Innovations Regarding Personal Needs and Preferences

An assumption underlying the *scoring inference* is that students are able to answer items or perform tasks without barriers. For example, when students need to work with formulas, they should not be limited to demonstrate their skills because of the complexity of the formula-editor that is used in an assessment context. This can also occur when a device that is used in the testing condition is different from the one a student is used to. As an example, someone who is used to an Android phone usually has problems in using an iPhone and the other way around. In an assessment situation these unnecessary difficulties cause for construct irrelevant variance since the score does not reflect the true ability of the student anymore, but is impacted by the ability to cope with other technical devices. One of the solutions in current assessments is a Bring Your Own Device (BYOD) policy where students can use devices and tools that they worked with during the learning phase. For students with special needs, this means that they can also use their own tools for accessibility, such as screen reader software or a refreshable braille device. We acknowledge that this strengthens the claim that underlies the scoring inference, but at the same time, it raises questions about comparability and might weaken other inferences.

The generalization inference is an example of an inference that might be weakened through a BYOD policy or allowing tools for accessibility. This is, when students bring their own devices and use their personal software and peripherals to ensure accessibility, the claim regarding comparability is challenged. Especially when items or tasks are not suited for these different modes (touch screen devices vs. mouse-controlled devices) or when item presentation varies over different devices. This causes items to differ in terms of necessary cognitive load and therefore in their difficulty for students.

To strengthen the *extrapolation inference*, we would like the assessment situation to be as authentic as possible. For example, when we want to say something about someone's ability to sing a song at home—the most authentic way is to let them sing a song. However, there are still aspects that would prevent us from being able to make that claim. What if the person is really nervous when there is an audience or when it is necessary to use a microphone? Therefore, even if a task is authentic, there is still an inference to be made about the possibility to extrapolate the observed behavior into potential behavior on the target domain. As mentioned before, it becomes more common for students to be able to use their own device and tools that they are used to work with during class. This bridges some of the extrapolation gaps between the assessment context and learning situation and therefore could positively impact the extrapolation inference.

When an assessment is over, a decision about students is made. Within the decision inference it is assumed that is possible to give meaning to the test results. This is done through norms, cut scores and score reports. These score reports usually consist of a visualization of the assessment score interpretation. However, it can also include feedback that is aimed to guide students' further learning. An advantage of technological advancement is the possibility to provide students with personalized feedback. The feedback that should be presented is not only selected based on item answers and ability estimates, but can also be selected based on analysis of learning patterns and learning preferences. When this is done in a formative assessment context or a context of classroom assessment, assessment results are translated into meaningful actions right away, strengthening the claim that the results can be used for the intended purpose. This is not only the case for formative assessment, also for summative assessment or classification purposes, the possibility to combine different data sources to decide on the next best step strengthens the inference.

1.4.2 Validity Argument of Technology-Enhanced Assessments

A validity argument consists of an integral evaluation of all sources of evidence and a critical appraisal of the claims and inferences. This is necessary because a design choice or a particular source of evidence can support an inference and at the same time threaten another. For example, the use of simulation-based assessment might be

Table 1.1 Opportunities (+) and threats (−) for validity

	Scoring	Generalization	Extrapolation	Decision
Items and tasks				
Simulations	−	−	+	−
Multi-media enhanced tasks	−	−	+	−
Hybrid tasks	−	−	+	−
Test construction, assembly and delivery				
Automated item generation			+	
Adaptive engines	+	+		
(online) Proctoring	+			
Personal needs and preferences				
Tools for accessibility	+	−	+	
Bring your own device	+	−	+	
Personalized feedback				+

evaluated positive in light of the extrapolation inference, but gives reason for concern for the scoring and generalization inference. Table 1.1 shows this in more detail. The technological innovations discussed in Sect. 1.2 of this chapter are listed as well as the four inferences. For every inference it is noted whether an innovation has the potential to be an opportunity (+) or a threat (−). Some technological innovations were not discussed in relation to the inference or are not applicable and are left empty.

The validity argument aims to present a balanced case to come to a conclusion about the overall validity of the test scores. When the results from Table 1.1 are taken into account, it stands out that the evidence most needed for innovative assessment is to strengthen the scoring, generalization and decision inference. Questions that should be addressed are, for example, can innovative items be scored in a transparent way that includes all relevant aspects of the task? Is the sample of behavior representative enough to justify claims that go beyond the assessment situation? Is comparability between test versions and test conditions plausible? And is it possible to make decisions about students performances that are both explainable and meaningful?

1.5 Concluding Remarks

In this chapter, we discussed the impact of several technological innovations from a validity perspective. Validity and validation are defined from a perspective of the argument-based approach and according to this approach several inferences are distinguished. For every inference, claims were specified and related to a limited set of technological trends from educational assessment practice. Some of these practices are strengthening the validity claims, others are weakening them.

In general, one could say that the scoring and generalization inference are of major concern when using these innovative techniques. Also, the use of innovative assessment tasks, such as simulations, multi-media enhanced tasks or hybrid assessment tasks is quite ambiguous from a validity point of view: it strengthens the extrapolation inference, but weakens the scoring, generalization and decision inference. It is important to note that this could be solved relatively easy by providing evidence that rejects the assumptions of incomparability between tasks or shows how these innovative tasks can be scored. An advantage of the technological advancement in this context, is that the new data that these tasks provide us with, and the new data analysis techniques that are available, can be of help in creating this evidence and to study these claims more extensively.

Furthermore, we stress that this is a general reflection of an IUA. For every assessment it is necessary to build a custom IUA specifying the inferences, claims and assumptions relevant to that assessment and its intended interpretation and use of the test scores. Moreover, every assessment probably holds its' own unique combination of innovative features that might combine differently than the ones presented here. Building an interpretive argument for a specific computer-based assessment is helpful in deciding what the weakest inferences are and where to target validation efforts on.

One thing that stands out, is that many innovations are practice-based. Test developers, assessment organizations, start-ups and even technological companies develop techniques to improve assessment. However, little are evaluated and reported on in a systematic way, let alone published in scientific journals. This is an important aspect of validity, the use of evidence-based techniques that are tested and proven lessens the burden to gather additional data and strengthens our claims. Also, when innovations or new techniques from other fields are used in assessment, it is necessary to study and publish on the impact of these advancements.

We acknowledge that there are many more advancements that were not discussed. Some of these will strengthen and some will weaken validity. Also, some of these innovations are very expensive to build or use, others are easy to implement. We conclude that a lot of technological advancements hold potential to improve assessments considerably, we should, however, not forget that a lot of assessment-challenges can be addressed by traditional types of assessments and items as well. Only when it is necessary, these innovations might be able to truly add value. And when it seems that these innovations bring new problems that seem unsolvable, keep in mind: traditional assessments have problems of their own, those are simply the problems we are familiar with by now.

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1–34.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*.

Foster, A. (2017, June 20). *What national security can teach us about protecting our exams*. Retrieved from https://www.caveon.com/2017/06/21/what-national-security-can-teach-us-about-protecting-our-exams/.

Gierl, M. (2013). *Advances in automatic item generation with demonstration* [PowerPoint slides]. Retrieved from https://www.taotesting.com/wp-content/uploads/2014/09/TAO-Days-AIG-October-2013.pdf.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Hao, J., Smith., L., Mislevy, R., Von Davier, A., & Bauer, M. (2016). *Taming log files from game/simulation-based assessments: Data models and data analysis tools*. ETS Research Report Series, 1–17. https://doi.org/10.1002/ets2.12096.

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2018). *A general framework for the validation of embedded formative assessments*. Manuscript submitted for publication.

Johnston, C. (2018, January 22). *Amazon opens a supermarket with no checkouts.* Retrieved from https://www.bbc.com/news/business-42769096.

Kaleagasi, B. (2017, March 9). *A new AI can write music as well as a human composer: The future of art hangs in the balance.* Retrieved from: https://futurism.com/a-new-ai-can-write-music-as-well-as-a-human-composer.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurment* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39–64). Charlotte, NC: Information Age Pub.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000.

Kane, M. T., & Wools, S. (2019). Perspectives on the validity of classroom assessments. In S. Brookhart & J. McMillan (Eds.), *Classroom assessment and Educational measurement.* Abingdon, Oxon: Routledge.

Laughlin Davis, L., Kon, X., & McBride, Y. (2015). *Device comparability of tablets and computers for assessment purposes.* Paper presented at National Council on Measurement in Education, Chicago.

Levy, R. (2013). Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. *Educational Assessment, 18,* 182–207. https://doi.org/10.1080/10627197.2013.814517.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing INC.

Llosa, L. (2007). Validating a standards-based assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24,* 489–515. https://doi.org/10.1177/0265532207080770.

Mayrath, M. C., Clarke-Midura, J., Robinson, D. H., & Schraw, G. (Eds.). (2012). *Technology-based assessment of 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age Publishing.

Measured Progress/ETS Collaborative. (2012). *Smarter balanced assessment consortium: Technology enhanced items*. Retrieved from https://www.measuredprogress.org/wp-content/uploads/2015/08/SBAC-Technology-Enhanced-Items-Guidelines.pdf.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Molenaar, D. (2015). The value of response times in item response modeling. *Measurement: Interdisciplinary Research and Perspectives, 13*(3–4), 177–181. https://doi.org/10.1080/15366367.2015.1105073.

OAT. (2018). *Userguide TAO—Portable custom interactions*. Retrieved from: https://userguide.taotesting.com/3.2/interactions/portable-custom-interactions.html.

OECD. (2017). What is collaborative problem solving? In *PISA 2015 results* (Vol. V)*: Collaborative problem solving*. Paris: OECD Publishing. https://doi.org/10.1787/9789264285521-7-en.

Osmo Play. (2017). Retrieved from: https://www.playosmo.com/en/tangram/.

Schoech, D. (2001). Using video clips as test questions: The development and use of a multimedia exam. *Journal of Technology in Human Services, 18*(3–4), 117–131. https://doi.org/10.1300/J017v18n03_08.

Shermis, M., & Hammer, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In *Annual National Council on Measurement in Education Meeting* (pp. 1–54).

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Smarter Balanced Assessment Consortium. (2012). *Technology-enhanced items guidelines*. Developed by Measured Progress/ETS Collaborative. Retrieved from: https://www.measuredprogress.org/wp-content/uploads/2015/08/SBAC-Technology-Enhanced-Items-Guidelines.pdf.

Van Der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23*(1), 21–29. https://doi.org/10.1177/01466219922031149.

Velazco, C. (2018, June 6). *Google's reservation-making AI will be making calls soon.* Retrieved from: https://www.engadget.com/2018/06/27/google-duplex-assistant-public-testing/?guccounter=1.

Wools, S., Eggen, T. J. H. M., & Béguin, A. A. (2016). Constructing validity arguments for test combinations. *Studies in Educational Evaluation, 48,* 10–16. https://doi.org/10.1016/j.studeduc.2015.11.001.

Wools, S., Eggen, T. J. H. M., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 8,* 63–82.

Zupanc, K., & Bosnic, Z. (2015). Advances in the field of automated essay evaluation. *Informatica* (Slovenia), 39.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 2
A Framework for Improving the Accessibility of Assessment Tasks

Erik Roelofs

Abstract In constructing tests it is vital that sources of construct irrelevant variance are minimized, in order to enable valid interpretations about the test taker. One important source of construct irrelevance is inaccessibility to the test and its items. In this chapter a framework is presented for design and review of test items, or more broadly, assessments tasks, to ensure their accessibility. An application is presented in the context of theory exams to obtain a drivers' license in the Netherlands.

2.1 Accessibility of Assessments

Access in the context of educational testing refers to the opportunity for a student to demonstrate proficiency on a target skill (e.g., reading, mathematics, science). Accessibility is not seen as a static test property but instead, as the result of an interaction among test features and person characteristics that either permit or inhibit student responses to the targeted measurement content (Kettler et al. 2009; Ketterlin-Geller 2008; Winter et al. 2006).

In more general terms, accessibility is considered as a prerequisite to validity, the degree to which a test score interpretation is justifiable for a particular purpose and supported by evidence and theory (AERA, APA and NCME 2014; Kane 2004).

Limited accessibility can have various causes, that may threaten the validity score interpretation in different ways, depending on the assessment purpose.

A first source of limited accessibility pertains to the situation where test takers do not yet master the target skills. If this is the case, a conclusion might be that the test taker needs to go through an extended or adapted learning process. There would not be a threat of validity. However, if the test in general was judged as too difficult for the intended target group, this could be a case of misalignment between test content and intended outcomes, which can be considered as a threat to validity.

E. Roelofs (✉)
Cito, Arnhem, The Netherlands
e-mail: Erik.Roelofs@cito.nl

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_2

Second, a lack of 'hard' access capabilities is a well-documented source of access problems (Sireci et al. 2005). Due to some sort of disorder or handicap a student is not able to process task information or to respond to the task. For instance, test takers with a visual or auditory impairment, a motor disorder, ADD, autistic spectrum disorder, or dyslexia, may lack necessary capabilities to have access to a test item. For these test takers, the validity of inferences is compromised when they cannot access the items and tasks administered to them and with which they are required to interact. To improve accessibility for students with special needs, an example of reducing barriers and increasing accessibility, is the provision of a read-aloud assessment administration.

A third source of limited accessibility is lack of access skills that can be developed through education, but of which it is questionable whether they belong to the target skill or competency (Abedi and Lord 2001; Kettler et al. 2011). For instance, assessment tasks for math may place a high burden on reading skill and may subsequently cause access problems. The question is whether students in this case do not have the necessary level of target skill or whether they lack the access skill of reading comprehension. Some of the accessibility problems can be prevented by defining the target skill related to the assessment purpose well in advance, and determine whether access support is necessary, without impacting the measurement of the target skill.

A fourth source of limited accessibility is related to flaws in task presentation itself, that may either result from constructors' mistakes or by misconceptions about a task presentation feature. By their very design assessment tasks themselves can be inaccessible to some and even all students. Design flaws relate to errors, inconsistencies, omissions in the assignment or in the task information, or in response options and cause extra processing load for students (Beddow et al. 2008).

In the remainder of this chapter we concentrate on the design principles for improving assessment accessibility, that help to avoid unnecessary, construct irrelevant task load, and thereby improve the validity of claims about test takers' target skills.

2.2 Principles that Underlie Accessible Assessment Design

2.2.1 Principles from Universal Design

To address the challenge of designing and delivering assessments that are accessible to and accurate for a wide range of students, principles of universal design (UD; Mace 1997) have been applied to the design, construction, and delivery of tests. The core tenet of UD is to create flexible solutions that avoid post hoc adaptation by considering from the start the diverse ways in which individuals will interact with the assessment process. Dolan and Hall (2001, 2007) therefore proposed that tests be designed to minimize potential sources of construct-irrelevant variance by supporting the ways that diverse students interact with the assessment process. Thompson et al. (2002) adapted Mace's original elements from architectural design to derive seven elements of accessible and fair tests: (1) inclusive assessment population; (2) precisely defined constructs; (3) accessible, nonbiased items; (4) items amenable to accommodations; (5) simple, clear, and intuitive instructions and procedures; (6) maximum readability and comprehensibility; and (7) maximum legibility. Ketterlin-Geller (2005, p. 5) provides a more generic definition of universal design for testing: an "integrated system with a broad spectrum of possible supports" that permits inclusive, fair, and accurate testing of diverse students.

2.2.2 *Principles from Cognitive Load Theory*

In cognitive load theory it is assumed that any instructional task, including an assessment task, has a certain amount of intrinsic task load. This is the natural complexity that the task possesses, due to the complexity of the task content (Sweller 2010). In Fig. 2.1 this is depicted by the green part of the left stacked bar, representing the total task load. The intrinsic task load is strongly related to the number of task elements that need to be processed simultaneously by a test taker. However, test items may also address learnable access skills (yellow colored area) which are not part of the target skill per se, such as reading a text or understanding a picture. In addition, the test item may address hard capabilities (red shaded area), such as eyesight, quality of hearing and short term memory. Depending on the target group, these types of task load may be seen as inevitable, or as avoidable or even irrelevant. The part of the load that can be avoided is hypothetically depicted through dotted areas of the left bar. A final source of extrinsic task load includes load that is caused by errors or flaws in parts of the item itself (blue dotted area). This source of extrinsic task load should always be avoided.

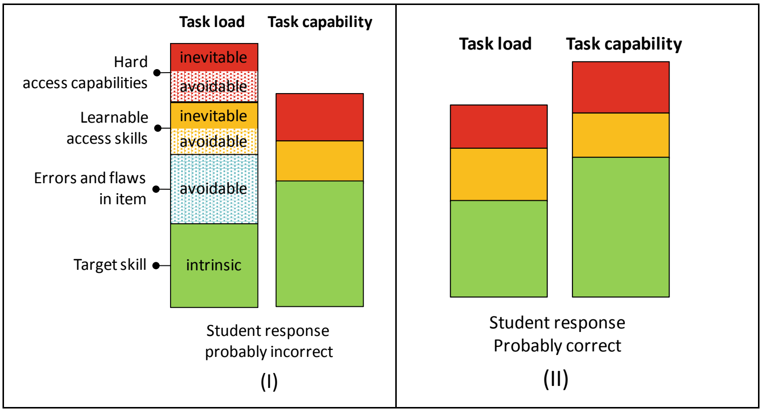


Fig. 2.1 Intrinsic and extrinsic task load versus test taker task capability in an assessment task

A test taker solving an assessment task is confronted with the total task load of an item and uses his or her own available combination of hard access capabilities, access skills, and target skills, his or her 'task capability'. Part of the task capability level is the ability to deal with avoidable extrinsic task load caused by the assessment task presentation over and above the level of the intrinsic task load, which is handled by applying the target skill. However, in the hypothetical situation in Fig. 2.1 (part I) the test taker will probably not be able to succeed because his task capability level falls short of the total level of task load.

When, however, all avoidable sources of extrinsic task load have been stripped off, and the amount of intrinsic task load remains constant, the test taker's level of task capability in this hypothetical situation is high enough (see part II of Fig. 2.1), to solve the task correctly.

Extrinsic task load in assessment tasks (test items, assignments) is for a large part caused by the way a task is presented. As discussed, test takers are forced to use skills that are not necessarily related to the target skill, but which do require cognitive resources. According to cognitive load theory (CLT) different cognitive mechanisms can be involved in causing extrinsic load.

- Quantitative overload: more information is presented than is strictly necessary for the assessment task.
- Deception: information sets off thinking directions that prevent test takers from arriving at a (correct) solution.
- Missing information: the test taker cannot get a complete problem representation, due to missing information.
- Split attention: test takers have to keep elements of information from different locations in their short term memory, to be able to respond to the task.
- Redundancy effects: the task presents redundant information that causes an extra memory load. E.g. redundant verbal explanation of an already self-evident picture.
- Interference-effects: verbal and visual information contradict or differ in nature, causing additional processing load for the test taker.

In the next section we present a framework for item design and item screening in which we intend to separate sources of extrinsic task load from intrinsic task load. Moreover, guidelines are provided to reduce extrinsic task load in different parts of an assessment task, taking the perspective of the test taker who solves a task.

2.3 Evaluating and Improving Accessibility of Assessment Tasks from a Test Takers' Perspective

In order to prevent unnecessary extrinsic task load it is important to evaluate assessment tasks (items, assignments) on possible sources of inaccessibility as described above, and modify these tasks accordingly. For that purpose two excellent tools have already been developed, the Test Accessibility and Modification Inventory (TAMI; Beddow et al. 2008) and the TAMI Accessibility Rating Matrix (ARM; Beddow et al.

2013). Beddow et al. (2011) report about how the use of the tools in the context of biology items improved the accessibility, while at the same time the psychometric quality of the items was improved.

The TAMI-tool used to evaluate and improve item accessibility is predominantly structured into of five categories, that represent the typical format of items: (1) item stimulus, (2) item stem, (3) visuals, (4) answer choices, (5) page/item layout. In addition, TAMI has a sixth dimension, referred to as Fairness, which refers to a broad collection of issues, such as respect for different groups, the risk of construct-irrelevant knowledge, emotion or controversy arousing content, the use of stereotypes, and overgeneralizations.

We extended the TAMI-model to use it in the context of item development and item piloting in two ways (Roelofs 2016). First, the structure of the screening device is built around the cognitive and self-regulative processes that test takers engage in, when solving an assessment task. By doing so, the item writer takes the perspective of the test taker. This idea stems from the cognitive labs methodology, which is used during item piloting. The methodology employs procedures, such as thinking aloud protocols by test takers, that provide insight into respondents' thought processes as they respond to test items. In cognitive laboratory interviews evidence is gathered to ensure that target audiences comprehend task materials-survey scenarios, items/questions, and options as they were designed. lf the evidence suggests that task materials are not being understood as survey and test developers designed them, then modifications can be made in the wording of questions and options to ensure that the correct interpretation is optimized (Leighton 2017; Winter et al. 2006; Zucker et al. 2004).

Second, in addressing the accessibility of parts of an item, we deliberately separated guidelines in order to create intrinsic task load and to prevent extrinsic task load. This is done, because the question what content belongs to the target skill, intrinsic task load, and what belongs to the access skill is often a matter of definition. In solving mathematical word problems, for instance, reading can be seen as part of the target skill, as an inevitable necessary access skill, or as a skill that may hinder task solution, if not mastered at a proper level (Abedi and Lord 2001; Cummins et al. 1988).

An accessible assessment task supports the test taker to perform at his best, without giving away the answers on the items. Our heuristic model (see Fig. 2.2) depicts the test takers' response process to assessment tasks in a series of five connected task processes. The model is similar to the Winter et al. cognitive lab model (2006), but was adapted to account for the self-regulation activities, including monitoring, adjustment and motivation activities that take place during solving tasks (Butler and Winne 1995; Winne 2001). The processes in our model of task solution entail: (1) orientation on the task purpose and requirements, (2) comprehending task information, (3) devising a solution, (4) articulate a solution. During these task processes, that need not be carried out strictly sequentially, a self-regulation process including planning, monitoring and adjustment takes place (Winne 2001). The model of test takers' actions and thoughts is heuristic and general in nature and can be applied to different subject domains.

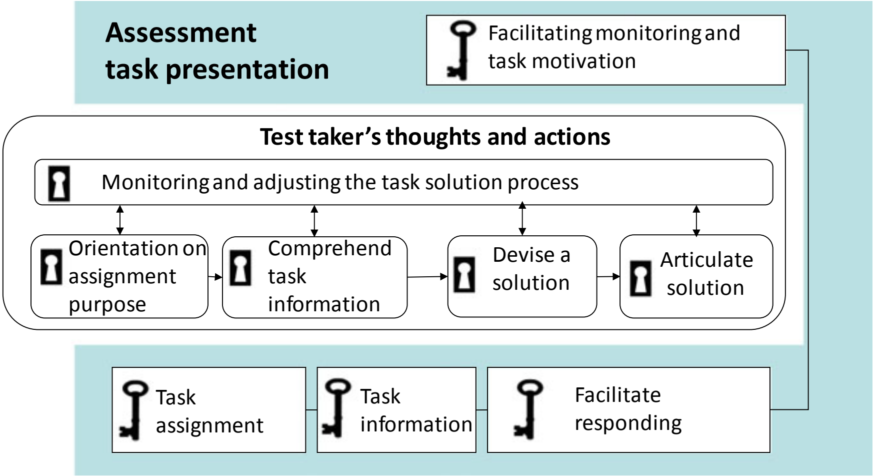

Assessment task presentation
Facilitating monitoring and task motivation
Test taker's thoughts and actions
Monitoring and adjusting the task solution process
Orientation on assignment purpose
Comprehend task information
Devise a solution
Articulate solution
Task assignment
Task information
Facilitate responding

Fig. 2.2 How assessment task presentation acts on test takers' thoughts and action the during task solution process

From an accessibility perspective it is the item constructors' task to support access to all sub processes of task solution by means of an optimal assessment task presentation. This involves the second ordering structure in our model and relates to similar categories as used in the TAMI-model, although reworded to cover any type of assessment task: assignment, task information, response facilitation, and lay-out/navigation.

Using this model and building on the recommendations by Beddow et al. (2008) we developed a checklist of requirements by which the most common problems of task presentation that act on test takers' response processes can be addressed (Roelofs 2016). The content of the recommendations and the checklist are summarized in Tables 2.1, 2.2, 2.3 and 2.4.

2.3.1 Supporting Orientation by a Clear Assignment

During orientation on an assessment task the test taker tries to find out the required outcomes of the task and expectations about the way these outcomes can be achieved. A clear assignment will support this orientation, including a clear statement about the product, the approach, the tools to be used, the criteria of quality and accuracy to be taken into account.

In order to create appropriate intrinsic task load in the presentation of the assignment it is essential that test takers are stimulated to show intended target behavior, including the steps to be taken, strategies to be used, tools to be used, outcomes to be accomplished, according to a specified level of accuracy. Reducing extrinsic load

Table 2.1 Guidelines for presenting the assignment

A1 Assignment: Provide information about the expected product, activities to be carried out, the purpose of the task *Create intrinsic task load* • Make sure the assignment is logically aligned with the task information • Make sure the assignment fits with the aimed target behavior, needed to demonstrate the skill *Reduce extrinsic task load* • Use specific words. Limit the number types of information • Length: keep assignment as concise as possible • Align the thinking direction of the task information with the assignment
A2 Guidelines for task execution: provide guidelines for a requested line of action, thinking steps to be shown, refer to the use of response tools and resources (see R4) *Create intrinsic task load* • Make sure that guidelines stimulate test takers to show intended specific target behavior, such as steps, tasks, strategy use, outcome, product • Clarify which tools and resources can or must be used (e.g. scrap paper) *Reduce extrinsic task load* • Give clear directions about the approach to be used. Make sure the requested responding behavior comes naturally, by facilitating response behavior instead of long explanations
A3 Task outcome requirements: clarify requirements for presentation, quantitative and qualitative outcome requirements, accuracy of solutions *Intrinsic task load* • Align requirements with intended task authenticity and performance conditions • Align accuracy of solutions or the result (number of words, accuracy margins) with what is intended *Reduce extrinsic task load* • Provide clear and concise directions about performance requirements

in orientation is accomplished by optimizing the presentation of the assignment. It is important to align the thinking direction of the assignment with the task information, the use of clear, concrete and concise language in giving directions. In general, we content that it is better to have self-evident response methods than to use (long) explanations of how to respond.

2.3.2 Supporting Information Processing and Devising Solutions

During the process of comprehending task information (second process) and the process of considering and devising solutions (or answers, third process), the test taker actually processes information that is relevant for solving the task, including some starting situation. Decisions are made about which information is essential for task solution, and strategies are chosen and applied to arrive at a solution, either in separate steps or by directly responding. Task information to be presented can relate to multiple objects, persons, symbols, actions that the test taker needs to act on. These can be presented within different contexts, that differ in authenticity, and conveyed

Table 2.2 Guidelines for presenting task information

I1 Task information: Present information from/about objects, persons, situations, that the test taker needs to act on. Align complexity level with the target group level

Create intrinsic task load

- Task information enables the onset of the target skill
- All necessary task information is given to carry out a task
- The amount of task information is balanced with the size of the task
- Level of complexity of task information is aligned with the developmental level of the target group

Reduce extrinsic task load

- Prevent mental overload, misguidance or deception, including:
  - Too high text complexity for the task purpose or the target group
  - Redundant textual information
  - Misleading textual information
  - Unnecessary complex visual information
  - Irrelevant visual information
  - Conflicting information

I2 Contexts: use contexts that are essential to the target skill and the target group

Create intrinsic task load

- Presented actors, objects, materials apply to the intended task situations
- Task contexts and their levels of authenticity align with what is intended
- Use context only when it is functional for demonstrating the target skill
- Choose the necessary level of context authenticity: real, simulated, no context

Reduce extrinsic task load

- Prevent the use of distracting, non-functional context-information
- Prevent confusing, stereotyping, exclusive, implausible contexts

I3 Media/stimuli: choose media that convey necessary task information faithfully

Create intrinsic task load

Make sure the media and stimuli represent the targeted task situation needed to demonstrate a target skill. E.g. the intended reading text types, math contexts, their presentation (plain text, web page, video clips, drawings, photos, animations)

Reduce extrinsic task load

- Provide maximum perceptibility of images, sound, speech
- Provide readability of characters, printed text, font
- Usability of media: adapt controls until these work intuitively
- Availability of tools to perceive stimuli (e.g. microscope, headphones)

Enable rewinding audio-visual stimuli, unless target skills require only one attempt

I4 Problem definition and support aligned with the target skill and the target group

Create intrinsic task load

- Define and structure the problem in line with the intended target skill, the target group, and desired performance conditions
- Determine how much regulative support is given to a test taker, based on intended level of independent task performance. Determine about supporting information, scaffolding of solution steps, or hints

Reduce extrinsic task load

- Present support and hints concisely and prevent spread-out of information

Table 2.3 Guidelines for facilitating responding

R1 Target material that the test taker has to act on *Create intrinsic task load* • Choose the material to be acted on: information, objects, materials, devices, persons, to the intended target behavior, task situations, and the level of authenticity *Reduce extrinsic task load* • Delete unnecessary target material, which is not to be acted upon in a task
R2 Interactivity of responses *Create intrinsic task load* • Determine whether the onset of the target skill requires an interactive response with visible consequences for: the actor, recipients, objects, present in the task context. Examples: marking of sentences, moving an object, an action with a visible result *Reduce extrinsic task load* • Make sure no unnecessary interactions can be carried out during responding
R3 Support in responding *Create intrinsic task load* • Make sure the level of support in responding corresponds with the intended performance conditions, as regards the target skill and the target group *Reduce extrinsic task load* • Prevent that support is redundant or oppositely too concise or even distractive
R4 Strategical and technical tools *Create intrinsic task load* • Add all necessary strategic and technical tools that enable test takers to express the target skill, in line with requirements about: the features of task situations; authenticity of the task context; specified performance conditions *Reduce extrinsic task load* • Make sure the tools work as expected and do not require highly developed handling skills
R5 Response mode and sensory channel *Create intrinsic task load* • Use a response format that enables the best expression of the target task behavior • Use the most appropriate response mode that expresses relevant task behavior (verbal, visual, psychomotor) • MC-options contain necessary information, are plausible, congruent with stem and task information, logically ordered, distinguishable; contain one correct option *Reduce extrinsic task load* • Replace indirect response modes by direct modes • Prevent the necessity to make mode switches or direction changes in responding: from verbal to visual; non-corresponding orders • MC-items: construct distractors that do not cause confusion or mental overload
R6 Response tools *Create intrinsic task load* • Choose response tools that enable a natural expression of a target skill. The required behavior in responding to tasks is supported by tools like drawing tools, formula-editors, word processor, chat apps, game controls *Reduce extrinsic task load* • Make sure the test taker can express the answer fluently and intuitively • Prevent any technical obstacles or difficulties in operability of response tools

Table 2.4 Guidelines regarding item lay-out and test navigation

SR1 Perceptibility-readability of screens, pages, item parts *Create intrinsic task load* • Presentation is aligned with the intended situation in which the test taker is supposed to perform • Consider: to what extent does increasing perceptibility of task information change the necessary level of intrinsic task load for measuring the target skill *Reduce extrinsic task load* • Make sure that screens, pages, item parts have a recognizable lay-out and content • Make sure the test taker does not get stuck in the task, because of the chosen level of perceptibility, readability of item parts, causing mental overload or confusion
SR2 Lay-out of item parts *Create intrinsic task load* • Make sure that presentation of the task elements represent task features of the target situation for the intended target group • Make sure that improving the accessibility of information by structuring its presentation does not change the necessary intrinsic task load that is part of the target skill (e.g. in reading comprehension) *Reduce extrinsic task load* • Make sure that the test taker does not get stuck in the task, because of the chosen lay-out and structuring of task elements, causing mental overload (e.g. spread of information, confusing tables, schemes)
SR3 Navigation through the test *Create intrinsic task load* • Make sure that the type of navigation support provided to the test taker represents the target situation for the intended target group *Reduce extrinsic task load* • Make sure the test taker does not get stuck in the task, because of a lack of support in monitoring own progress in item solving and navigation

through different modes of presentation, that may address different senses resulting in different processing.

In order to create appropriate intrinsic task load (see Table 2.2), it is essential that information is necessary for the onset of the target skill, including solution processes at a level that fits to the developmental level of the target group. Media need to deliver a faithful presentation of the information. The degree of problem definition and structuring is need to be aligned with what is intended for the target group.

Reducing extrinsic task load in task information includes avoiding the use of distracting, non-functional context-information, including the use of confusing, stereotyping, exclusive or implausible contexts. Images, sound, and speech should be maximally noticeable. Printed characters, fonts need to be well readable, media controls should work intuitively, and if necessary, tools to perceive stimuli (e.g. microscope, headphones) need to function well.

2.3.3 Facilitating Responding

In the fourth phase the test taker prepares and expresses responses on the assignment. It needs to be clear on which target material he or she needs to act, how and in which mode is to be responded, and which sensory channels should be used, with or without the use of strategical tools (e.g. dictionaries, search engines), and technical tools (E.g. calculators, rulers, carpentry tools).

In order to create appropriate intrinsic task load (see Table 2.3) it is essential that all required facilities for responding activate the behavior necessary to observe the target skill in the intended task situations. This relates to the choice of material to be acted on, the level of support given (e.g. hints, structure), the format and the sensory channel that best expresses the intended behavior. Response tools should enable a natural expression of that behavior.

Extrinsic task load is reduced by stripping all sources of distraction, confusion, or mental overload. This can be done by deleting material, that is not to be acted upon in a task, avoiding unnecessary interactions, avoiding response channel-switches (from verbal-to visual), avoiding non-corresponding orders between task information text and (multiple choice) response options, using direct response modes. Finally, the response tools themselves should function technically fluently and intuitively.

2.3.4 Facilitating Monitoring and Adjusting

During all phases of responding, an overarching process of monitoring and adjusting of the task solution process takes place. In various domains, these self-regulation skills are seen as an inseparable part of the target skill, e.g. reflecting and revision in writing (Deane 2011).

In general, the presentation of the assignment already implies support of self-regulation: orientation on task requirements and anticipated task outcomes form the basis of monitoring by the test-taker. In addition, lay-out of the individual assessments tasks and the navigation through the test can support self-regulation on the part of the test taker.

In order to create appropriate intrinsic task load it is essential that navigation design and item formatting optimally support information processing, responding and self-regulation. This is to be achieved through presenting information that is aligned with the intended situation in which the test taker is supposed to perform.

In order to prevent extrinsic task load it is essential that the test taker does not get stuck in the task, because of low perceptibility, readability of item parts, causing mental overload, or confusion. A well-known phenomenon is that of split attention, which takes place when test takers have to look for information elements that are spread over different locations in order to respond to the task (Ginns 2006). Finally, navigation can be assisted by providing feed-back about the progress status (status bars, finished and unfinished items) and by providing tools to navigate through the test without getting lost.

2.4 An Application of the Test Accessibility Framework: The Dutch Driving Theory Exam

In this section we describe an application of our accessibility improvement framework in the context of the Dutch driving theory exam.

2.4.1 Innovations in the Dutch Traffic Theory Exam for Car Drivers

During the last decade the Dutch theory exam for prospective car drivers (B license) has undergone major changes. First of all, driving a car is seen as a complex task that is inter-twined with daily life tasks. In addition to rule application and vehicle handling, strategic skills are seen as vital for proficient driving. Examples of these are aligning life-level goals with travel purposes, reflecting on the effects of own driving behavior for traffic safety, planning and adjusting the route according to travel purposes. Throughout Europe (Hatakka et al. 2002; Peräaho et al. 2003) the learning outcomes for driver training have been changed accordingly, which has had major implications for educational designs and driver exams. The content of driver exams, both regarding the practical and theoretical exams, has been changed. In Dutch practical exams, test takers are expected to choose their routes independently and demonstrate self-reflection on the quality of driving (Vissers et al. 2008). In theory exams, attention is paid to the measurement of higher order skills, specifically hazard prediction in traffic, in addition to the application of traffic rules in meaningful traffic situations.

Second, technological assets for item construction, presentation and delivery have developed fast during the last decade (Almond et al. 2010; Almond et al. 2010; Drasgow et al. 2006; Drasgow and Mattern 2006). The use of computer-based online item authoring software enabled the use of innovative interactive items types, including hotspot, drag and drop, and timed response items, that more closely resembled the required responses as used in daily traffic. It is expected that both the change in content and in item types could have benefits for validity, which in the end improves driver safety.

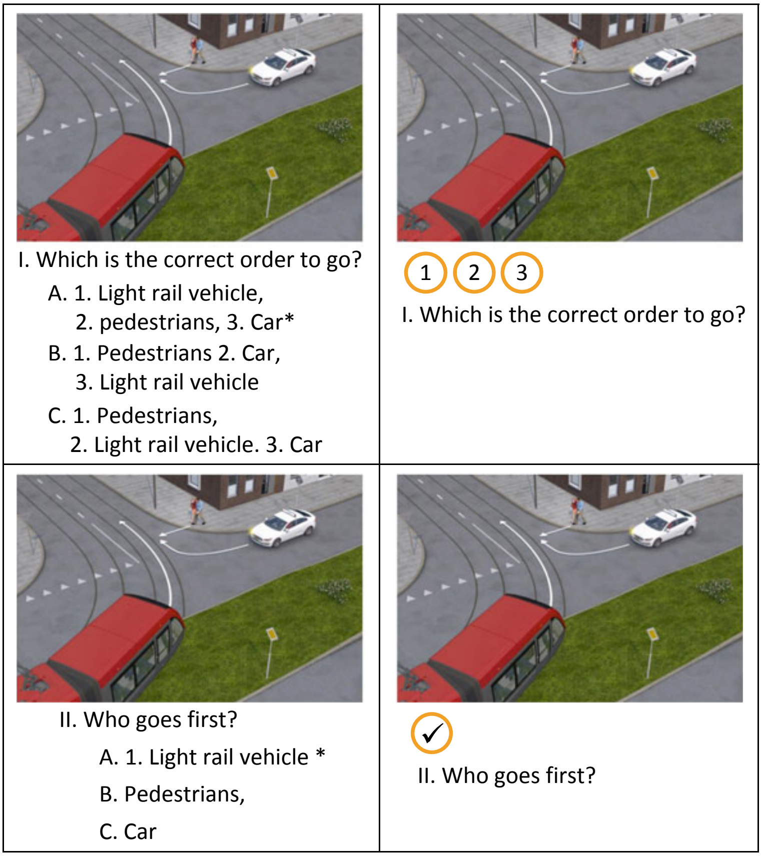
In practice, theory exam delivery has changed throughout many European countries from linear classroom-based delivery into individual computer-based, where its content follows a European Directive on driving licenses (European parliament and council 2006). With regard to the Dutch theory exam, a first step in the transition of linear paper-based into non-linear computer-based exams involved the redesign of the existing item collection, allowing online delivery in approximately 35 examination centers throughout the country.

The theory exam consists of two parts. The first part is a 25 item hazard perception test, which will not be discussed in this chapter. The second part, a 40 item subtest regarding traffic rules embedded in a traffic situation, is discussed here. The cut-score for passing this traffic rules test is a 35 out of 40 items correct score. The items typically consist of a specific traffic situation, presented in a picture, where the driver of a ‘leading car’, generally seen from the back, is about to carry out a traffic task (e.g. merging, turning, crossing, parking, overtaking) on a road section or an intersection. The candidate takes the perspective of the driver of this leading car.

Shortly after the introduction of the new computer-based delivery system a first evaluation study was carried out in 2014, in which the accessibility of a part of the redesigned items was reviewed, using the checklist described in Sect. 1.2 of this chapter. Results indicated possible sources of reduced accessibility (Zwitser et al. 2014). In general, most of the theory items were considered as accessible. Items that required a decision about the order of who goes first, second, third, fourth on an intersection using traffic rules and traffic signs, were seen as potentially less accessible. Other evaluative outcomes related to the stimuli used: in some cases the reviewers noted the use of too complex pictures, that did not correspond with the usual drivers’ perspective, e.g. a helicopter view of the car to determine dead angles around it.

2.4.2 Applied Modifications in the Response Mode of Theory Items

Following the results of the 2014 item screening, items that related to the application of Dutch traffic rules on intersections have been modified in order to improve the accessibility. Before 2014, test-takers for the Dutch theory exam were presented rule-of-way-at-intersections items in which they had to choose from verbal options (see left part if Fig. 2.3). The stem could be: “Which is the correct order to go?”. Test takers had to choose the correct option, which was a verbal enumeration of road users. In the 2014 study it was noted that the verbal response method could add extra extrinsic task load: keeping the verbal order in short term memory and compare it to the visual presentation of the road users in the item stimulus. Therefore the response mode for the rule-of-way-at-intersections items was changed into a direct visual response. As a response to the two respective item stems, test takers were either asked to drag corresponding rank numbers to the vehicle to indicate right of way order or to drag a check symbol towards the road user in the picture, who can go first at the intersection.



I. Which is the correct order to go?

A. 1. Light rail vehicle, 2. pedestrians, 3. Car*

B. 1. Pedestrians 2. Car, 3. Light rail vehicle

C. 1. Pedestrians, 2. Light rail vehicle. 3. Car

1 2 3

I. Which is the correct order to go?

II. Who goes first?

A. 1. Light rail vehicle *

B. Pedestrians,

C. Car

✓

II. Who goes first?

Fig. 2.3 Two items adapted form verbal into visual response mode

2.4.3 Psychometric Indications of Accessibility Improvement?

In studies regarding item modification, no clear cut indications are given of what counts as psychometric proof of improvement in item accessibility. Following prior studies of Beddow et al. (2013) and Beddow et al. (2011) we make the following argument.

From an accessibility perspective, a decrease in item difficulty after a modification is acceptable when it is caused by a reduction in extrinsic task load. In this case, the modified item had been more difficult for the wrong reasons, for instance because it required access skills unrelated to the task for the intended target group or because of inconsistencies in item presentation. Oppositely, a decrease in difficulty after modification would be less acceptable, when caused by a parallel reduction of intrinsic task load, e.g. for instance because certain knowledge is no longer needed to respond correctly to the item. Vice versa, an increase in item difficulty is acceptable as long as there has been no increase in extrinsic task load, and as long as the increase in difficulty is balanced by other items in the intended test.

Ideally, item discrimination will at least remain the same or preferably increase after modification. The reduction of extrinsic task load can be expected to come with a more clear coverage of the target skill.

In the current study, the reasoning above was used, to investigate changes in item response format and their effect on item difficulty and item discrimination. The leading research question was:

To what extent are differences in response format, i.e. verbal or visual, related to item difficulty and item discrimination, when item type and content have been accounted for?

2.4.4 Item Selection

In this study, only items regarding the rule of way at intersections, involving three or four road users were taken into consideration, because the verbally presented orders within these items were expected to cause extrinsic task load.

For intersection items with three or four road users, including the leading car perspective, three different item stems have been used in either response mode. (1) “Which is the correct order to go?” (2) “Who goes first?” (3) “To whom do you have to give way?”

2.4.5 Data Collection

To enable the estimation of item parameters for the subset of chosen items we used the test data of 345 versions of the Dutch theory test, administered in the period between fall 2015 and spring 2018. The test data pertained to 294.000 test takers. Due to the fact that new items were added to and old items were removed from the item bank, the number of versions in which an items appeared, varied between 1 and 344, with a median value of 5 versions.

The test data did not contain data from accommodated versions of the theory test, i.e. individually administered versions, read aloud by test assistants.

2.4.6 Data Analyses

For the purpose of this study the test data of 107 intersection items were used, in which three or four road users arrive at the intersection.

Table 2.5 Prediction of Item difficulty (*Beta*) by item stem type and response mode in a multiple regression analysis

Variables in equation	*B*	*SE*		*t*	Sig.
Constant	−.22	.12		−1.89	.06
Stem: Which is the correct order to go?	.34	.17	.19	1.99	.05
Excluded variables					
Response mode: verbal (0)- visual (1)			.028	.197	.84
Stem: Who goes first?			−.043	−.35	.73

Multiple $R = .19$; Df regression = 1; Df residual = 105

Using Dexter, an R-application for Item-response calibrations (Maris et al. 2019), necessary item parameters have been estimated. A one parameter model was used to estimate beta-values. Mean item-rest correlations (Rir-values) were calculated for each item across the versions, where it had appeared. Using multiple regression analyses the predictive effects of two item features for item difficulty and item discrimination were assessed: item stem (three types, see Sect. 2.3.3) and response mode (verbal vs. visual).

2.4.7 Results

Multiple regression analyses with item difficulty (Beta-parameter) as a dependent variable (see Table 2.5) revealed that items in which test takers had to determine the correct order to go on the intersection were more difficult than items with a different stem, where the first to go was to be determined (Beta = .19, $p = .05$). Response mode was not associated with item difficulty (Beta = .028, $p = .84$)

Multiple regression analyses with item discrimination (item-rest correlation) as a dependent variable (see Table 2.6) showed that response mode was significantly associated with item discrimination (Beta = .04, $p = .00$). Items with a visual (drag and drop) response had higher item-rest correlations than items with a multiple choice verbal response. The type of stem used was not related to item discrimination.

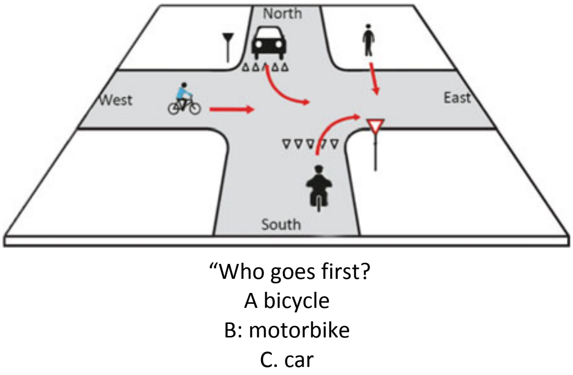
In Fig. 2.7 the item parameters for 16 paired items are displayed in a scatterplot. These pairs involve identical tasks and traffic situations, but differ in response type (visual response vs. verbal mode). It can be noticed from the upper part of Fig. 2.7 that for most of the items the change in difficulty is relatively low, after response mode change, with one clear exception. "Item A" in Fig. 2.4 is clearly more difficult in the visual response mode (beta-parameter = .13) than in the verbal mode (beta = −2.29). The content of this item is schematically depicted in Fig. 2.4. In order not to expose the original items, the original picture has been re-drawn. Item A involves four road users approaching a 4-way intersection. The traffic signs mean that participants on the East-West road have right of way vis a vis the participants in the North-South road. A pedestrian on the sidewalk North-left crossing into south direction, a car

Table 2.6 Prediction of Item discrimination by item stem type and response mode in a multiple regression analysis

	B	*SE*	*Beta-weight*	*t*	Sig.
Constant	.21	.01		32.40	.000
Response mode: verbal (0)- visual (1)	.04	.01	.35	3.76	.000
Excluded variables					
Stem: Which is the correct order to go?			−.17	−1.24	.218
Stem: Who goes first?			.06	.54	.588

Multiple $R = .35$; Df regression = 1; Df residual = 105

Fig. 2.4 Sketch of "item A"

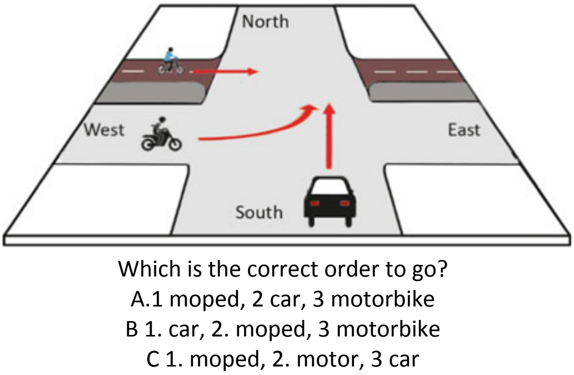


"Who goes first?
A bicycle
B: motorbike
C. car

turning left from the north lane, a motorbike from the south lane turning right and a cyclist from the west crossing in east direction. The question is who goes first out of four road users. The higher difficulty in the visual mode is most likely explained by the fact that in the verbal mode, the pedestrian is not mentioned in the response options. For pedestrians, the general rule of way does not apply. It only applies to drivers or riders of a vehicle. Because the pedestrian is not mentioned as an option, the test takers do not need to take into account this traffic rule. In the visual response mode however, the pedestrian needs to be considered, in order to indicate the road user who goes first, which increases the intrinsic task load.

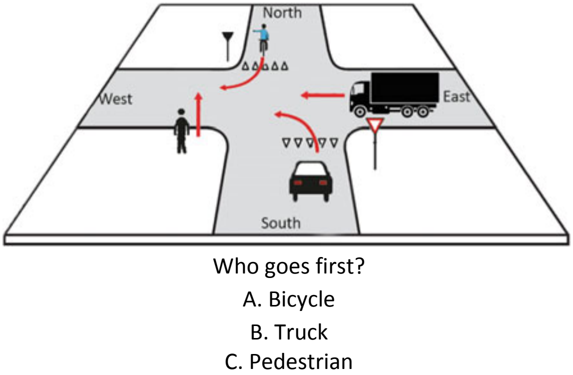
In the bottom part of Fig. 2.7, the Rir-values for 16 pairs of identical items, that were changed from a verbal into an visual response mode, are displayed in a scatterplot. In general it can be noted that most items have higher Rir-values after the adaptation of the response mode from verbal into visual. Item A, described above, and B and C have the highest increase in Rir-value. For item A the extra rule regarding the pedestrian, that needs to be considered in the visual response mode, will probably lead to an increase in possible options, contributing to the higher Rir.

Item B, the content of which is depicted in Fig. 2.5, shows a clear increase in difficulty (from 1.07 to 1.70), where at the same time the Rir-value increases (from .13 to .23). In item B three road users arrive at a non-signalized 4-way intersection,



Which is the correct order to go?
A.1 moped, 2 car, 3 motorbike
B 1. car, 2. moped, 3 motorbike
C 1. moped, 2. motor, 3 car

Fig. 2.5 Sketch of "item B"



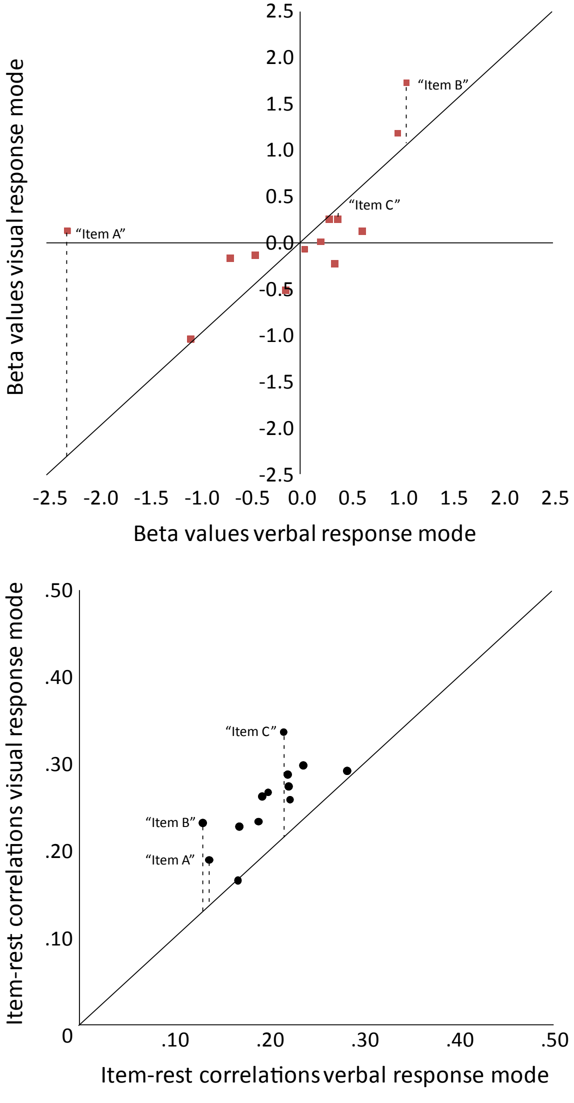
Who goes first?
A. Bicycle
B. Truck
C. Pedestrian

Fig. 2.6 Sketch of "item C"

including a motorbike and a cyclist from the same direction (west), where the motorbike intends to make a left turn north, and the cyclist intends to cross eastwards. A car intends to cross from the south lane to the north lane. The explanation of the increased Rir is very likely the increased number of possible orders to go in the visual mode compared to the verbal mode. However, the Rir-value is still below what can be considered as appropriate.

Item C (depicted in Fig. 2.6) has the largest increase in Rir (from .22 to .34), but interestingly enough a relatively stable difficulty value (Beta decrease from .35 to .25). In this item, four road users arrive from four different directions at a 4-way intersection. The traffic signs mean that drivers and riders on the East-West road have right of way vis a vis drivers and riders in the North-South road. A car from south intends to turn left towards the west lane, a pedestrian is crossing from south to north on the left sidewalk, a cyclist turns right from the north lane into the west lane, and a truck crosses from the east lane towards the west lane.

Fig. 2.7 Item difficulty (betas) and item rest correlations for item pairs in a verbal and visual response model



Both in the visual and verbal mode all applicable traffic rules need to be applied to arrive at the correct answer, although the car is not mentioned in the options. In this case, there is a possible exclusive response mode effect, in which the visual response takes away construct irrelevant task load, causing an increase in item discrimination (Fig. 2.6).

2.5 Discussion

In this chapter we presented a framework for design and review of accessible assessment tasks. We deliberately use the term “assessment task” to refer to tasks with any size, meant to be used in an assessment, that can have any form, be it a test or a performance assessment. Access in the context of educational testing refers to the opportunity for a student to demonstrate proficiency on a target skill. The framework has three basic elements: (1) cognitive load; (2) the test taker task solution process; and (3) parts of the assessment task presentation.

First, using the concept of cognitive load (Sweller 2010), we contended that each assessment task, intended to measure a target skill such as a mathematical word problem or solving an item about traffic rules on an intersection has a certain amount of intrinsic task load and extrinsic task load. Improving accessibility basically means optimizing the amount of intrinsic task load and minimizing the amount of extrinsic task load. It is up to the test designer to specify, what task elements are considered as intrinsic task load, and what comprises acceptable levels of intrinsic load for the target group test taker. In addition, elements that are seen as unavoidable extrinsic task load needs to be specified.

Second, using insights from cognitive laboratory studies we proposed to take the test takers task solution processes in mind in such a way as to facilitate these processes in assessment task design. In our framework, five general mental processes were considered that test takers go through when solving a problem. These entailed: (1) orientation on the task purpose and requirements, (2) comprehending task information, (3) devising a solution, (4) articulating a solution and (5) self-regulation including planning, monitoring and adjustment of the solution process. The fifth part of the process model was added by applying self-regulation theories on task execution (Winne 2001). Depending on the target domain (mathematics, science, reading, driving a car), these task solution processes may be broken down into further detail and specified for the skills content at hand. Additionally, test takers may make short cuts and arrive automatically at solutions, because sub processes are automated or take place unconsciously.

A third element of our model pertains to the parts of the assessment task presentation. We distinguished the assignment, task information, facilitation of responding, and item lay-out and test navigation. Again, these parts can be broken down further into subcategories or elaborated further. For instance, group tasks with collective outcomes and individual test taker contributions are not mentioned in our framework.

Combining these three elements we presented an approach for item design and screening, summarized by means of a checklist. Principles of existing instruments, such as the Test Accessibility and Modification Inventory (TAMI, Beddow et al. 2008) were applied in the checklist. In the checklist we concentrated on removing or avoiding extrinsic task load from assessments tasks, without changing the level of intrinsic task load. The idea behind it was that part of the inaccessibility is caused by the way assessments tasks are presented.

We reported a study in the context of theoretical driving exams, specifically regarding the application of rules of way on traffic intersections. This is a relatively complex skill, because Dutch traffic highway code has many different rules and signs that regulate the right of way on intersections. The transition of paper-based exams towards individually administered computer-based exams came with some item modifications after a first item review study (Zwitser et al. 2014). One of these modifications was the change of response mode regarding intersection items from verbal options into direct visual (drag and drop) response in order to meant to improve item accessibility. Using our framework we investigated whether this modification, meant to reduce the reading load and the necessity to change from visual orders to verbal orders, had effects on item difficulty and item discrimination. Reasoning from our accessibility framework, it would be desirable that a change in response mode would not affect the difficulty, but it should affect item discrimination. Reducing extrinsic task load should result in an improved target skill representation, which can be expected to show in an increase in item discrimination parameters.

The study including 107 items showed that the response mode mattered to item discrimination, as measured by the corrected item total correlation, although the effects were not spectacular. Regression analyses showed that items with a visual response mode discriminated better than those with a verbal response mode, when item stem differences had been accounted for. In addition we found that for items that address identical intersection tasks and that had identical stems, those with a visual response types discriminated significantly better than those with verbal response types. At the same time response mode was not significantly related to item difficulty, as measured by the IRT beta coefficient. There were a few exceptions for this result. We found that for one item the changed response type resulted in higher difficulty. This result could be explained by the increase of the number of traffic rules that needed to be applied to arrive at an correct answer. In the verbal option mode, one of the traffic rules did not need consideration, because the road user was not mentioned in the options.

This finding suggested that modifying items in order to improve accessibility needs to be done in combination with a look at the intrinsic task load at the same time. In follow-up research we intend to develop predictive models for item difficulty and item discrimination that combine intrinsic and extrinsic task features. In follow-up research it is our aim to identify task features that comprise intrinsic task load and task elements that comprise unavoidable and irrelevant task load. By specifying these task features in advance, it can be determined which elements of the assessment tasks need some kind of access skill and which address target skills.

To conclude this chapter some theoretical, practical and methodological limitations and challenges can be mentioned. First, it is agreed upon that accessibility is not seen as a static test property but instead, as the result of an interaction among test features and person characteristics that either permit or inhibit student responses to the targeted measurement content (Kettler et al. 2009). In the current study on traffic intersection items, the data did not allow us to look for specific interactions between test takers features and item features. A research question could have been whether test takers with reading or attention problems benefitted more from the changed response modes than other test takers.

Second, a validity issue is to which extent changes in accessibility affect extrapolation of test results to target situations (Kane 1992, 2004). Do changes in accessibility correspond with accessibility in the target situations? The question is, which are absolutely unavoidable sources of extrinsic task load? What if the target situations contain comparable sources of extrinsic task load? Applying to our example regarding driving on intersections, there may be all kinds of distracting information, e.g. adverse weather conditions, unexpected maneuvers of other road users, crying kids in the back of the car. To extrapolate from the test towards performance in the practice domain, we probably need a sound definition of the target skill and the target situation. In this example, drivers need to be situationally aware, prepared to notice changes in the traffic situation. A theory test as described, with a picture and a question like "who goes first", cannot account for these situational changes. It was not designed for the purpose of taking into account situational awareness (Crundall 2016; McKenna and Crick 1994). It would however be possible to design an interactive simulator-driven test with scenarios that presents tasks, where application of traffic rules is part of a larger set of target skills, including determining whether it is safe to cross, taking into account both the rule and the speed and position of other road users.

This brings us to a final, much broader issue, involving the question how to map the development of access skills and target skills over time, using developmental trajectories and what this means for designing construct relevant and accessible assessments, that inform learning processes (Bennett 2010; Pellegrino 2014; Zieky 2014). How do target skills develop over time? To what extent do skills develop into access skills for other skills, such as reading skill for math? How do these skills get intertwined to form new target skills? Which intermediate learning progressions take place? In order to inform learning design for students with different needs in different developmental stages, we to need to find answers to several design questions: what constitutes necessary intrinsic task load to cover target skills? What goes on in test takers during task solution? What levels of access skills can be assumed and which sources of extrinsic task load are unavoidable or unnecessary? In order to find answers to these questions in the future, a combination of approaches provide a more comprehensive framework, including an evidence centered design of assessments (Mislevy and Haertel 2006) using cognitively based student models (Leighton and Gierl 2011) elaborated in specified item models (Gierl and Lai 2012) to be tested in multiple educational contexts.

Acknowledgements I am grateful to Frans Kamphuis, Timo Bechger, and Angela Verschoor for the item calibration activities they have carried out.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219–234.

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., et al. (2010a). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities— A foundation for research. *The Journal of Technology, Learning, and Assessment, 10*(5), 1–41.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2010b). Enhancing the design and delivery of assessment systems: A four process architecture. *The Journal of Technology, Learning, and Assessment, 1*(5), 1–63.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test accessibility and modification inventory*. Nashville, TN: Vanderbilt University.

Beddow, P. A., Kurz, A., & Frey, J. R. (2011). Accessibility theory: Guiding the science and practice of test Item design with the test-taker in mind. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students bridging the gaps between research, practice, and policy* (pp. 163–182). New York: Springer.

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2013). Test accessibility: Item reviews and lessons learned from four state assessments. *Education Research International, 2013,* 1–12.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research & Perspective, 8*(2–3), 70–91.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245.

Crundall, D. (2016). Hazard prediction discriminates between novice and experienced drivers. *Accident Analysis and Prevention, 86,* 47–58. https://doi.org/10.1016/j.aap.2015.10.006.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20,* 405–438.

Deane, P. (2011). *Writing assessment and cognition*. Research Report ETS RR–11-14. Princeton: Educational Testing Service.

Dolan, R. P., & Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessment. *IDA Perspectives, 27*(4), 22–25.

Dolan, R. P., & Hall, T. E. (2007). Developing accessible tests with universal design and digital technologies: Ensuring we standardize the right things. In L. L. Cook & C. C. Cahalan (Eds.), *Large-scale assessment and accommodations: What works* (pp. 95–111). Arlington, VA: Council for Exception Children.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram & R. Hambleton (Eds.), *Computer-based testing and the internet* (pp. 59–76). Hoboken, NJ: Wiley.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Washington, DC: American Council on Education/Praeger Publishers.

European parliament and the council of 20 december 2006 on driving licences. (2006). Directive 2006/126/EC (recast). Official Journal of the European Union (2016), December 30, L 403/18- L 403/13. Retrieved at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006L0126&from=en.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing, 12*(3), 273–298.

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction, 16*(6), 511–525.

Hatakka, M., Keskinen, E., Gregersen, N. P., Glad, A., & Hernetkoski, K. (2002). From control of the vehicle to personal self-control: Broadening the perspectives to driver education. *Transportation Research Part F, 5,* 201–215.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2,* 135–170.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment, 4*(2), 1–23.

Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*(3), 3–16. https://doi.org/10.1111/j.1745-3992.2008.00124.x.

Kettler, R. J., Braden, J. P., & Beddow, P. A. (2011). Test-taking skills and their impact on accessibility for all students. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students bridging the gaps between research, practice, and policy* (pp. 163–182). New York: Springer.

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education, 84,* 529–551.

Leighton, J. (2017). *Using think-aloud interviews and cognitive labs in educational research*. New York, NY: Oxford University Press.

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

Mace, R. (1997). *The principles of universal design* (2nd ed.). Raleigh, NC: Center for Universal Design, College of Design. Retrieved May 20, 2010, from http://www.design.ncsu.edu/cud/pubs_p/docs/poster.pdf.

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2019). DEXTER. Data management and analysis of tests, version 0.8.4. Manual Retrieved at: https://cran.r-project.org/web/packages/dexter/dexter.pdf.

McKenna, F. P., & Crick, J. L. (1994). *Hazard perception in drivers: A methodology for testing and training*. TRL Contractor Report (313).

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25,* 6–20.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*(2), 65–77.

Peräaho, M., Keskinen, E., & Hatakka, M. (2003). *Driver competence in a hierarchical perspective: Implications for driver education*. Turku: Turku University, Traffic Reseacrh.

Roelofs, E. C. (2016). Naar richtlijnen voor het ontwerp van toegankelijke toetsopgaven [Towards the design of accessible test items]. *Examens, 13*(3), 4–11.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457–490.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22,* 123–138.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [2-1-11], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

Vissers, J., Mesken, J., Roelofs, E. C., & Claesen. (2008). New elements in the Dutch practical driving test: a pilot study. In L. Dorn (Ed.), *Driver behavior and training* (Vol. III, pp. 37–50). Hampshire: Ashgate Publishing Limited.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 153–189). Mahwah, NJ: Lawrence Erlbaum Associates.

Winter, P. C., Kopriva, R. J., Chen, Ch S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences, 16,* 267–276.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa, 20,* 79–87.

Zucker, S., Sassman, C., & Case, B. J. (2004). *Cognitive Labs*. Pearson Technical report. New York, NY. Retrieved at http://images.pearsonassessments.com/images/tmrs/tmrs_rg/cognitivelabs.pdf.

Zwitser, R., Roelofs, E., & Béguin, A. (2014). *Rapportage van het onderzoek naar de psychometrische kwaliteit van de huidige CBR theorie-examens [Research into the psychometric quality of the current Dutch driving theory-exams].* Arnhem: Cito.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 3
The Design and Validation of the Renewed Systems-Oriented Talent Management Model

Arnold J. Brouwer, Bernard P. Veldkamp and Marieke Vroom

Abstract Within the field of talent management, it is common to make use of assessments in order to select and appoint the best fitting human talent at the right moment and for the right time. Characteristic for current assessment instruments is that they primarily focus on measuring the talents and motives of (potential) employees. There is no known psychometric test instrument available that links these human characteristics to the strategy and purpose of an organisation. In order to bridge this gap, this chapter introduces the systems-oriented talent management (STM) model. The STM model is elaborated into three STM diagrams in which business purpose and human talent are aligned from both a managerial and a psychological perspective. The management building blocks framework and the systems theory, applied to the field of organisation, are used to describe the interrelations between the components of the STM model. Lexical-semantic analyses are conducted to link human talent to the core elements of an organisation's purpose. This study contributes to achieving a sustainable match between organisations and their employees that, in our rapidly changing world, is able to move along with continuous adaptations in business strategy and the associated adjustments in team compositions and tasks. However, not only the workplace is subject to change. The field of education is affected as well, and the pupils and students of today are the employees of tomorrow. Therefore, this chapter also explores the possible application of the STM in the educational world.

A. J. Brouwer (✉) · M. Vroom
RCEC, Vaassen, The Netherlands
e-mail: a.j.brouwer@rcec.nl

B. P. Veldkamp
University of Twente, Enschede, The Netherlands

© The Author(s) 2019
B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_3

3.1 Introduction

3.1.1 Problem Situation and Purpose of the Study

Our ever more rapidly changing world demands a lot from organisations, and in particular their HRM specialists, to find, retain and promote the right people for the right positions. In their search for support, a growing number of organisations turn to talent management, a specialisation defined as the process of discovering, developing and retaining top talent (Michaels et al. 2001). Talent management focuses on the recruitment and selection of the right people, helps employees develop in their professional roles and guides them to the next step in their careers. The aim is to be able to continuously anticipate the internal and external changes that all organisations face (Berger and Berger 2011).

Within the field of talent management, it is common to select employees by conducting assessments. Measuring talents and predicting success rates of candidates for certain positions within companies is a valuable way to select potentials and to prevent mismatches (Berger and Berger 2011; Schoonman 2013). This procedure is founded in psychological testing techniques, defined as systematic and standardised procedures for measuring human characteristics of a candidate (Smith and Robertson 1986; Michaels et al. 2001).

Although the field is equipped with many reliable and valid test instruments, it is surprising that the majority of these measures solely map the human characteristics side of the match between the organisation and its worker. This is done by either identifying inner characteristics ('attributes and/or attitudes') or diagnosing visible behaviour and skills ('abilities'; McDonnell and Collings 2011). Together with intelligence tests that measure cognitive capabilities, these attributes, attitudes and abilities are presumably the strongest predictors for success or failure (Schmidt and Hunter 1998; Schmidt et al. 2016).

Next to these instruments that measure people's inner and/or visible characteristics, the field of management science designed different models to represent an organisation as a mixture of jointly interacting management building blocks (Galbraith 2002; Hendricks and Singhal 1996; Tillema and Markerink 2006; Nieuwenhuis 2006). Even though this resulted in clear managerial constructs that, from their nature, could potentially be measured as well, there is no known psychometric test instrument available that links the talents and motives of (potential) employees to the strategy and purpose of an organisation.

Because such an instrument is missing, HR consultants and other assessors have to depend on their own knowledge and expertise in psychology and business administration to make this connection (Cable and Yu 2007; Van Beirendonck 2010). In practice, this is a difficult task, which often results in a limited match between those characteristics and the organisation. Vice versa, HR consultants also face difficulties in indicating exactly what the organisation is looking for in terms of human qualities. Not rarely this leads to broad, and sometimes contradictory, organisational strategy and culture models and individual function and competence profiles. Who has not read mission statements or descriptions of core values in which both staff and clients hardly recognise themselves or the firm? And who has not seen the advertisements in which the candidate sought after is a result-oriented go-getter who is independent, but also works well in teams, and is not only flexible but also meticulous?

Consequently, the long-term results of strategy and culture change programmes and recruitment and selection procedures often are unsatisfactory. While at short-term, there seems to be a match between the employee and the organisation, problems arise when the job content is adjusted or the organisation changes. As a solution, companies often opt for top-down management interventions, short employment contracts, or onboarding: a period in which employees learn the knowledge, skills and behaviours to be effective in the renewed organisational context or in their new function, to first check whether it really works. As a result, the organisation is confronted with high costs and both parties might end up disappointed.

To integrate psychological questionnaires that test human characteristics with models for representing organisations in a series of managerial blocks, Brouwer (2012) introduced the systems-oriented talent management (STM) model. This model is elaborated into three initial STM-scan diagrams in which business purpose and human talent are aligned from both a managerial and a psychological perspective.

The three initial STM diagrams were implemented as an online testing instrument, named STM-scan. This initial version of the STM-scan has been used over 1000 times as talent management instrument within different Dutch companies. Multiple intermediate evaluations established that this initial version of the STM-scan was experienced as a helpful assessment instrument for answering talent management questions regarding adoption and/or adjustment of corporate strategy and culture, recruiting and selecting new personnel, coaching and developing employees, and, outplacement and career advice to employees. For example, in Brouwer (2018, Chap. 6) interviews were held with a panel of four talent management experts who have several years of experience with the STM: a recruiter, a HRD consultant, a career coach and a business coach. The recruiter explains that the foremost value of STM lies in getting a sharp image of someone's qualities in a short amount of time. "And when you have a clear picture of what is needed [within the organisation], you can quickly come to the conclusion whether or not there is a fit." According to the career coach this accelerates the intervention process: "I dare say that it saves me at least a few appointments in a coach assignment." All four experts assent that the STM can provide insight into individuals, teams and organisations, and gives a clear image of their strengths and weaknesses. It makes it possible to look at people "for who they are, and not for what they show," says the business coach. This helps both the test taker and the test professional to more quickly find detailed answers to talent management questions. "That is often what is most important for clients who make use of the STM: they want to achieve their goals as effectively as possible. […] Ultimately, organisations hire people to realise something, not just because they like to employ people," notes the HRD consultant (pp. 174–175).

Since the initial STM was best practice oriented rather than evidence-based, a series of qualitative and quantitative studies was done to (re)design and validate this first model (Brouwer 2018). This resulted in a renewed evidence-based STM model and its elaboration in three renewed STM diagrams, that could take the place of the initial three STM-scan diagrams. The renewed STM not only provides a theoretical framework to align human talent and organisational purpose, but also supports HR consultants and other professionals with tools that measure both human talent and organisational purpose, and quantify their connection.

The management building blocks framework (MBBF; Nieuwenhuis 2006) and systems theory (Katz and Kahn 1966; Meadows 2008) are used to elaborate the interrelations within the new STM model. This provides a renewed way of linking human talent to the core elements of an organisation's purpose (Barile 2006, 2008; Mele et al. 2010), and is assumed to result in a fit between corporate and personal identity instead of a fit between a person's visible skills and a specific job profile that in this rapidly changing world is subject to continuous alteration.

3.2 Theoretical Framework

3.2.1 The Management Building Blocks Framework

Within the field of management science there are different models for representing an organisation as a dynamic and a constantly adapting organism. Widely recognised models are the 7S model (Peters and Waterman 1998), the Star Model (Galbraith 2002) and the European Foundation for Quality Management (EFQM) excellence model (Hendricks and Singhal 1996) with its application for the Dutch market in the INK (Instituut Nederlandse Kwaliteit) management model (Tillema and Markerink 2006). As visualised in Fig. 3.1, a common feature in these models is the use of a set of five management building blocks that jointly interact as a value chain, describing the composition of and joint interactions within the primary business process. This value chain is defined as the management building blocks framework, or MBBF (Nieuwenhuis 2006).

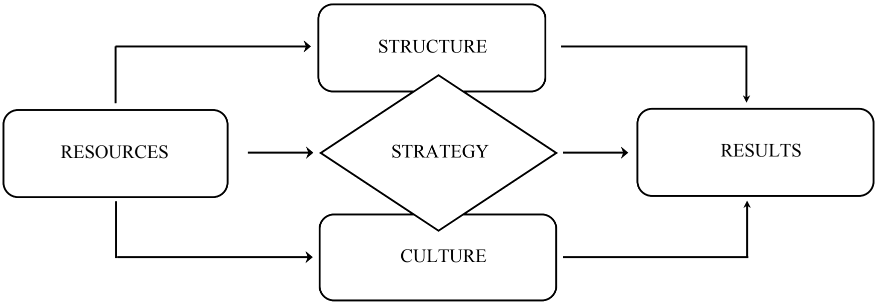


Fig. 3.1 The management building blocks framework, or MBBF (Nieuwenhuis 2006)

From a managerial point of view, the first building block of the MBBF, *structure*, is seen as organisational structure. This describes the way activities such as task allocation, coordination and supervision are directed towards the achievement of organisational aims (Pugh 1990). Structure affects organisational actions towards reaching the organisation's purpose in two ways: (1) it provides the foundation on which operating procedures and routines rest, and, (2) it determines which individuals get to participate in which decision-making processes, and to what extent their actions shape the organisation's purpose (Jacobides 2007).

The second building block, *culture*, is seen as organisational culture and defined as a set of shared mental assumptions that guide interpretation and action in organisations by prescribing appropriate behaviour for various situations (Ravasi and Schultz 2006). This set of shared assumptions is tangible in the organisation's climate, which is the shared meaning organisational members attach to the events, policies, practices and procedures they experience and the behaviours they see being rewarded, supported and expected (Ehrhart et al. 2014). Whereas organisational culture represents the predefined and desired image, the organisational climate embodies the actual identity.

The third building block, *strategy*, defined as business strategy, is known as the formulation and implementation of the organisation's purpose and initiatives taken by employees on behalf of its stakeholders (Nag et al. 2007). According to Wit and Meyer (2011), business strategy consists of two dimensions: (1) the strategy process, expressed in the amount of effectiveness of the organisation design, and (2) the strategy content, measured from the outcome of its employees' contribution.

For the execution of the organisation's purpose, an organisation needs the best fitting human talent, represented in the fourth building block, *resources*. The effect of the execution of the organisation's purpose is visualised in the interaction between *strategy*, consisting of *structure* and *culture*, and the building block *resources*. Its outcome is expressed in the fifth building block of the MBBF, *results*.

To elucidate the relationship between the organisation's purpose and human talent in order to find ways to gain the desired results, a systems-oriented view on the composition of and joint interactions between these five management building blocks is needed. This is found in the systems theory.

3.2.2 Systems Theory

Systems theory is an interdisciplinary theory about every system in nature, society and many scientific domains, as well as a framework with which phenomena can be investigated from a holistic approach (Capra 1997). It encompasses a wide field of research with different conceptualisations and areas of focus. Katz and Kahn (1966) applied systems theory to the field of organisations. Studied from its managerial context, systems theory is a theoretical perspective that helps to analyse the organisation seen as a whole and not as simply the sum of its elementary parts (Meadows 2008). Every organisation is seen as a group of interconnected and interrelated parts that jointly perform the organisation's purpose and that, mutually, are related to other organisations in its environment (Barile 2006, 2008; Mele et al. 2010).

From this perspective, the MBBF can be seen as a sub-system of five interrelated constructs. In order to design a model to align these constructs, the present study continues with a systems-oriented elaboration of the different interrelations between the five building blocks, studied from both their position in and contribution to the MBBF. This results in the design of the renewed evidence-based STM model and its elaboration in three renewed STM diagrams of the initial three STM-scan diagrams as introduced in Brouwer (2012).

3.2.3 Evidence-Based Systems-Oriented Talent Management

Looking at the MBBF (Fig. 3.1) from a systems-oriented perspective reveals three different paths between the building blocks *resources* and *results*. The first runs from *resources* via *structure* to *results*. The second path runs from *resources* via *culture* to *results*, and the third goes from *resources* via the higher order construct *strategy* towards *results*. The joint approach of these three paths forms the central idea behind the renewed STM. In this way, STM aligns human talent, found in *resources*, with the organisation's purpose, found in *structure*, *culture* and *strategy*, in order to achieve the predefined *results*.

Path 1: Structure

Organisational structure is operationalised into organisational effectiveness, defined as the efficiency with which an organisation is able to meet its objectives. It is about every employee doing what he or she does best. The main measure of organisational effectiveness for a business is generally expressed in terms of how well the achieved results compare with the predefined goals (Pedraza 2014).

Perceived from its position within the MBBF, organisational effectiveness is linked to the organisation's focus and the way the organisation is structured to achieve its goals (Yu and Wu 2009). This is related to the foundation of organisational structure (Pugh 1990). Approached from its contribution to the MBBF, Mitchell (2012) sees organisational effectiveness as a logic model that specifies how resources produce activities and output, which in turn will lead to outcomes. This is associated with the decision-making processes of organisational structure (Jacobides 2007).

In order to describe the first path of the MBBF, the interaction between the building blocks *resources* and *structure* needs to be unravelled. The interplay between *resources* and *structure* leads to a specific outcome, seen as the building block *results*. Since organisational effectiveness is considered as the effectuation of organisational structure, a corresponding type of resources element is required to study their interconnection. This can be found in personality facets underlying the five personality factors openness, conscientiousness, extraversion, agreeableness and neuroticism of the five factor model (FFM; Costa and McCrae 1985). More specified facets help to express the many meanings and components of personality. For example, the six facets of extraversion are warmth, gregariousness, assertiveness, activity, excitement seeking and positive emotions.

Path 2: Culture

Organisational culture can be approached from both an integral and an individual perspective. Whereas organisational culture defines the values and behaviours an organisation requests from its employees, organisational climate focuses on the employees' actual experiences and the attitudes or workstyles they see being rewarded and encouraged by the company (Ehrhart et al. 2014). Organisational culture represents the organisation's demonstrated image from the outside-in, found in the building block *culture*, and organisational climate embodies its actual present identity from the inside-out, linked to the building block *resources*. In studying its position in and contribution to the MBBF, organisational climate can be delineated from both a strategic approach and from a molar approach. The first approach considers organisational climate as the individual perception and representation of the work environment, focusing on a specific outcome (Kuenzi and Schminke 2009; Gimenez-Espin et al. 2013). The second approach concentrates on capturing the generic sense of the experiences people have at work (Schneider and Reichers 1983). Jointly, the strategic and molar approach convert organisational climate into a construct that represents both the position (Ravasi and Schultz 2006) and the contribution (Ehrhart et al. 2014) of the building block *culture* to the business purpose.

To describe the relationship of organisational climate with human talent, a second corresponding human characteristic is needed. This can be found in work values of the universal values model, or UVM (Schwartz 1992), which represents human motives and beliefs people have concerning what situations and actions are desirable in the working environment, such as independence, variety and income.

Path 3: Strategy

Business strategy, studied from an integral organisational perspective, dissects the organisation's purpose in organisational effectiveness and organisational climate. Within the MBBF, the two building blocks *structure* and *culture* jointly form their higher-order building block *strategy*. Whereas business strategy is seen as the execution of these two constructs, a same way of composing the building block *resources* is needed in order to study the interconnection between personality facets of the FFM and work values of the UVM. This was found in the concept of competences, which is the combination of these two human characteristics. Competences are defined as the sum of a person's abilities, intrinsic gifts, skills, knowledge, experience, intelligence, judgment, attitude, character and drive (Michaels et al. 2001). In the STM, this is elaborated in a set of 16 key competences, including entrepreneurship, initiative, creativity and empathy.

To link the individual competences to the integral business strategy, a higher-order construct, consisting of both a business strategy and a competence element, is required. This was found in the theory of team roles, which was first introduced by Belbin in 1981 as the result of a study on the question why some (management) teams succeed and others fail. He defined team roles as tendencies to behave, contribute and

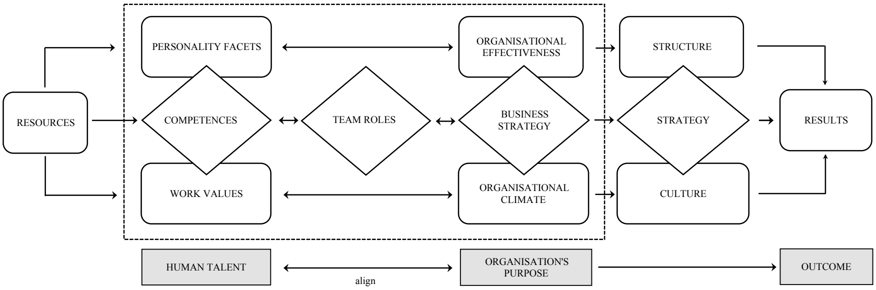


Fig. 3.2 The relationships between the STM elements and the five building blocks of the MBBF

interrelate with others in a particular way and used it to identify people's behavioural strengths and weaknesses in the workplace (Belbin 2010). In the STM, Belbin's eight team roles are defined in terms of work-related human activities, including innovate, activate and coordinate.

The different relationships between the STM elements and the building blocks are presented in Fig. 3.2. The three paths from *resources* to *results* as described above have been elaborated in the three renewed STM diagrams (Figs. 3.3, 3.4 and 3.5).

3.3 Renewed STM Diagrams

3.3.1 Renewed STM Diagram 1: Aligning Organisational Structure and Human Talent

Figure 3.3 presents the renewed version of the first STM diagram, which elaborates the systems-oriented interplay between the building blocks *structure* and *resources* of the MBBF in the relationship between organisational effectiveness and personality facets. The building block *structure*, seen as organisational structure, becomes tangible in the construct organisational effectiveness. Its position, dealt with as one of the management processes within the MBBF, is elaborated in the Deming quality circle (Deming 1986), known as the plan—do—check—act (PDCA) cycle, which is a frequently used problem-solving model in the field of quality management. In the cycle:

1. *plan* is the act of identifying opportunities and ways for improvement;
2. *do* refers to the actions necessary to effect the change;
3. *check* is the verification of whether the changes resulted in the desired improvements; and
4. *act* refers to what one does in response to the effects that are observed.

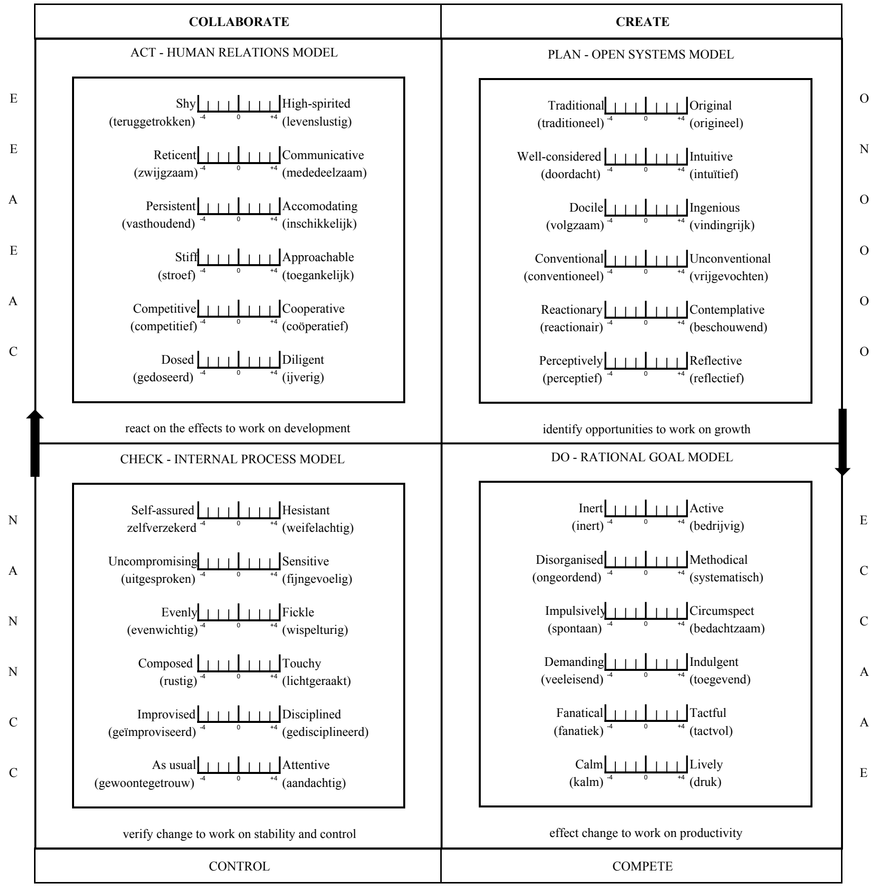

COLLABORATE
CREATE
ACT - HUMAN RELATIONS MODEL
PLAN - OPEN SYSTEMS MODEL
Shy (teruggetrokken) -4 0 +4 High-spirited (levenslustig)
Reticent (zwijgzaam) -4 0 +4 Communicative (mededeelzaam)
Persistent (vasthoudend) -4 0 +4 Accomodating (inschikkelijk)
Stiff (stroef) -4 0 +4 Approachable (toegankelijk)
Competitive (competitief) -4 0 +4 Cooperative (coöperatief)
Dosed (gedoseerd) -4 0 +4 Diligent (ijverig)
Traditional (traditioneel) -4 0 +4 Original (origineel)
Well-considered (doordacht) -4 0 +4 Intuitive (intuïtief)
Docile (volgzaam) -4 0 +4 Ingenious (vindingrijk)
Conventional (conventioneel) -4 0 +4 Unconventional (vrijgevochten)
Reactionary (reactionair) -4 0 +4 Contemplative (beschouwend)
Perceptively (perceptief) -4 0 +4 Reflective (reflectief)
E
E
A
E
A
C
O
N
O
O
O
O
react on the effects to work on development
identify opportunities to work on growth
CHECK - INTERNAL PROCESS MODEL
DO - RATIONAL GOAL MODEL
Self-assured zelfverzekerd -4 0 +4 Hesistant (weifelachtig)
Uncompromising (uitgesproken) -4 0 +4 Sensitive (fijngevoelig)
Evenly (evenwichtig) -4 0 +4 Fickle (wispelturig)
Composed (rustig) -4 0 +4 Touchy (lichtgeraakt)
Improvised (geïmproviseerd) -4 0 +4 Disciplined (gedisciplineerd)
As usual (gewoontegetrouw) -4 0 +4 Attentive (aandachtig)
Inert (inert) -4 0 +4 Active (bedrijvig)
Disorganised (ongeordend) -4 0 +4 Methodical (systematisch)
Impulsively (spontaan) -4 0 +4 Circumspect (bedachtzaam)
Demanding (veeleisend) -4 0 +4 Indulgent (toegevend)
Fanatical (fanatiek) -4 0 +4 Tactful (tactvol)
Calm (kalm) -4 0 +4 Lively (druk)
N
A
N
N
C
C
E
C
C
A
A
E
verify change to work on stability and control
effect change to work on productivity
CONTROL
COMPETE

Fig. 3.3 Renewed STM diagram 1: the alignment of organisational structure and human talent, elaborated in the relationship between organisational effectiveness and personality facets

The contribution of organisational structure to the building block *results* becomes tangible in the competing values framework, or CVF (Quinn and Rohrbaugh 1983). The CVF is a cycle for process improvement that consists of four models:

1. the *open systems model*, in which growth, new resources, and external support are worked on by maintaining flexibility and availability;
2. the *rational goal model*, where productivity and efficiency are worked on through goal setting and planning;
3. the *internal process model*, in which stability and control are worked on through information management and coordination; and
4. the *human relations model*, in which human resources are developed by maintaining cohesion and morale (Cameron and Quinn 2011).

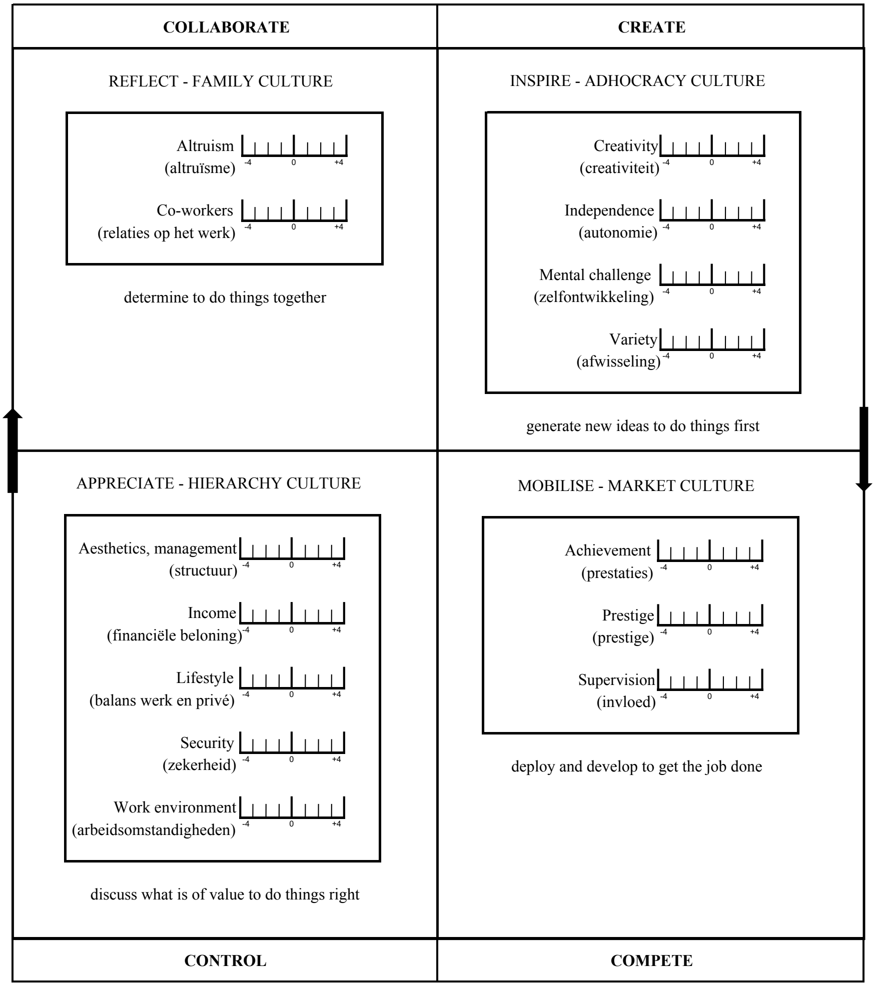

COLLABORATE
CREATE
REFLECT - FAMILY CULTURE
Altruism (altruïsme) -4 0 +4
Co-workers (relaties op het werk) -4 0 +4
determine to do things together
INSPIRE - ADHOCRACY CULTURE
Creativity (creativiteit) -4 0 +4
Independence (autonomie) -4 0 +4
Mental challenge (zelfontwikkeling) -4 0 +4
Variety (afwisseling) -4 0 +4
generate new ideas to do things first
APPRECIATE - HIERARCHY CULTURE
Aesthetics, management (structuur) -4 0 +4
Income (financiële beloning) -4 0 +4
Lifestyle (balans werk en privé) -4 0 +4
Security (zekerheid) -4 0 +4
Work environment (arbeidsomstandigheden) -4 0 +4
discuss what is of value to do things right
MOBILISE - MARKET CULTURE
Achievement (prestaties) -4 0 +4
Prestige (prestige) -4 0 +4
Supervision (invloed) -4 0 +4
deploy and develop to get the job done
CONTROL
COMPETE

Fig. 3.4 Renewed STM diagram 2: the alignment of organisational culture and human talent, elaborated in the relationship between organisational climate and work values

The personality facets corresponding to these models emerged from the translation of each of the personality facets of the FFM (Costa and McCrae 1985) into sets of Dutch non-normative and work related synonyms and antonyms (called synsets), derived from the Dutch Idioticon of Personality (De Raad and Doddema-Winsemius 2006). This idioticon is a lexical matrix derived from the 1203 trait terms of Brokken (1978), which is used as a vocabulary to describe a person's nature.

The interconnections between the management cycle of organisational effectiveness and the lexical corresponding personality facets is grafted in the competing

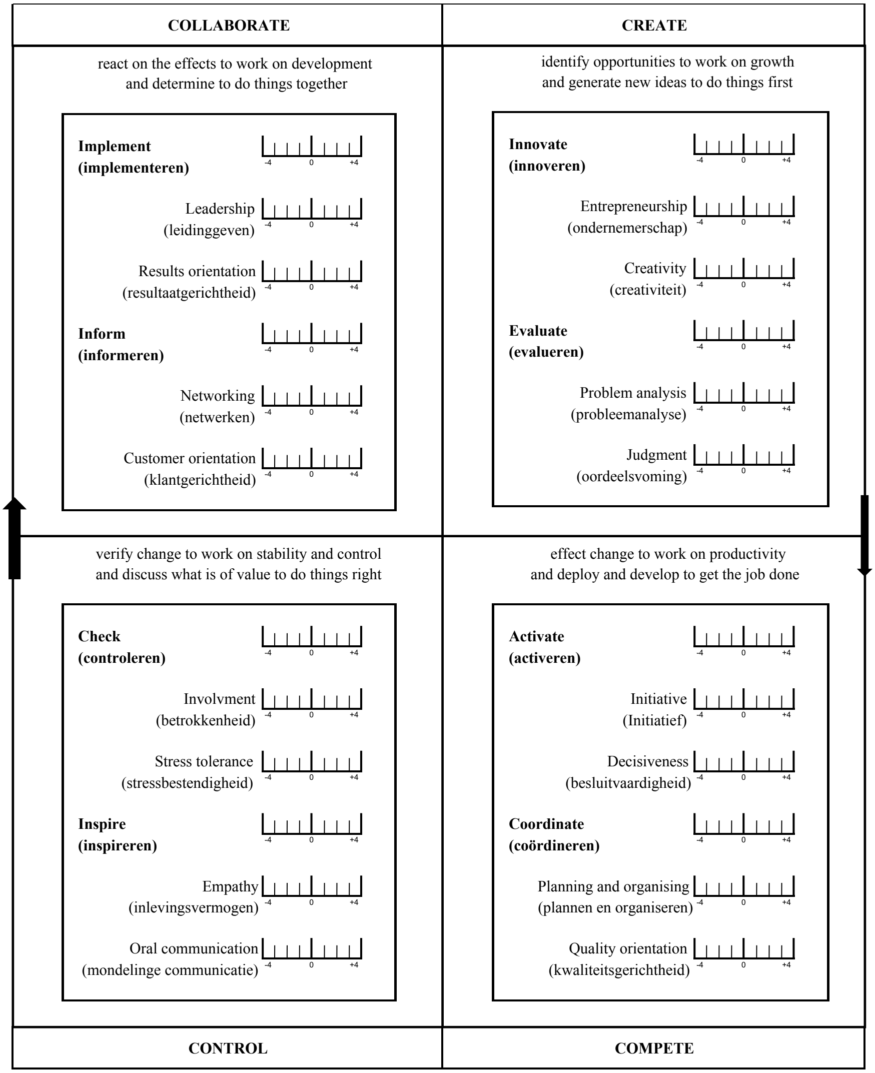

COLLABORATE
CREATE
react on the effects to work on development and determine to do things together
Implement (implementeren)
Leadership (leidinggeven)
Results orientation (resultaatgerichtheid)
Inform (informeren)
Networking (netwerken)
Customer orientation (klantgerichtheid)
identify opportunities to work on growth and generate new ideas to do things first
Innovate (innoveren)
Entrepreneurship (ondernemerschap)
Creativity (creativiteit)
Evaluate (evalueren)
Problem analysis (probleemanalyse)
Judgment (oordeelsvoming)
verify change to work on stability and control and discuss what is of value to do things right
Check (controleren)
Involvment (betrokkenheid)
Stress tolerance (stressbestendigheid)
Inspire (inspireren)
Empathy (inlevingsvermogen)
Oral communication (mondelinge communicatie)
effect change to work on productivity and deploy and develop to get the job done
Activate (activeren)
Initiative (Initiatief)
Decisiveness (besluitvaardigheid)
Coordinate (coördineren)
Planning and organising (plannen en organiseren)
Quality orientation (kwaliteitsgerichtheid)
-4
0
+4
CONTROL
COMPETE

Fig. 3.5 Renewed STM diagram 3: the alignment of business strategy and human talent, elaborated in team roles as the junction in the relationship between business strategy and competences

values leadership model, or CVLM (Cameron et al. 2014). The CVLM expresses the combination of the PDCA and the CVF in four verbs that represent human activity:

1. *create*, defined as 'doing new things' and considered as the junction of 'plan' and the open systems model;
2. *compete*, specified as 'doing things now' and perceived as the link between 'do' and the rational goal model;
3. *control*, determined as 'doing things right' and perceived as the junction of 'check' and the internal process model; and
4. *collaborate*, or 'doing things that last', considered as the link between 'act' and the human relations model.

With this, the first renewed diagram is a representation of the individual aptitude for the different phases of organisational effectiveness.

To convert this diagram into a computer based testing instrument, the sum score of each underlying five factor personality facet, measured with a five factor personality questionnaire such as the Dutch Personality test (NPT; Van Thiel 2008a), can be calculated as the sum of the raw scores on the set of corresponding items on a five-point Likert scale of that specific personality facet. The sum score then can be converted into a standardised Z-score on a bandwidth of -4 until $+4$, comparable to the range of -4σ until $+4\sigma$, defined as four standard deviations from the mean within a normal distributed sample. In this way it is possible to visualise the individual score between two opposite facets. For example, the raw score for the facet 'original', derived from the factor 'openness' and measured with a five factor test, is based on the responses on ten items with a raw sum score between 10 and 50. If a candidate has a raw score of 28 points, this results in a standardised Z-score of 2.23, which implies that the individual score bandwidth on the synset traditional versus original is [-1.77 until 2.23]. This means that the candidate has a somewhat higher aptitude for original ($2.23/4.00 = 0.56$) than for traditional ($-1.77/-4.00 = 0.44$).

The four CVLM models each consist of six synsets of two opposite personality facets, derived from the FFM. Figure 3.3 shows that the first model 'create' consists of five synsets of the factor openness and one synset of the factor neuroticism (well considered vs. intuitive). The second model 'compete' consists of two synsets of the factor extraversion (inert vs. active and calm vs. lively) and two synsets of the factor conscientiousness (disorganised vs. methodical and impulsively vs. circumspect). The last two synsets of 'compete' were derived from the factor agreeableness. The model 'control' consists of one synset of the factor agreeableness (uncompromising vs. sensitive) and of two synsets of the factor conscientiousness (improvised vs. disciplined and as usual vs. attentive). The other three synsets were derived from the factor neuroticism. 'Collaborate', the fourth model, consists of one synset derived from conscientiousness (dosed vs. diligent) and of two synsets from agreeableness (persistent vs. accommodating and competitive vs. cooperative). The other three synsets were derived from extraversion. On average, each of the four models is built on facets of three factors of the FFM.

To find relations between various aspects of organisational effectiveness and personality facets (Brouwer 2018, Chap. 2), a lexical-semantic analysis was performed.

Lexical analyses address a language's lexicon, or the collection of words in a language (Murphy 2003). It can be used to study word meanings, the relationships between (groups of) words and sentences, and to test semantic distances. Semantic distances are defined as the number of lexical steps required to relate the meaning and longer utterance of key terms. To calculate the semantic distances between personality facets and the four CVLM models, Open Wordnet was used, which is a lexical database that groups words together by joint meanings. The strongest relationships were found for 'collaborate', the weakest for 'compete' and 'control'. For all models, the resulting lexical-semantic distance was within four steps of Wordnet, which shows that four or less semantics were needed to link the models of organisational effectiveness to the corresponding personality facets. The ordering of personality facets in the four models was also confirmed by stepwise multiple linear regression analyses, predicting the work values from the personality facets (Brouwer and Veldkamp 2018).

3.3.2 Renewed STM Diagram 2: Aligning Organisational Culture and Human Talent

Figure 3.4 presents the renewed version of the second STM diagram. It visualises the elaboration of the systems-oriented interplay between the building blocks *culture* and *resources* of the MBBF in the relationship between organisational climate and work values. The building block *culture* becomes tangible in the construct organisational climate. Its function, seen as the motivational aspects behind the management cycle of organisational effectiveness, is elaborated in the IMAR-cycle (INK 2008). The IMAR (for inspire—mobilise—appreciate—reflect) is a method to design and control organisational climate from a 'level of excellence' perspective. The cycle is interpreted as follows:

1. *inspire*, which is the act of stimulating the mind and generating new ideas;
2. *mobilise*, or the act of deploying and developing the capabilities of all stakeholders in and around the organisation;
3. *appreciate*, or the act of discussing with stakeholders what is really of value; and
4. *reflect*, which is the act of discussing what matters, what will be possible or difficult to do, and what to do about anything that is decided on.

The impact of organisational climate on the building block *results* gets tangible in the four models of the organisational culture assessment instrument, or OCAI (Cameron and Quinn 2011). This quantitatively based organisational culture survey provides a framework for clarifying the underlying relationships between organisational climate and its effects on the performance of the organisation. The culture models consist of:

1. *adhocracy culture*, a culture that is dynamic and entrepreneurial, in which people concentrate on doing things first;
2. *market culture*, which is a results-oriented culture that focusses on getting the job done;

3. *hierarchy culture*, where the culture is structured and controlled, with the intention of doing things right; and
4. *family culture*, where the culture is characterised by mentoring and nurturing, with the aim of doing things together.

The corresponding work values emerged from the UVM. The interconnections between the management cycle of organisational climate and the lexical corresponding work values are also grafted in the four models of the CVLM. With this, the second renewed diagram is an illustration of the individual affinity with the different phases of organisational climate.

To convert this diagram into a computer based testing instrument, the sum score of each work value, measured with a universal values questionnaire (the Dutch work values test; NWT; Van Thiel 2008b), can be calculated as the sum of the raw scores on the set of corresponding items on a five-point Likert scale of that specific work value. The sum score can then be converted into the standardised Z-score on a bandwidth of −4 until +4, comparable to the range of −4σ until +4σ. This makes it possible to visualise the individual score on that specific work value. For example, the work value 'independence', measured with a universal values test, is built on eight items with a raw sum score between 8 and 40. If a candidate has a raw score of 23 points, this results in a standardised Z-score of 0.53. This implies that the individual score bandwidth on the work value independence runs from [−4.00 until 0.53], which means that the candidate has a slightly more than average affinity with independence (4.53/8.00 = 0.57).

The four CVLM models each consist of a set of work values, comparable to the clustering of work values found in earlier research (Robinson and Betz 2008; Van Thiel 2008b). A lexical-semantic analysis was performed to find relations between various aspects of organisational climate and work values (Brouwer 2018, Chap. 3). The strongest relationships were found for 'compete', the weakest for 'control'. For all models, the resulting lexical-semantic distance was between 3.2 and 4.4 steps of Wordnet. This ordering of work values in four models was also confirmed by the stepwise multiple linear regression analyses, predicting the work values from the personality facets (Brouwer and Veldkamp 2018).

3.3.3 Renewed STM Diagram 3: Aligning Business Strategy and Human Talent

Figure 3.5 presents the renewed version of the third STM diagram, showing the elaboration of the systems-oriented interplay between the building blocks s*trategy* and *resources* of the MBBF in team roles as the junction in the relationship between business strategy and competences. The building block s*trategy* is viewed as the joint process-oriented and human-contribution approach of both organisational effectiveness and organisational climate. The former, known as the process dimension of *strategy*, is elaborated in the combination of the PDCA-cycle and the IMAR-cycle.

The latter, defined as the contribution of *strategy* to the building block *resources*, becomes visible in the combination of the four models of the CVF and the OCAI.

The corresponding competences emerged from the combination of underlying personality facets of the FFM and work values of the UVM, representing the attribute- and attitude elements of the competence. The individual competences and the integral business strategy, are linked through the higher-order construct of team roles, that consists of both a business strategy and a competence element. As shown in the first two renewed STM diagrams (Figs. 3.3 and 3.4), the interconnections between the management cycle of business strategy and the lexical corresponding competences are grafted in the four models of the CVLM. With this, the third renewed STM diagram is an illustration of the individual contribution to the different phases of the business strategy.

To make the third diagram applicable as a testing instrument, the sum score of each competence can be calculated as the sum of the raw scores on the set of corresponding underlying personality facets, measured with a five factor personality questionnaire, and work values, measured with a universal values questionnaire, on a five-point Likert scale. This sum score is converted into the standardised Z-score on a bandwidth of -4 until $+4$, comparable to the range of -4σ until $+4\sigma$, defined as four standard deviations from the mean within a normal distributed sample. In this way, it is possible to visualise the individual score on that specific competence. The sum score on each team role is the average of the standardised sum score of the two underlying competences. For example, the competence creativity is built on the three personality facets reflective, original and ingenious and on the four work values mental challenge, creativity, independence and variety. This, for instance, results in an average standardised Z-score of 1.75, which implies that the individual score bandwidth of the candidate on the competence creativity runs from [−4.00 until 1.75]. This means that the candidate has an aptitude of $(5.75/8.00 = 0.72)$ for creativity. If, for example, the average standardised Z-score for the competence entrepreneurship was 0.52, then the candidate would have an aptitude of $(5.75 + 4.52)/2 = 5.14/8.00 = 0.64$ for the team role innovate.

The four CVLM models each consist of two team roles that are both built on two competences. Each competence is constructed from a set of underlying personality facets and work values. This classification of competences and team roles was confirmed by both the lexical-semantic analyses and the stepwise multiple linear regression analyses, predicting the work values with the personality facets (Brouwer 2018, Chap. 5).

First, the lexical-semantic relation between four CVLM models, 16 key competences and eight team roles was established. Then, data were collected to substantiate this relation. Factor analyses showed that the ordering of key competences could be clustered in the lexical corresponding four CVLM models. A second factor analysis of the eight team roles resulted in three clusters of team roles. It appeared that the first two models of CVLM, ‘create’ and ‘compete’, showed overlap. However, the comparable strength of the factor loadings of the team roles innovate and evaluate on the one hand and the team roles activate and coordinate on the other hand, seem to suggest that these are two different clusters. This implies that the team roles can also be divided into the four models.

For both the key competences and the team roles, strong reliability in terms of Cronbach's alpha (α) was found. The key competences show an average α of 0.798 within a range of [0.745–0.855]. The team roles present an average α of 0.811 within a range of [0.777–0.839]. According to both the Standards for Educational and Psychological testing (AERA, APA and NCME 2014) and the Dutch Committee on Tests and Testing (COTAN), a Cronbach's alpha of 0.8 or higher indicates a high reliability and is required for selection assessment tools (Evers et al. 2010). Therefore, both the composition of the key competences in underlying personality facets and work values, and the composition of the team roles out of the combination of two key competences, can be measured in a reliable way. A nuance should be made for the team roles, since the increase in the number of combined personality facets and work values behind the team role in itself contributes positively to the reliability coefficient.

3.4 Implications of the STM for Educational Measurement

STM is developed as an instrument to align human talent and business purpose. Psychological tests are used to measure human talent, and five managerial building blocks (Nieuwenhuis 2006) are applied to develop three renewed STM diagrams that connect human *resources* and *results* through interconnecting managerial building blocks. The logic that underlies the STM can be generalised to other applications. It could, for example, be applied in the educational world.

The purpose of education could be formulated as to have learners graduate with the skills needed to succeed. The results could be expressed in how well these skills are mastered. Dependent on the kind of education; primary, secondary or tertiary, different skills need to be obtained. Primary education prepares children for secondary education, secondary education prepares them for college, and college prepares students for the labour market. To operationalise these skills, learning goals have been formulated and curricula have been developed and implemented. For the Dutch educational system, for example, the learning goals that have to be mastered by the end of primary education are well documented for the various domains (Rijksoverheid 2006). The goals for the various levels of secondary education are specified in syllabi at course level (www.examenblad.nl). All of these learning goals are content oriented. With respect to language, for example, pupils are expected to be able to gain information from spoken language and to represent the information in a structured way, either orally or written, by the end of primary education. Recently, a discussion was started to extend this set of goals by incorporating 21st century skills.

Human talent in education is often expressed in terms of how well children perform with respect to educational tests. These tests could be developed by a teacher, teacher teams, publishers who developed the learning materials, or by specialized test publishers. The content of the tests is derived from the general learning goals.

Since these goals have to be formulated for the end of primary, secondary or tertiary education, they are very general. Therefore, they have been operationalised in learning lines, that specify when children or students have to be able to master sub-goals. These sub-goals can be translated into test goals. Finally, for these test goals, items can be developed that are administered to children or students in educational tests. The learning lines are generally content oriented, which is also reflected in the items. For this process, see also the 12-steps test development framework of Downing and Haladyna (2009).

One of the first things that can be noted when the STM approach is applied to the educational system, is that human talent is much broader defined than content related ability, which is the topic of most educational assessments. STM studies human talent by including personality factors, work values, and competences. These personality factors can be much broader than cognitive abilities. Besides, competences and work values could also contribute to educational outcomes.

In the STM model, personality factors are operationalized by the FFM. The first STM diagram models the relation between the facets of the FFM and organisational effectiveness. The third diagram models the relation between the facets and business strategy via competences and team roles. Even though the concepts of organisational effectiveness and business strategy might be hard to translate to the world of educational sciences, the more general building blocks of *structure* and *strategy* play an important role in education as well. *Structure* could be related to what pupils and students learn, *strategy* can be related to how they learn.

The relation between the FFM and academic performance is studied extensively (Poropat 2009). Academic performance was found to correlate significantly with agreeableness ($r = 0.07$), conscientiousness ($r = 0.19$), and openness ($r = 0.10$). Conscientiousness correlated the highest, almost as high as intelligence ($r = 0.23 - r = 0.56$). Even though a significant relation between personality and performance has been reported, personality factors are almost never taken into account in assessment in the educational system. Two reasons can be given. The first one is a statistical reason: the correlations are relatively small, even conscientiousness only explains four percent of the variance. Secondly, educational assessments primarily focus on whether cognitive goals have been reached. However, this might change in the near future. As found in the study of Komarraju et al. (2009) on the connection between personality and how students learn (strategy), the relation between personality and what students learn (structure) could become more important when education focuses on competences and 21st century skills as well, rather than content related learning goals only.

Within the STM, competences are related to business strategy via team roles. The concept of competences has been applied in the educational context as well. Mulder et al. (2007) provide an overview of various traditions that used the concept of competences. They distinguished between a behaviourist, general and cognitive approach, where the STM model seems to fit within the behaviouristic approach. Even though there has been a considerable tradition of using competences in education, especially in vocational education, there are still many educational programmes in which the learning goals have not been formulated in terms of competences. STM demonstrates how competences can be used to align talent and purpose. It therefore supports a wider use of competences in formulating purpose of learning as well.

The third personality characteristic that is distinguished in the STM is work values. The STM showed how work values are related to culture and contribute to results. Adding the building block *culture* to a model that relates human talent to educational outcomes is not very common in educational measurement. The most general factors that are taken into consideration are student characteristics, teacher characteristics, school characteristics and teaching characteristics. Culture is related to why and when students learn. Therefore, work values could be studied as student characteristics influencing learning results as well. However, in most studies, parental education, parental occupational status, income, and social economic status are taken into consideration instead. In order to develop 21st century skills, the what, the why and the when are drivers behind the how. Knowing when a student will be intrinsically motivated to learn (why) and knowing when a student will be most effective in his/her learning process (what), provides both teachers and school systems with insights how to optimise the learning results in terms of both cognitive and social-emotional skills that are needed to thrive in the modern work-environment.

Overall, when the STM is translated to the educational context, quite a few interesting observations can be made. The STM provides a more general framework for aligning talent and outcomes. Educational measurement could therefore be enriched by adopting this broader approach. Two applications might especially profit from such an approach. In the Netherlands, the transition from primary to secondary education is quite a hurdle for many children. Secondary education in the Netherlands distinguishes various levels and the problem is to assign each child to his/her appropriate level. Based on performance in primary education, the teacher in the final grade advises which level would suit a child best. After this advice, children participate in one of five national tests at the end of primary education, as a kind of objective second opinion. The results of these national tests can be used to adjust the advice to a higher level. The national tests mainly focus on performance in language and math. The STM showed that adding other personality factors, work values and competences might result in a more appropriate advice.

A second application is in the design of teaching. The STM shows that optimal alignment of talent and purposes leads to the best outcomes. The assessment of human talent is an important ingredient in this process. When translated to the educational context, this implies that educational measurement is a considerable ingredient in the alignment of talent and educational results. There is quite some debate about the role of educational measurement in the teaching process. Teachers often complain that too much time is spent on testing rather than teaching. Besides the learning goals that have been defined for primary and secondary education, schools tend to focus on the more general development of children as good citizens. Teaching to the test does have a very negative connotation in this debate (Onderwijsraad 2013). From the STM it can be learned, however, that efficient use of educational measurement in the educational process will lead to better outcomes. Aligning learning goals and educational measurement, as was proposed by Scheerens (2016), could further increase the outcomes of the educational processes.

3.5 Conclusion, Limitations, and Recommendations

3.5.1 Conclusion

This chapter introduced the design of the renewed evidence-based STM model and its elaboration in three renewed STM diagrams, that could take the place of the initial three diagrams as introduced in Brouwer (2012). The management building blocks framework (MBBF; Nieuwenhuis 2006) and the systems theory (Katz and Kahn 1966; Meadows 2008) were used to elaborate the interrelations within the new STM model. This provided in a renewed way of linking human talent to the core elements of the organisation's purpose (Barile 2006, 2008; Mele et al. 2010), and is assumed to create a fit at the level of a joint corporate and personal identity instead of at the level of a specific job profile that in our ever-changing world is subject to continuous alteration.

The first renewed STM diagram (Fig. 3.3) is seen as an improved and evidence-based rendition of the initial 2012 version of the first Dutch STM-scan diagram. In both versions, the 24 personality facets are lexically derived from the 24 FFM labels including their lexical antonym. Whereas in the initial diagram, the antonyms were directly derived from the Van Dale Thesaurus, a lexicon consisting of synonyms and antonyms, the antonyms in the renewed version were derived from the list of characteristics describing both poles of each of the five factor personality facets, as documented in the Dutch 'Idioticon of Personality' (De Raad and Doddema-Winsemius 2006). In the initial 2012 version, the 24 personality facets and their antonyms were clustered in the four steps of the primary business process that stem from the business purpose (idea—plan—form—action). In the renewed version of the first diagram, business purpose is further detailed in the content- and contribution side of organisational effectiveness, and is considered the effectuation of the building block *structure* found in the MBBF. The lexical-semantic linking of this block to the personality facets through the optimal path similarity results in the first path that runs from *resources* through *structure* to *results*.

The second renewed STM diagram (Fig. 3.4) is introduced as an improved and evidence-based version of the initial second 2012 STM-scan diagram. In both editions, the work values are lexically derived from the UVM and clustered in four higher-order culture types. In the initial version, this resulted in four culture types with four corresponding fundamental attitudes, representing an individual's social orientation. Within the renewed version, the work values were clustered in higher-order culture types, similar to the ordering of work values of Schwartz (1992), Ros et al. (1999), Robinson and Betz (2008), Daehlen (2008), and Van Thiel (2008b). Business purpose is further detailed in the content- and contribution side of organisational climate and is considered the effectuation of the building block *culture* found in the MBBF. The lexical-semantic linking of this block with the work values through the optimal path similarity leads to the second path that runs from *resources* through *culture* to *results*.

The third renewed STM diagram (Fig. 3.5) is presented as an improved and evidence-based version of the initial third STM-scan diagram. In the initial 2012 version, the 24 personality facets and their antonyms of the first STM diagram were used to calculate the amount of disposition for a series of 24 competences. Each competence was built on a combination of three of the 24 personality facets measured in the first STM diagram. The 24 competences were clustered in eight team roles, each consisting of three competences representing the four steps of the primary business process that stem from the business purpose. Within the renewed third STM diagram, combinations of both personality facets (first diagram) and work values (second diagram) are used to calculate the amount of disposition for a series of 16 key competences. The algorithms of combinations of personality facets and work values per key competence are confirmed by both the lexical-semantic classification, found in Brouwer (2018, Chaps. 2 and 3), and the linear regression models found in Brouwer and Veldkamp (2018). Each of the 16 key competences is related to a series of lexical-semantic synonyms found in existing competence frameworks in which the different competences are clustered in: (1) strategic competences, (2) tactical competences, (3) communicating competences, and (4) operational competences. The 16 key competences are clustered in eight team roles, each consisting of two key competences.

In the new version of the third diagram, business purpose is further detailed in the process-oriented and human-contribution approach of business strategy. The process-oriented approach is dealt with as the content side of both organisational effectiveness and organisational climate. The human-contribution approach is seen as the contribution side of both organisational effectiveness and organisational climate. This operationalisation of the construct business strategy is considered the effectuation of the building block *strategy* found in the MBBF. The renewed versions of the STM diagrams contribute in more detail to the alignment of the organisation and its employee.

All in all, the design and validation of the renewed STM contributes to a sharp and objective picture of the match between people and the organisation, by linking human characteristics to managerial building blocks. The three renewed STM diagrams jointly make up a potential new version of the systems-oriented assessment instrument STM-scan, both in Dutch and in English. In completing a five factor personality test, such as the NPT and an universal values inventory, like the NWT, the individual contribution to the four models of organisational effectiveness, organisational climate and business strategy can be measured and reported in the three renewed STM diagrams that jointly represent three alignment paths between *resources*, the organisation and its intended *results*.

3.5.2 Application to Educational Measurement

In order to modernise education, it is required to measure the what, why, when and how of the learning process. It is not only important to know what a student should learn in order to participate in the working environment, but also when and under which circumstances he or she is motivated to learn. This will provide earlier and better insight in how every individual reaches maximal learning results in his or her own way. This specifically concerns the results that consist of both cognitive and social-emotional skills, which will prepare students for the sustainable employability our rapidly changing world asks for. Therefore, it is necessary to develop skills that help to use acquired knowledge in different ways and situations. This will contribute to the prevention of school- and work dropout and the entailed costs.

3.5.3 Limitations

Before turning to the recommendations and implications of this study, there are some limitations to take into account. The renewed evidence-based systems-oriented talent management model is built on three different paths between the building blocks *resources* and *results*, found in the MBBF. Therefore, the intermediate building blocks *structure*, *culture* and *strategy* are theoretically linked to organisational effectiveness, organisational climate and business strategy. The building block *resources* is detailed in personality facets, work values, competences and team roles. The different relationships, found in the three paths between *resources* and *results* are partly established on the basis of interpreting different text corpuses. This could imply that other existing lexical-semantic relations, that might argue against the present used relations, may have been overlooked. However, since the majority of the lexical-semantic relationships are empirically substantiated by stepwise multiple regression analyses, confirmatory factor analyses and multitrait multimethod matrixes, this supports the internal consistency reliability and construct validity of the renewed STM diagrams.

A second limitation of this study is that the renewed STM diagrams so far have not yet been operationalised in a new version of the web-based STM-scan assessment instrument. Therefore, the utility of the renewed model could not yet be evaluated based on practical experience.

A third limitation concerns the definition of competences. The educational literature shows that different competency models were developed in different traditions (Mulder et al. 2007). This is also the case for organisational practice, where large companies often create their own competency-language. In the development of STM, the overlap between a series of those different models was studied. This resulted in a set of 16 key competences, with underlying synonyms. This makes the STM an instrument with which customer-specific competency-languages can be converted to the 16 key competences. In this way it is possible to measure different competence models in the same way. However, in practice this turns out to be a difficult task, that asks for a thorough understanding and knowledge of the definitions for and distinctions between the different competences.

3.5.4 Recommendations and Implications

Effective test development requires a systematic, well-organised approach to ensure sufficient validity evidence to support the proposed inferences from the test scores. Therefore, it is recommended to use the 12-steps test development framework of Downing and Haladyna (2009), in order to implement the renewed STM diagrams in practice. Part of this development process should be the training of STM experts and the implementation of a periodic evaluation cycle of the psychometric quality and utility of the renewed assessment instrument.

Another recommendation concerns the starting point of the measurement. In its current composition, the renewed STM model is filled with the test results of an individual's scores on a set of human characteristics, which builds the relationship with the management building blocks of the MBBF. In addition, the field of management science has introduced different inventories for measuring a combination of these management building blocks (Cameron and Quinn 2011; Cameron et al. 2014). An interesting follow-up study would be to investigate whether the STM model can also be measured the other way around by predicting human characteristics with the help of a managerial inventory. The present study lays the foundation for this follow-up research.

In conclusion, the renewed evidence-based STM model and diagrams introduced in this chapter, contribute to the future bridging of the gap between psychological questionnaires for testing human characteristics and models for unravelling managerial building blocks. In doing so, insight is provided in how adaptive enterprises ought to be organised these days and how to give shape to the corresponding upscaling that is required of their talent management experts. Subsequently, the study joined the debate on how the educational field ought to implement a wider approach on learning outcomes, by studying the structure (what), culture (why and when) and the strategy (how) sides of educating and measuring 21st century skills next to cognitive skills.

References

American Educational Research Association AERA, American Psychological Association APA, National Council on Measurement in Education NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Barile, S. (2006). *The company as a system. Contributions on the viable system approach, part I*. Torino: Giappichelli.

Barile, S. (2008). *The company as a system. Contributions on the viable system approach, part II*. Torino: Giappichelli.

Belbin, M. R. (2010). *Management teams: Why they succeed or fail, 3rd revised ed*. London: Taylor & Francis Ltd.

Berger, A., & Berger, D. R. (2011). *The talent management handbook* (2nd ed.). New York: McGraw-Hill.

Brokken, F. B. (1978). *The language of personality,* unpublished thesis. University of Groningen.

Brouwer, A. J. (2012). *STM-scan, handboek voor het systeemgericht managen van talent*. Vaassen, Nederland: A.J. Brouwer.

Brouwer, A. J. (2018). *Systems-oriented talent management. A design and validation study* (Doctoral dissertation). Retrieved from https://doi.org/10.3990/1.9789036546843.

Brouwer, A. J. & Veldkamp, B. P. (2018). How age affects the relation between personality facets and work values of business and private bankers. *Journal of Work and Organizational Psychology*. Advance online publication https://doi.org/10.5093/jwop2018a20.

Cable, M., & Yu, K. Y. T. (2007). How selection and recruitment practices develop the beliefs used to assess fit. In C. Ostroff & T. A. Judge (Eds.), *Perspectives on organizational fit* (pp. 155–182). New York: Taylor & Francis-Group LLC.

Cameron, K. S., & Quinn, R. E. (2011). *Diagnosing and changing organizational culture: Based on the competing values framework* (3rd ed.). San Francisco: Jossey-Bass.

Cameron, K., Quinn, R. E., & Degraff, J. (2014). *Competing values leadership* (2nd ed.). Camberley, Surrey, UK: Edward Elgar.

Capra, F. (1997). *The web of life*. New York: Doubleday-Anchor Book.

Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO personality inventory manual*. Odessa, FL: Psychological Assessment Resources.

Daehlen, M. (2008). Job satisfaction and job values among beginning nurses: A questionnaire survey. *International Journal of Nursing Studies, 45,* 1789–1799.

De Raad, B., & Doddema-Winsemius, M. (2006). *De Big 5 persoonlijkheidsfactoren: Een methode voor het beschrijven van persoonlijkheidseigenschappen*. Amsterdam: Uitgeverij Nieuwezijds.

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Downing, S. M., & Haladyna, T. M. (2009). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). New York: Routledge.

Ehrhart, M. G., Schneider, B., & Macey, W. H. (2014). *Organizational climate and culture: An introduction to theory, research, and practice*. London: Taylor & Francis.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests*. Amsterdam: NIP/COTAN.

Galbraith, J. R. (2002). *Designing organizations: An executive guide to strategy, structure and process*. New York: Wiley.

Gimenez-Espin, J. A., Jiménez-Jiménez, D., & Martínez-Costa, M. (2013). Organizational culture for total quality management. *Total Quality Management & Business Excellence, 24*(5–6), 678–692.

Hendricks, K., & Singhal, V. (1996). Quality awards and the market value of the firm: An empirical investigation. *Management Science, 42*(3), 415–436.

Instituut Nederlandse Kwaliteit (INK). (2008). *Introductie, inhoud en toepassing van het INK managementmodel.* Zaltbommel, Nederland: INK projectgroep 'vernieuwing INK'.

Jacobides, M. G. (2007). The inherent limits of organizational structure and the unfulfilled role of hierarchy: Lessons from a near-war. *Organizational Science, 18*(3), 455–477.

Katz, D., & Kahn, R. L. (1966). *The social psychology of organizations*. New York: Wiley.

Komarraju, M., Karau, S. J., & Schmeck, R. R. (2009). Role of the Big Five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences, 19*(1), 47–52.

Kuenzi, M., & Schminke, M. (2009). Assembling fragments into a lens: A review, critique, and proposed research agenda for the organizational work climate literature. *Journal of Management, 35*(3), 634–717.

McDonnell, A., & Collings, D. G. (2011). The identification and evaluation of talent in MNE's. In H. Scullion & D. G. Collings (Eds.), *Global talent management*. New York: Routledge.

Meadows, D. H. (2008). *Thinking in systems: A primer*. London: Chelsea Green Publishing.

Mele, C., Pels, J., & Polese, F. (2010). A brief review of systems theory and their managerial applications. *Service Science, 2*(1), 126–135.

Michaels, E., Handfield-Jones, H., & Axelrood, B. (2001). *The war for talent*. Boston: Harvard Business School.

Mitchell, G. E. (2012). The construct of organizational effectiveness: Perspectives from leaders of international nonprofits in the United States. *Nonprofit and Voluntary Sector Quarterly, 42*(2), 324–345.

Mulder, M., Weigel, T., & Collins, K. (2007). The concept of competence in the development of vocational education and training in selected EU member states: A critical analysis. *Journal of Vocational Education & Training, 59*(1), 67–88.

Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge, UK: Cambridge University Press.

Nag, R., Hambrick, D. C., & Chen, M. J. (2007). What is strategic management, really? Inductive derivation of a consensus definition of the field. *Strategic Management Journal, 28*(9), 935–955.

Nieuwenhuis, M. A. (2006). *The art of management*. Retrieved on January 15, 2014 from http://www.the-art.nl.

Onderwijsraad. (2013). *Een smalle kijk op onderwijskwaliteit* [In Dutch], Onderwijsraad.

Pedraza, J. M. (2014). *What is organisational effectiveness and how an organization could achieve it*. Retrieved December 1, 2017, from https://www.researchgate.net/post/What_is_organisational_effectiveness_How_an_organisation_could_achieve_it.

Peters, T. J., & Waterman, R. H. (1998). *In search of excellence, lessons from America's best run companies*. New York: Warner Books.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322.

Pugh, D. S. (1990). *Organization theory: Selected readings*. Harmondsworth: Penguin.

Quinn, R. E., & Rohrbaugh, J. (1983). A spatial model of effectiveness criteria: Towards a competing values approach to organisational analysis. *Management Sciences, 29*(3), 363–377.

Ravasi, D., & Schultz, M. (2006). Responding to organizational identity threats: Exploring the role of organizational culture. *Academy of Management Journal, 49*(3), 433–458.

Rijksoverheid. (2006). Kerndoelenboekje [learning goals primary education]. Retrieved from: www.rijksoverheid.nl/documenten/rapporten/2006/04/28/kerndoelenboekje.

Robinson, C. H., & Betz, N. E. (2008). A psychometric evaluation of super's work values inventory-revised. *Journal of Career Assessment, 16*(4), 456–473.

Ros, M., Schwartz, S. H., & Surkiss, S. (1999). Basic individual values, work values, and the meaning of work. *Applied Psychology: An International Review, 48*(1), 49–71.

Scheerens, J. (Ed.). (2016). *Opportunity to learn, curriculum alignment and test preparation: A research review*. Berlin: Springer.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmidt, F. L., Oh, I. S., & Shaffer, J. A. (2016). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research findings.* Working paper ResearchGate, 1–73.

Schneider, B., & Reichers, A. E. (1983). On the etiology of climates. *Personnel Psychology, 36,* 19–39.

Schoonman, W. (2013). *Mensen beoordelen, voor HR professionals*. Amsterdam: De Witte Ridders.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology, 25,* 1–65.

Smith, M., & Robertson, I.T. (1986). *The theory and practice of systematic staff selection*. London: Macmillan Press

Tillema, K., & Markerink, F. (2006). *Gericht presteren met het INK-managementmodel, van visie naar actie*. Deventer: Kluwer.

Van Beirendonck, L. (2010). *Iedereen content, nieuwe paradigma's voor competentie- en talentmanagement*. Schiedam, Nederland: Uitgeverij Lamnoo.

Van Thiel, E. (2008a). *Handleiding Nederlandse Persoonlijkheidstest*, Nijmegen Nederland: 123test B.V.

Van Thiel, E. (2008b). *Handleiding Nederlandse Werkwaardentest*, Nijmegen, Nederland: 123test B.V.

Wit, B., & Meyer, R. (2011). *Strategy Synthesis, resolving strategy paradoxes to create competitive advantage*. Hamsphire: Cengage Learning EMEA.

Yu, T., & Wu, N. (2009). A review study on the competing values framework. *International Journal of Business and Management, 4,* 37–42.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

cc BY

Chapter 4
Assessing Computer-Based Assessments

Bas Hemker, Cor Sluijter and Piet Sanders

Abstract Quality assurance systems for psychological and educational tests have been available for a long time. The original focus of most of these systems, be it standards, guidelines, or formal reviewing systems, was on psychological testing. As a result, these systems are not optimally suited to evaluate the quality of educational tests, especially exams. In this chapter, a formal generic reviewing system is presented that is specifically tailored to this purpose: the RCEC review system. After an introduction with an overview of some important standards, guidelines, and review systems, and their common backgrounds, the RCEC review system for the evaluation of educational tests and exams is described. The underlying principles and background of this review system are explained, as well as the reviewing procedure with its six criteria. Next, the system is applied to review the quality of a computer-based adaptive test: Cito's Math Entrance Test for Teachers Colleges. This is done to illustrate how the system operates in practice. The chapter ends with a discussion of the benefits and drawbacks of the RCEC review system.

4.1 Introduction

Quality assurance systems for psychological and educational tests have been available for a long time. These systems have their origins in the need to serve the public interest. They provide professional users with information to determine whether these instruments are suitable for the user's purpose. Quality assurance systems come in different forms. A common differentiation is between codes, guidelines, standards, and review systems (e.g., Roorda 2007). Codes are a cohesive set of behavioral rules with which test authors are expected to comply in order to make good and fair tests. As such, they are different from the other three as they do not reflect on the test

B. Hemker (✉) · C. Sluijter
Cito, Arnhem, The Netherlands
e-mail: bas.hemker@cito.nl

P. Sanders
RCEC, Vaassen, The Netherlands

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_4

itself. Guidelines are intended to show how a test should be developed. Standards are slightly different as they describe a level of quality that should be attained by a test on the aspects deemed relevant. Review systems critically evaluate a psychological or educational test in order to make it possible to decide whether or not it has sufficient fit to purpose.

The first document containing a set of systematic evaluations of tests was the 1938 Mental Measurements Yearbook (Buros 1938). This volume contained a set of critical reviews of various psychological tests, questionnaires, and rating scales then in use. It was intended to assist professionals to select and use the most appropriate psychological test for their specific problem. Spanning a period of almost eight decades, its twentieth edition was published in 2017 (Carlson et al. 2017). Nowadays, the system used to review all instruments in the Mental Measurements Yearbook is accompanied by a profusion of other quality assurance systems.

Well-known guidelines on the development and use of psychological tests were developed and are maintained by the International Test Commission (ITC). This is an association of national associations of psychologists, test committees, and other organizations and individuals promoting the proper development and use of tests. The ITC came into existence in the 1970s (see, Oakland et al. 2001). The ITC now oversees six different guidelines, including guidelines for test use (ITC 2013), test security (ITC 2014), and computer-based and internet testing (ITC 2005).

The best known standards to date are the Standards for Educational and Psychological Testing. These standards have been jointly published in different editions since 1966 by three institutes: the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The most recent publication dates from 2014 (APA et al. 2014). The different editions were preceded by separate documents called “Technical recommendations for psychological tests and diagnostic techniques” (APA 1954) and “Technical recommendations for achievement tests” (AERA and NCME 1955). These publications primarily addressed the developmental process of tests as well as the type of information publishers should make available to test users in manuals (Camara and Lane 2006).

The European Federation of Psychologists’ Associations (EFPA) developed and maintains a review system for the description and evaluation of psychological tests. Its development started in 1989, but the first formal version was published much later (Bartram 2002). The most recent version was issued in 2013 (EFPA 2013). The current system was greatly influenced by criteria used earlier by the British Psychological Society and the Dutch Committee on Testing (COTAN) of the Dutch Association of Psychologists. This latter body, which was founded in 1959, itself has a long tradition in evaluating tests. COTAN started publishing test reviews in 1969 (Nederlands Instituut van Psychologen 1969). The most recent revision of the system was published almost a decade ago (Evers et al. 2010a, 2010b). In 2019 COTAN initiated work on a new revision. For a detailed overview of the history of the introduction of review models for psychological tests in Europe, the interested reader is referred to Evers (2012).

An analysis of the content of these and other current quality assurance systems, for instance, those of the Educational Testing Service (2014), the Association for Educational Assessment Europe (2012), and Cambridge Assessment (2017), demonstrates that all systems show substantial overlap. This is not surprising, because they all reflect in one way or another the theoretical and practical developments in psychological and educational measurement from the first half of the last century up until now, as captured in, for instance, the consecutive editions of Educational Measurement (Brennan 2006; Lindquist 1951; Linn 1989; Thorndike 1971). Unfortunately, all these systems also have a basic flaw that comes forth from their origins. Because they all had psychological tests as their original focus, the evaluation of educational tests, and especially exams as a specific subset, raises several issues

An exam can be defined as a formal investigation by a licensed body into the knowledge base, abilities, attitudes, skills and/or competencies of candidates. In order to receive a diploma or a certificate, candidates have to demonstrate a certain level of mastery on a set of assignments that are representative of the total domain of assignments for (certain parts of) the corresponding curriculum. This definition points out a generic difference between psychological and educational tests: the construct that is intended to be measured.

Educational tests are more often than not direct measures of human behavior, such as mathematical ability, reading comprehension, or spelling ability, while the constructs that are the subject of psychological tests have a more theoretical character like intelligence or neuroticism. This has as a direct consequence that criteria having to do with this aspect of test validity have a different orientation and weight for educational and psychological tests.

Another difference is that for most psychological constructs, some sort of stability over time is assumed, whereas the constructs measured in education are subject to change. While students are learning, they are expected to change over time in their ability—hopefully increasing after more education. On the other hand, the ability may also decrease, for example, due to a lack of practice, or simply forgetting things that once were learned. What is being measured is often a snapshot in time. The temporality is also reflected by the single use of many exams: they are used once, at one moment in time, never to be used again for the same purpose.

A lack of stability is also reflected in the constructs themselves as they can change over time as well: what was considered relevant in mathematics can change over time and over levels of ability. On the other hand, even if educational tests and exams change over time, regulators want to compare results over time. This means that in comparison with psychological tests, equating procedures to maintain stable norms are especially important. All this has consequences for the way tests and exams are reviewed.

4.2 The RCEC Review System for the Evaluation of Computer-Based Tests

The Research Center for Examinations and Certification (RCEC) has developed an analytical review system that is specifically tailored to evaluating the quality of educational tests, and particularly exams (Sanders et al. 2016). It was in large part inspired by the aforementioned COTAN review system. An overview of the principles and background of the RCEC review system is presented below, including the description of the six criteria the system uses.

The RCEC review system has three main characteristics in common with other review systems such as the EFPA system and the COTAN system. First, it focusses on the intrinsic quality of the instrument itself and not on the process of test development. It evaluates the quality of items and tests, but not the way they are produced. Of course, the scientific underpinning of the test reveals much about the process of test development, but this is only reviewed in the light of the impact of the process on the quality of the items and the test as a whole. Secondly, the review system works with a set of criteria with which an educational test should comply in order to be considered to be of sufficient quality. The third characteristic is that the review is completely analytical, or can even be considered to be actuarial. For each criterion, the reviewer answers a series of questions by giving a rating on a three-point scale: insufficient—sufficient—good. The review system contains clarifications of the questions. Instructions are provided to ensure that the scores ensuing from these questions are as objective as possible. For each of the six criteria the system applies, the ratings are combined through specific rules that yield a final assessment of the quality of each criterion, again on this three-point scale.

The nature of the selected criteria and their specific description is where the RCEC review system differs from others. Other systems, like the EFPA system and originally the COTAN system as well, focus more on psychological tests. As already mentioned, other criteria apply, or have a different weight when it comes to educational tests and especially exams. The RCEC review system consequently differs from other systems in the wording of the criteria, the underlying questions, their clarifications, the instructions, and the scoring rules. This is done in order to have the best fit for purpose, i.e., the evaluation of educational assessment instruments and exams in particular.

The reviewing procedure shares some features with other reviewing systems. Like the Buros, EFPA, and COTAN systems, two reviewers evaluate an educational test or exam independently. The reviewers are non-anonymous. Only professionals who are certified by RCEC after having completed a training in using the review system are allowed to use it for certification purposes. Note that all three authors of this chapter have this certificate. All cases are reviewed by the overseeing Board. They formulate the final verdict based on the advice of the reviewers.

The criteria of the RCEC system are:

- Purpose and use;
- Quality of test and examination material;

- Representativeness;
- Reliability;
- Standard setting, norms, and equating;
- Administration and security.

There is overlap with the criteria of other quality assurance systems. ‘Purpose and use’, ‘Quality of test and examination material’, ‘Reliability’, and ‘Administration and security’ can also be found in other systems. The most notable difference between the RCEC review system and other systems rests in the criterion of ‘Representativeness’ which corresponds with what other systems refer to as (construct and criterion-related) validity, but uses a different approach, especially for reviewing exams. Since these are direct measures of behavior rather than measures of constructs, the focus of this criterion is on exam content. Another difference is that within the criterion of ‘Standard setting, norms, and equating’, more attention is given to the way comparability over parallel instruments is ensured. It details how equivalent standards are being set and maintained for different test or exam versions.

Below, the criterion ‘Purpose and use’ is discussed in detail. This criterion is emphasized, because it is often taken for granted. Its importance cannot be overstated, as in order to produce a quality educational test or exam, it is vital that its purpose is well-defined. For the other five criteria, a shorter overview is given. Similar to the first criterion, these criteria are also found in other review systems. In this overview, special attention is given to the criteria as applied to computerized tests. This is done because the application of the review system is demonstrated by the evaluation of the quality of a computerized adaptive test (CAT).

A detailed description of the whole RCEC review system can be found at www.rcec.nl. Currently, the review system is only available in Dutch. An English version is planned.

4.2.1 Purpose and Use of the Educational Test or Exam

The golden rule is that a good educational test should have one purpose and use only. The exception to this is a situation where different purposes are aligned. For instance, a formative test can help in making simultaneously decisions on an individual, group, or school level, simultaneously. Discordant purposes and uses (e.g., teacher evaluation versus formative student evaluation) should not be pursued with one and the same educational test. This would lead to unintended negative side effects. In most cases, the purpose of educational tests and exams is to assess whether candidates have enough knowledge, skills, or the right attitudes. The use of an educational test concerns the decisions that are made based on the score obtained.

There are three questions used to score a test on this criterion:

- Question 1.1: Is the target population specified?
- Question 1.2: Is the measurement purpose specified?
- Question 1.3: Is the measurement use specified?

Question 1.1. has to do with the level of detail in the description of the test or exam target groups(s). Age, profession, required prior knowledge, and the level of education can also be used to define the target group. Without this information, the evaluation of the language used in the instructions, the items, the norm, or cut scores of the test becomes troublesome. Question 1.1 relates to who is tested and when. A test or exam gets a rating ‘Insufficient’ (and a score of 1) for this question when the target group is not described at all, or not thoroughly enough. This rating is also obtained when the educational program of studies for the target group is not described. A test gets a rating ‘Sufficient’ (a score of 2) only when the educational program the test is being used for is stated. It receives a rating ‘Good’ (a score of 3) for this question if not only the educational program but also other relevant information about the candidates is reported. This detailed information includes instructions on the application of the test to special groups, such as students having problems with sight or hearing.

An educational test should assess what candidates master after having received training or instruction. This is what question 1.2 refers to. What candidates are supposed to master can be specified as mastery of a construct (e.g., reading skill); of one of the subjects in a high school curriculum (e.g., mathematics); of a (component of a) professional job; or of a competency (e.g., analytical skills in a certain domain). A test that measures a construct or a competency needs to present a detailed description with examples of the theory on which the construct or competency is based. This implies that tautological descriptions like ‘this test measures the construct reading skills’ do not suffice. The construct or competency has to be described in detail and/or references to underlying documents have to be presented. The relevance of the content of the test or exam for its intended purpose should be clarified. A blueprint of the test can be a useful tool in this regard. A rating ‘Insufficient’ is given when the measurement purpose of the test is not reported. A rating ‘Sufficient’ is given when the purpose is reported. A rating ‘Good’ is given when in addition to this, a (detailed) description of constructs, competencies, or exam components is supplied as described above.

Educational tests or exams can be used in many ways. Each use refers to the type of decision that is being made based on the results of the test(s) and the impact on the candidate. Common uses are selection or admittance (acceptance or refusal), classification (different study programs resulting in different certificates or degrees), placement (different curricula that will result in the same certificate or degree), certification (candidates do or do not master a certain professional set of skills), or monitoring (assessment of the progress of the candidates). Question 1.3. is dichotomously scored: either the use of the test is reported in enough detail (‘Good’), or it is not (‘Insufficient’).

The overall evaluation of the description of the purpose and use of the test is based on the combination of scores on the three questions. The definite qualification for this criterion is ‘Good’ if a test receives a score of 3 on all three questions, or if two questions have a score 3 while the third one a score of 2. If Question 1.3 is scored 3 and the other two are scored 2, the qualification ‘Sufficient’ is given. Finally, the qualification is ‘Insufficient’ if one of the three questions was awarded a score of 1. This means that all three items are knock-out questions.

4.2.2 Quality of Test Material

All test material (manual, instructions, design, and format of items, layout of the test, etc.) must have the required quality. The items and the scoring procedures (keys, marking scheme) should be well defined and described in enough detail. The same holds for the conditions under which the test is to be administered.

The following key questions are considered:

- Question 2.1: Are the questions standardized?
- Question 2.2: Is an objective scoring system being used?
- Question 2.3: Is incorrect use of the test prevented?
- Question 2.4: Are the instructions for the candidate complete and clear?
- Question 2.5: Are the items correctly formulated?
- Question 2.6: What is the quality of the design of the test?

The first two questions are knock-out questions. If on either one of the two, a score of 1 is given, the criterion is rated ‘Insufficient’ for the test.

The RCEC review system makes a distinction between paper-and-pencil tests and computer-based tests. Some remarks on the application of the system for a CAT can be made. First, the next item in a CAT should be presented swiftly after the response to the previous item(s). In evaluating a CAT, Question 2.2 implies that there should be an automated scoring procedure. Secondly, Question 2.3 implies that software for a CAT should be developed such that incorrect use can be prevented. As the routing of the students through the test depends on previously given answers, going back to an earlier item and changing the response poses a problem in a CAT. Finally, Question 2.6 refers to the user interface of the computerized test.

4.2.3 Representativeness

Representativeness relates to the content and the difficulty of the test or exam. This criterion basically refers to the content validity of the test: do the items or does the test as a whole reflect the construct that is defined in Question 1.2. The key question here is whether the test (i.e., the items it contains) is actually measuring the knowledge, ability, or skills it is intended to measure. This can be verified by the relationship between the items and the construct, namely, the content. This criterion is evaluated through two knock-out questions:

- Question 3.1: Is the blueprint, test program, competency profile, or the operationalization of the construct an adequate representation of the measurement purpose?

- Question 3.2: Is the difficulty of the items adjusted to the target group?

Note that this criterion has a structurally different approach compared to corresponding criteria from review systems with their focus on psychological tests. Question 3.1 specifically refers to the content of a test or exam: it should be based on what a candidate has been taught, i.e., learning objectives. As these learning objectives often are not specific enough on which to base the construction of a test, classification schemes, or taxonomies of human behavior are used to transform the intended learning objectives to objectives that can be tested. Since educational tests, and especially exams are generally direct measures of behavior rather than measures of constructs, priority is given here to the content of the test or exam. In a CAT this also means that extra constraints have to hold to assure that candidates get the appropriate number of items for each relevant subdomain.

Question 3.2 asks whether the difficulty of the items, and thus the difficulty of the test or exam, has to be adjusted to the target group. In practice, this means that a test should not be too difficult or too easy. Particularly in a CAT, where the difficulty of the question presented is targeted to the individual taking the test, this should be no problem. The only issue here is that there should be enough questions for each level of difficulty.

4.2.4 Reliability

The previous two earlier criteria focus mainly on the quality of the test items. The evaluation of reliability involves the test as a whole. It refers to the confidence one can have in the scores obtained by the candidates. Reliability of a test can be quantified with a (local) reliability coefficient, the standard error of measurement, or the proportion of misclassifications. The first of the three questions is a knock-out question:

- Question 4.1: Is information on the reliability of the test provided?
- Question 4.2: Is the reliability of the test correctly calculated?
- Question 4.3: Is the reliability sufficient, considering the decisions that have to be based on the test.

In the case of a CAT, traditional measures for reliability do not apply. A CAT focusses on minimizing the standard error of measurement by following an algorithm that sequentially selects items that maximize the statistical information on the ability of the candidate, taking into consideration a set of constraints. The information function drives the selection of items, and the evaluation of the standard error of measurement is one of the important criteria to stop or to continue testing. Thus, without a positive answer on question 4.1, a CAT is not possible. Question 4.3 can be interpreted in a CAT by checking whether the stopping rule is appropriate given the purpose and use of the test, and whether there are sufficient items to achieve this goal.

4.2.5 Standard Setting and Standard Maintenance

This criterion reviews the procedures used to determine the norms of a test, as well as how the norms of comparable or parallel tests of exams are maintained. Norms can be either relative or absolute. If the norms were previously determined but need to be transferred to other tests or exams, equivalence and equating procedures need to be of sufficient quality. There are separate questions for tests or exams with absolute or relative norms.
Questions for tests with absolute norms:

- Question 5.1: Is a (performance) standard provided?
- Question 5.2a: Is the standard-setting procedure correctly performed?
- Question 5.2b: Are the standard-setting specialists properly selected and trained?
- Question 5.2c: Is there sufficient agreement among the specialists?

Questions for tests with relative norms:

- Question 5.3: Is the quality of the norms sufficient?
- Question 5.3a: Is the norm group large enough?
- Question 5.3b: Is the norm group representative?
- Question 5.4: Are the meaning and the limitations of the norm scale made clear to the user and is the norm scale in accordance with the purpose of the test?
- Question 5.5a: Is the mean and standard deviation of the score distribution provided?
- Question 5.5b: Is information on the accuracy of the test and the corresponding intervals (standard error of measurement, standard error of estimation, test information) provided?

Questions for maintaining standards or norms:

- Question 5.6: Are standards or norms maintained?
- Question 5.6a.: Is the method for maintaining standards or norms correctly applied?

A CAT can have absolute or relative norms, depending on the purpose and use of the test. However, for a CAT, the evaluation of the way the standards or norms are maintained most definitely needs to be answered, as each individual candidate gets his or her unique test. It is mandatory that the results from these different tests are comparable in order to make fair decisions. In CAT, this equating is done through item response theory (IRT). Question 5.6a relates to whether IRT procedures have been applied correctly in the CAT that is being reviewed.

4.2.6 Test Administration and Security

Information on how to administer the test or exam and how to assure a secure administration should be available for the proctor. The key concern is whether the design of the test is described in such a way that, in practice, testing can take place under standardized conditions, and whether enough measures are taken to prevent fraud. The questions for this criterion are:

- Question 6.1: Is sufficient information on the administration of the test available for the proctor?
- Question 6.1a: Is the information for the proctor complete and clear?
- Question 6.1b: Is information on the degree of expertise required to administer the test available?
- Question 6.2: Is the test sufficiently secured?
- Question 6.3: Is information on the installation of the computer software provided?
- Question 6.4: Is information on the operation and the possibilities of the software provided?
- Question 6.5: Are there sufficient possibilities for technical support?

Question 6.1 refers to a proper description of what is allowed during the test. Question 6.2 refers to the security of the content (e.g., for most practical purposes, it should not be possible for a candidate to obtain the items before the test administration), but also refers to preventing fraud during the test. Finally, security measures should be in place to prevent candidates altering their scores after the test is administered.

This means that it should be clear to a test supervisor what candidates are allowed to do during the administration of a CAT. In order to get a 'Good' on this criterion, it must be made clear, for example, whether the use of calculators, dictionaries, or other aids is allowed in the exam, what kind of help is allowed, and how to handle questions from the examinees. The security of CAT is also very much dependent on the size and quality of the item bank. A CAT needs measures to evaluate the exposure rate of items in its bank. Preferably, measures for item parameter drift should also be provided.

4.3 Reviewing a Computer Based Test

The usefulness of a review system is best demonstrated by its application. Therefore, Cito's Math Entrance Test for Teachers College (WISCAT-pabo) is evaluated below with the RCEC review system. This was done by two independent certified reviewers, who did not differ as far as the ratings on all criteria are concerned. The WISCAT-pabo is a compulsory high stakes CAT in the Netherlands, developed by Cito. It is a test of arithmetic for students in their first year of primary school teacher education. The test has been in use for over a decade with regular updates of its item bank. Candidates get three attempts to score above the cut score. If they fail the test, they cannot continue their teacher training. The psychometric advantages of computer-based adaptive testing in this instance are obvious: efficiency, high measurement precision, and prevention of the test content becoming exposed. The instrument is reviewed separately for each criterion. The review is for the most part based on a number of sources: information from the WISCAT-pabo manual (Straetmans and

Eggen 2007) that contains a detailed technical report, information on the Cito website (https://www.cito.nl/onderwijs/hoger-onderwijs/ho-toetsen-pabo/wiscat-pabo), and the reviewers taking the test several times.

4.3.1 Purpose and Use of the Test

The target population is well defined, consisting of incoming and first-year teachers college students. However, the manual does not provide information on whether or not the instrument is also suited for students with special needs. The test can be taken at various moments, but these are limited in number to assure security of the item bank.

The purpose of the test is defined as measuring the level of calculation skills of incoming and first-year students. A very detailed description of the construct of calculation skills is provided. Calculation skills are described for four domains for four different math levels: (1) Calculations and measures; (2) Geometry; (3) Information processing, statistics, and probability; and (4) Algebra, connections, graphs, and functions. The test tackles basic skills (counting, addition, subtraction, multiplication, division, powers, estimates, rounding), fractions, percentages, ratios, decimal numbers, among others. Within these domains, 21 subdomains are given and 178 descriptions of subskills.

Finally, the intended use of the WISCAT-pabo is extensively discussed in the manual. The instrument is intended to determine whether incoming students have sufficient arithmetical knowledge and skills to successfully develop the subject-specific and subject-didactic knowledge and skills to a level that is required to learn how to teach arithmetic to pupils in primary education. In addition, the instrument can serve a formative purpose when a candidate scores below the cut score. It then provides global indications on the level of mastery of several subdomains, making it possible for students to specifically focus on the subdomains they master the least.

The review yields full points for Questions 1.2 and 1.3. The score on Question 1.1 is 2 (‘Sufficient’), because very limited information was provided on the use of the test for special groups. The scoring rules yield ‘Good’ as the overall score for this criterion.

4.3.2 Quality of Test Material

The items are standardized. Each student gets a set of 50 items. The items are selected from a database of over 900 items. There are multiple-choice questions and short open-ended questions that are automatically scored. Items answered correctly yield a score of 1; items answered incorrectly get a score of 0. Both an overall score on the ability scale and indicative profile scores are generated. Because of the nature of the CAT, it is not possible for candidates to review earlier items. This is often seen as a disadvantage and is one of the reasons other methods of testing, such as multistage-testing are becoming more popular to use in high-stakes testing (van Boxel and Eggen 2017).

The instructions for the candidates are complete and clear. The Cito website provides a well-written six-page instruction. Not all 900 items in the item bank were evaluated for this review, but all items were developed by experienced item writers well acquainted with the specific subject matter. The quality of the content of the items as well as their psychometric quality is guaranteed. Items were developed through an established procedure in which experienced test developers thoroughly checked the items and all items were piloted with potential test takers. The psychometric evaluation took place through pretest procedures and continuous monitoring of new incoming data, Finally, the quality of the interface of the WISCAT-pabo can also be rated as 'Good'. Based on all the ratings on the questions and the application of the scoring rules, the instrument receives the rating 'Good' for this criterion.

4.3.3 Representativeness

A well-designed educational test or exam reflects the objectives of the (part of the) curriculum it is intended to measure. To achieve this, a test matrix is drawn up early in the design phase. This can be considered the blueprint of the test in which the test material is depicted by two dimensions, respectively the operations that students must be able to carry out and the subject matter. The test items must be evenly distributed over both dimensions.

In a CAT, the composition of the test primarily takes into account the reconciliation of the difficulty of the items with the provisional estimate of the student's skill. Without special measures, the computer will not pay attention to the distribution of the items on the subject matter and the operations. A straightforward way to guarantee this is the design of a constrained CAT in which the item bank is compartmentalized. From each part of this item bank, a specified minimum number of items is selected. A sufficient number of items needs to be available in each part of the bank, thus the developers of the CAT need to provide a balanced item bank. Otherwise, the algorithm does not produce tests that are representative for the relevant subdomains.

As the WISCAT-Pabo must generate tests that, in addition to an overall ability estimate, also provide an indication of the level of mastery of four subdomains, the CAT is designed in such a way that sufficient items from each subdomain are selected. In the WISCAT-Pabo, 50 items are presented to the students, with a very strict distribution over subdomains. Thus, all subdomains are covered. As this is a CAT and items are selected and presented based on the estimated ability of the candidate, the item difficulty is by definition at the right level for the candidates. Note that this optimal selection depends on the availability of sufficient items on the whole range of relevant abilities. The main purpose of the test is to check whether the candidates pass a specifically set cut-off score. It turns out that given the ability of the candidates this is somewhat in the middle of the ability range. Therefore it may not be too much of an issue whether there are enough items available for test takers of very high or very low ability. Also a procedure for exposure control is applied, and in case of over exposure items are to be replaced by new equivalent items. With over 900 items in the initial bank (18 times the test size), the distribution is also covered well. With maximum scores on all questions the review for this criterion results in a 'Good'.

4.3.4 Reliability (Measurement Precision)

Because this is a CAT that uses the item and test information at the heart of its procedure, measurement precision is definitely provided. Experience shows that with traditional paper-and-pencil calculation skills tests, a total number of 50 items yields high reliability. Additionally, early research has shown that CATs measure just as accurately with about half the number of items as traditional paper-and-pencil tests (Vispoel et al. 1994). The actual reduction depends on the composition of the item bank and the item selection criterion.

Thus, in the case of the WISCAT-pabo, the reliability is good. Studies performed by the authors also confirm this: the estimated mean reliability for each test is 0.91. This result was found both in simulation studies as well as in operational results. The manual provides all necessary formulas, based on relevant literature, such as Thissen (2000). It can be concluded that the way reliability was calculated is correct. In addition, the authors also provide the percentages of correct and incorrect pass-fail decisions, based on a simulation study. They show that the percentage of correct decisions is 91.54%, with about an equal percentage of candidates incorrectly passing or failing. These are good percentages, considering the passing rate of about 45%. With the additional opportunity to do a resit of the test, the number of students that fail the test while having sufficient ability is extremely small (about a fifth of a percent, after one resit). It is almost nonexistent after two resits. The review of this criterion therefore also results in a 'Good'.

4.3.5 Standard Setting and Standard Maintenance

The norms for the WISCAT-pabo are set by a combination of procedures, mainly relative. The cut score is set on the ability scale at the position of the 80th percentile of the grade 8 student population, the last grade in primary education. The underlying rationale is that based on experts' opinion, the ability of a student starting teacher training should at least be at the level of a good grade 8 student. A good grade 8 student for calculation skills was next defined at the minimum level for the top 20% of grade 8 students. This corresponding cut score on the ability scale of the WISCAT-Pabo was determined by an equating study relating the math results of 150,722 grade 8 students in 2002 on the End of Primary Education test to the results on the WISCAT-pabo. This size is most definitely large enough and, with over 87% of the total population included, it is also representative. The equating study used the OPLM model (Verhelst and Eggen 1989; Verhelst and Glas 1995), a variant of the two-parameter logistic test model. The design of the study was described in detail, as well as the meaning and the limitations of the norm scale. The results were related to the results of a sample of 4805 aspiring teachers. A wide variety of distribution characteristics was given, including the distribution of the ability of these aspiring teachers. As IRT takes such a crucial role in the procedure to set the norm, and as IRT is also crucial in the application of a CAT, it is obvious that IRT was used to maintain the standards. All procedures that were described were correctly applied. Rating all the questions in total, the review for this criterion results in a 'Good'.

4.3.6 Test Administration and Security

Most relevant information on the administration of the test is available on the Cito website. This includes the information on installing the computer software, the way the software operates, and possibilities for technical support. The safety measures include an aspect of the CAT algorithm which prevents the over- and under-utilization of items in the bank. Simply put, before a new test is started, part of the data bank is shielded from the test algorithm (counteracting overuse). The selection of items is based on a mixture of strictly adaptive and strictly random selection, while the relationship between the two modes shifts in the direction of adaptive selection with each successive item. This procedure can lead to a candidate being given an item, sometimes at the beginning of the test, which is obviously far too difficult or too easy, based on the currently estimated skills of that candidate. More detailed references, e.g., Sympson and Hetter (1985), and Revuelta and Ponsoda (1998), are given in the manual. Reviewing the responses to the questions, and the scoring rules of the review system, this criterion also yielded a rating 'Good'.

4.3.7 Review Conclusion

As the review for all criteria was positive, the conclusion is that the WISCAT-Pabo is fit for its purpose. Thus it can be used by Teachers Colleges in the Netherlands to decide whether or not starting students have sufficient calculation skills to continue their training. In addition, the WISCAT-Pabo can be used in a formative way by students scoring below the cut score to find out if there are specific subdomains in arithmetic that they should focus on.

4.4 Discussion

In the previous sections, the RCEC review system was described. Also an example of a specific review was presented. We would like to stress here that this was not a formal review and that it was only performed to make clear that reviewing the quality of educational tests and exams requires a structurally different approach than reviewing the quality of psychological tests. The field of educational measurement is still developing and improving. This means that the RCEC review system will have to be updated on a regular basis.

We hope that the RCEC review system will enjoy increasing popularity within the educational measurement community. One of the reasons is that the RCEC review system is designed to deal with one of the principal differences between exams and psychological tests. The content of the latter can remain the same over an expanded period of time, whereas the content of most exams can be deployed only once. The reason for this, of course, is that, in high-stake situations, exam content becomes known or is even made public after the first administration. This exposure makes it impossible for candidates to do a resit of the exam since future candidates can become directly acquainted with the exam content. Again, this has consequences for the application and weight of criteria like ageing of research findings. For exams, the issue of equating is much more relevant than updating the norm because it is outdated. The lifespan of a specific educational test is simply too short for that to happen.

Another reason we see for its increasing popularity is that the RCEC review system has a generic nature. It can be further adapted for reviewing different types of educational tests, including multiple-choice exams, open-ended exams, oral exams, performance exams, or, as we have shown here, computer-based educational tests. Furthermore, we would like to make a plea to not only use this system for reviewing educational tests and exams, but also as a concrete guideline for producing educational tests and exams. This could help inexperienced test and exam developers to increase their level of expertise in an efficient way, thus increasing the quality of the instruments they produce.

References

American Educational Research Association, & National Council on Measurement Used in Education. (1955). *Technical recommendations for achievement tests*. Washington, DC: National Educational Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Association of Educational Assessment—Europe. (2012). *European framework of standards for educational assessment 1.0*. Roma: Edizione Nova Cultura. Retrieved from www.

aea-europe.net, https://www.aea-europe.net/wp-content/uploads/2017/07/SW_Framework_of_European_Standards.pdf

Bartram, D. (2002). *Review model for the description and evaluation of psychological tests*. Brussels, Belgium: European Federation of Psychologists' Associations.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Buros, O. K. (Ed.). (1938). *The 1938 mental measurements yearbook*. Oxford, England: Rutgers University Press.

Camara, W. L., & Lane, S. (2006). A historical perspective and current views on the standards for educational and psychological testing. *Educational Measurement: Issues and Practice, 25*(3), 35–41. https://doi.org/10.1111/j.1745-3992.2006.00066.x.

Cambridge Assessment. (2017). *The Cambridge Approach to Assessment. Principles for designing, administering and evaluating assessment* (Revised Edition). Retrieved from http://www.cambridgeassessment.org.uk/Images/cambridge-approach-to-assessment.pdf.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2017). *The twentieth mental measurements yearbook*. Lincoln, NE: Buros Center for Testing.

Cito. (2018). *Toetsen voor de pabo - Rekentoets wiscat* [Tests for Teachers College—The Wiscat Math Entrance Test]. Retrieved from https://www.cito.nl/onderwijs/hoger-onderwijs/ho-toetsen-pabo/wiscat-pabo.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton: Educational Testing Service.

European Federation of Psychologists' Associations. (2013). *EFPA review model for the description and evaluation of psychological and educational tests. Test review form and notes for reviewers. Version 4.2.6*. Retrieved from http://www.efpa.eu/professional-development/assessment.

Evers, A. (2012). The internationalization of test reviewing: Trends, differences and results. *International Journal of Testing, 12*(2), 136–156. https://doi.org/10.1080/15305058.2012.658932.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, S. (2010a). *COTAN Beoordelingssysteem voor de kwaliteit van tests*. [COTAN Review system for the quality of tests]. Amsterdam: NIP. Retrieved from www.psynip.nl, https://www.psynip.nl/wp-content/uploads/2016/07/COTAN-Beoordelingssysteem-2010.pdf.

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010b). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10*(4), 295–317. https://doi.org/10.1080/15305058.2010.518325.

International Test Commission. (2005). *International guidelines on computer-based and internet delivered testing*. Retrieved from www.intestcom.org, https://www.intestcom.org/files/guideline_computer_based_testing.pdf.

International Test Commission. (2013). *ITC guidelines on test use. Version 1.2*. Retrieved from www.intestcom.org, https://www.intestcom.org/files/guideline_test_use.pdf.

International Test Commission. (2014). *International guidelines on the security of tests, examinations, and other assessments*. Retrieved from www.intestcom.org, https://www.intestcom.org/files/guideline_test_security.pdf.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington, D.C.: The American Council on Education.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York, NY, England: Macmillan Publishing Co, Inc.: The American Council on Education.

Nederlands Instituut van Psychologen. (1969). *Documentatie van tests en testresearch in Nederland* [*Documentation of tests and test research in the Netherlands*]. Amsterdam: Nederlands Instituut van Psychologen.

Oakland, T., Poortinga, Y. H., Schlegel, J., & Hambleton, R. K. (2001). International Test Commission: Its history, current status, and future directions. *International Journal of Testing, 1*(1), 3–32. https://doi.org/10.1207/S15327574IJT0101_2.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311–327. https://doi.org/10.1111/j.1745-3984.1998.tb00541.x.

Roorda, M. (2007). Quality systems for testing. In R. Ramaswamy & C. Wild (Eds.), *Improving testing: Process tools and techniques to assure quality* (pp. 145–176). London: Routledge.

Sanders, P. F., van Dijk, P., Eggen, T., den Otter, D., & Veldkamp, B. (2016). *RCEC Beoordelingssysteem voor de kwaliteit van studietoetsen en examens*. [Review system for the quality of educational tests and exams]. Enschede: RCEC.

Straetmans, G., & Eggen, T. (2007). *WISCAT-pabo Toetshandleiding*. [WISCAT-pabo Test manual]. Arnhem: Cito.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the annual conference of the Military Testing Association, San Diego.

Thissen, D. (2000). Reliability and measurement precision. In: Wainer, H. (Ed.), *Computerized adaptive testing. A primer* (pp. 159–184). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, D.C.: The American Council on Education.

van Boxel, M., & Eggen, T. (2017). *The Implementation of Nationwide High Stakes Computerized (adaptive) Testing in the Netherlands.* Paper presented at the 2017 conference of the International Association for Computerised Adaptive Testing, Niigata, Japan. Retrieved from http://iacat.org/implementation-nationwide-high-stakes-computerized-adaptive-testing-netherlands-0.

Verhelst, N. D., & Eggen, T. J. H. M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek.* [Psychometrical and statistical features of national assessment of educational progress] (PPON-rapport, nr. 4). Arnhem: Cito Instituut voor Toetsontwikkeling.

Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.). *Rasch models. Foundations, recent developments and applications* (pp 215–238). New York: Springer.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7*(1), 53–79. https://doi.org/10.1207/s15324818ame0701_5.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part II
Psychometrics

Chapter 5
Network Psychometrics in Educational Practice

Maximum Likelihood Estimation of the Curie-Weiss Model

M. Marsman, C. C. Tanis, T. M. Bechger and L. J. Waldorp

Abstract In network psychometrics undirected graphical models—such as the Ising model from statistical physics—are used to characterize the manifest probability distribution of psychometric data. In practice, we often find that it is extremely difficult to apply graphical models as the Ising model to educational data because (i) the model's likelihood is impossible to compute for the big data that we typically observe in educational measurement, and (ii) the model cannot handle the partially observed data that stem from incomplete test designs. In this chapter, we therefore propose to use a simplified Ising model that is known as the Curie-Weiss model. Unlike the more general Ising model, the Curie-Weiss model is computationally tractable, which makes it suitable for applications in educational measurement. The objective of this chapter is to study the statistical properties of the Curie-Weiss model and discuss its estimation with complete or incomplete data. We demonstrate that our procedures work using a simulated example, and illustrate the analysis of fit of the Curie-Weiss model using real data from the 2012 Cito Eindtoets.

5.1 Introduction

The item response theory (IRT) model is ubiquitous in the analysis of pupil responses to the questions in educational tests. In the model, a latent variable is posited to represent the ability of a pupil that is measured by the test, and then characterizes the probability that the pupil responds correctly or incorrectly to the test questions as a function of his or her ability. But these abilities are never directly observed. What we can estimate directly, however, are the proportion of pupils in a given population

M. Marsman (✉) · C. C. Tanis · L. J. Waldorp
Psychological Methods, University of Amsterdam, Nieuwe Achtergracht 129B,
PO Box 15906, 1001 NK Amsterdam, The Netherlands
e-mail: m.marsman@uva.nl

T. M. Bechger
Cito, Arnhem, The Netherlands

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_5

that obtain particular configurations of correct and incorrect responses to the test questions (Cressie and Holland 1983). Modeling these *manifest probabilities* has been the focus of several areas in the psychometric literature, such as that related to the Dutch Identity (Holland 1990; Hessen 2012; Ip 2002), log-multiplicative association models (Anderson and Vermunt 2000; Anderson and Yu 2007), and marginal models (Bergsma 1997; Bergsma and Rudas 2002). A recent addition to the psychometric literature on modeling manifest probability distributions is the network psychometric approach (van der Maas et al. 2006; Borsboom 2008; Epskamp 2017), which utilizes undirected graphical models to characterize the manifest probabilities and posits observables (e.g., responses to test questions) on a graphical or network structure.

In network psychometrics, the Ising (1925; Lenz 1920) model is a commonly used graphical model for binary random variables X, which may be used to encode the correct ($X = 1$) and incorrect ($X = 0$) responses to the questions in the test. The model is characterized by the following probability distribution over the configurations of k binary variables $\mathbf{X}$,

$$p\left(\mathbf{X}=\mathbf{x}\right)=\frac{\exp\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\mu}+\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{x}\right)}{\sum_{\mathbf{x}}\exp\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\mu}+\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{x}\right)}, \tag{5.1}$$

where the sum in the denominator ranges across all possible configurations $\mathbf{x}$ of $\mathbf{X}$, $\boldsymbol{\Sigma} = [\sigma_{ij}]$ is the symmetric $k \times k$ *connectivity matrix* that encodes the strength of interactions between the response variables—i.e., the network structure—and $\boldsymbol{\mu}$ is the $k \times 1$ vector that encodes influences from outside the network. In an educational testing context, the σ_{ij} may be taken to represent the latent processes that are shared by questions i and j, and μ_i may be taken to represent factors that are attributed to a specific question—e.g., its difficulty. This interpretation of the Ising model comes from recent work that characterizes it as the marginal distribution of a multidimensional IRT (MIRT) model (Marsman et al. 2015; Epskamp et al. 2018),

$$p\left(\mathbf{X}=\mathbf{x}\right)=\int_{\mathbb{R}^k}\prod_{i=1}^{k}p\left(X_i=x_i\mid\boldsymbol{\theta}\right)f\left(\boldsymbol{\theta}\right)\,\mathrm{d}\boldsymbol{\theta},$$

where $-\mu_i$ was shown to be equal to the MIRT model's difficulty parameter, and the square root of the r-th term in the eigenvalue decomposition of the connectivity matrix,

$$\boldsymbol{\Sigma}=\sum_{r=1}^{k}\lambda_r\mathbf{q}\mathbf{q}^{\mathsf{T}},$$

was shown to be equal to the loadings of the vector of responses on the r-th ability dimension. Viewed in this way, the two-parameter logistic (2PL) model corresponds to an Ising model with a rank one connectivity matrix.

The use of graphical models such as the Ising model in large-scale educational applications remains problematic, however. One major issue with the Ising model, for instance, is that the model is computationally intractable except for very small or highly constrained problems. This intractability resides in the model's normalizing constant, which requires the evaluation of 2^k distinct terms. With as little as 20 questions there already are over one million terms that need to be evaluated, and educational tests often consist of much more than 20 questions. This is particularly problematic for estimating the model's parameters, since closed form expressions are unavailable, and iterative procedures are needed to estimate them. The model's normalizing constant then needs to be evaluated several times in each of the iterations. Another important issue is that the Ising model consists of too many unknown parameters. With only $k = 20$ questions on the test there already are $\frac{1}{2}\left(k^2 + k\right) = 210$ free parameters, and with $k = 50$ test questions there are over one thousand free parameters. When the number of questions in the test increases, both the number of terms that are evaluated in the normalizing constant and the number of free parameters quickly grow. This makes the Ising model impractical for the large applications that are often encountered in educational measurement. Finally, the Ising model is not closed under marginalization (Marsman et al. 2017); that is,

$$\underbrace{p\left(\mathbf{X}^{(i)} = \mathbf{x}^{(i)}\right)}_{\text{not Ising model}} = \sum_{x_i} \underbrace{p\left(\mathbf{X} = \mathbf{x}\right)}_{\text{Ising model}},$$

where $\mathbf{x}^{(i)}$ is the vector $\mathbf{x}$ without element i. As a result, it is complicated to handle data that is collected with the incomplete test designs that are commonly used in educational testing.

As a way to work around these problems we will use a simplified Ising model that is known as the Curie-Weiss model (Kac 1968), in which the connectivity matrix is the scalar

$$\mathbf{\Sigma} = \sigma\, \mathbf{1}_k,$$

where $\mathbf{1}_k$ denotes a $k \times k$ matrix of ones, and there is a constant interaction $\sigma > 0$ among variables in the network. Even though the Curie-Weiss model may seem to be an oversimplification of the full Ising model, it is an incredibly rich model. For example, it provides an analytic expression for the marginal Rasch model (Bock and Aitken 1981), and is a special case of the extended Rasch model (Tjur 1982; Cressie and Holland 1983), two well-known psychometric models for the manifest probability distribution. But it also remains of interest in theoretical physics (e.g., Kochmański et al. 2013), and it is strongly related to the mean-field approximation that is often used to study theoretical properties of Ising models in the context of, for instance, magnetism. Moreover, it offers a pragmatic approximation to the full Ising model, as the Ising network can be factored into cliques—fully connected subgraphs—and the distribution of variables in such cliques can be closely approximated with a Curie-Weiss model. What matters here is that, unlike the full Ising model, the Curie-Weiss model is computationally tractable, which makes it suitable for applications in educational measurement.

The objective of the current chapter is to study the statistical properties of the Curie-Weiss model and discuss its estimation with complete or incomplete data. First, we introduce the model and analyze its statistical properties. In particular, we examine what happens when we either condition on or marginalize a part of the Curie-Weiss network. This also provides us with the opportunity to discuss its relation to the two Rasch-type models, and how its properties relate to these two models. Hereafter we show how to estimate the Curie-Weiss model and its asymptotic standard errors in both the complete data case and the incomplete data case, and then illustrate our procedures using simulated data. Data from a large-scale educational testing application—the 2012 Cito Eindtoets—is used to illustrate model fit procedures. We end the chapter with a discussion.

5.2 The Curie-Weiss Model

In a now famous series of lectures (Chrétien et al. 1968), Marc Kac lectured about the mathematical mechanisms of phase transitions (Kac 1968; Kac and Thompson 1966), and in particular the phase transitions that are observed in the study of magnetism. The Ising model is often used in this context to study dynamic properties such as the shift from a non-magnetic to a magnetic state when the material is cooled. But since it is very difficult to produce analytic results with the Ising model, Kac proposed to start with a simpler model based on the theories of magnetism of Pierre Curie and Pierre-Ernest Weiss. His version of the Curie-Weiss model is characterized by the probability distribution over the configurations of k random variables Y that take values in $\{-1, +1\}$,

$$p\left(\mathbf{Y} = \mathbf{y}\right) = \frac{\exp\left(\frac{J}{k}\sum_i\sum_j y_iy_j\right)}{\sum_{\mathbf{y}}\exp\left(\frac{J}{k}\sum_i\sum_j y_iy_j\right)},$$

where the interaction strength J/k depends on the size of the network and is constant throughout the network. In this context, the variables Y refer to the magnetic moments of electrons, which may point up ($Y = +1$) or point down ($Y = -1$). Since every variable in the Curie-Weiss network is related to all other variables, it is sometimes referred to as the *fully-connected Ising network* (Gould and Tobochnik 2010). For a detailed analysis of its dynamical properties we refer the interested reader to recent work by Kochmański et al. (2013).

Our interest in this chapter will focus on the statistical properties of the Curie-Weiss model, which inspired us to work with the following version of it

$$p\left(\mathbf{X} = \mathbf{x}\right) = \frac{\exp\left(\sum_i x_i\mu_i + \sigma\sum_i\sum_j x_ix_j\right)}{\sum_{\mathbf{x}}\exp\left(\sum_i x_i\mu_i + \sigma\sum_i\sum_j x_ix_j\right)}, \tag{5.2}$$

which differs in three important ways from the original version that was introduced by Kac. Firstly, our version of the Curie-Weiss model is used to model the distribution of {0, 1}-variables X instead of $\{-1, +1\}$-variables Y. This formulation suits the typical {0, 1} coding that is used to score responses to educational test items. Secondly, we have introduced the *external fields* $\boldsymbol{\mu}$ (c.f. Kochmański et al. 2013), which are used here to model differences in item difficulty. Moreover, since our interest is focused on its statistical instead of its dynamical properties—e.g., phase transitions—we will not investigate networks that increase in size. Therefore we have simplified the association strength from J/k to σ so that it does not explicitly depend on the network's size.

5.2.1 Some Statistical Properties of the Curie-Weiss Model

Before we continue with the problem of estimating the model's parameters $\boldsymbol{\mu}$ and σ, we will first review the model and its conditioning and marginalization properties. From our formulation of the Curie-Weiss model in Eq. (5.2) we readily find the simplified expression,

$$
\begin{aligned}
p\left(\mathbf{X}=\mathbf{x}\right) &= \frac{\exp\left(\sum_i x_i\mu_i + \sigma\sum_i\sum_j x_ix_j\right)}{\sum_{\mathbf{x}}\exp\left(\sum_i x_i\mu_i + \sigma\sum_i\sum_j x_ix_j\right)} \\
&= \frac{\left(\prod_{i=1}^k \exp\left(x_i\mu_i\right)\right)\exp\left(\sigma x_+^2\right)}{\sum_{\mathbf{x}}\left(\prod_{i=1}^k \exp\left(x_i\mu_i\right)\right)\exp\left(\sigma x_+^2\right)} \\
&= \frac{\left(\prod_{i=1}^k \exp\left(x_i\mu_i\right)\right)\exp\left(\sigma x_+^2\right)}{\sum_{s=0}^k \gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)},
\end{aligned} \tag{5.3}
$$

where x_+ is used to denote the sum score $\sum_i x_i$, and $\gamma_s\left(\boldsymbol{\mu}\right)$ is used to denote the elementary symmetric function of order s of the vector $\boldsymbol{\mu}$. The elementary symmetric function of order s of the vector $\boldsymbol{\mu}$ is defined here as

$$
\gamma_s\left(\boldsymbol{\mu}\right) = \sum_{\mathbf{x}\colon x_+=s}\prod_{i=1}^k \exp\left(x_i\mu_i\right),
$$

where the sum ranges across all configurations of $\mathbf{x}$ for which the total score x_+ is equal to s.

Observe that the normalizing constant of this version of the Curie-Weiss model

$$\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right) \exp\left(\sigma s^2\right),$$

is linear in the number of variables in the network. Even though this expression depends on the elementary symmetric functions $\gamma_s\left(\boldsymbol{\mu}\right)$, their computation using, for example, the summation algorithm (Fischer 1974; Verhelst et al. 1984) is of a quadratic order of complexity (Baker and Harwell 1996). As a result, the computation of the normalizing constant also has a quadratic order of complexity, which is a huge improvement over the exponential order of complexity of computing the normalizing constant of the Ising model. As a result, the normalizing constant of the Curie-Weiss model can be efficiently computed.

Let $A \subset \Omega = \{1, 2, \dots, k\}$ be a subset of the variables and $\bar{A}$ its complement, such that $\Omega = A \cup \bar{A}$. Consider the conditional distribution $p\left(\mathbf{X}^{(A)} \mid \mathbf{x}^{(\bar{A})}\right)$ of the variables in subset A given the remaining variables. We find

$$\begin{aligned}
p\left(\mathbf{X}^{(A)} = \mathbf{x}^{(A)} \mid \mathbf{x}^{(\bar{A})}\right) &= \frac{p\left(\mathbf{x}^{(A)}, \mathbf{x}^{(\bar{A})}\right)}{\sum_{\mathbf{x}^{(A)}} p\left(\mathbf{x}^{(A)}, \mathbf{x}^{(\bar{A})}\right)} \\
&= \frac{\left(\prod_{i\in A} \exp\left(x_i\mu_i\right)\right) \exp\left(\sigma\left[\left(x_+^{(A)}\right)^2 + 2x_+^{(A)}x_+^{(\bar{A})}\right]\right)}{\sum_{\mathbf{x}^{(A)}} \left(\prod_{i\in A} \exp\left(x_i\mu_i\right)\right) \exp\left(\sigma\left[\left(x_+^{(A)}\right)^2 + 2x_+^{(A)}x_+^{(\bar{A})}\right]\right)} \\
&= \frac{\left(\prod_{i\in A} \exp\left(x_i\left[\mu_i + 2\sigma x_+^{(\bar{A})}\right]\right)\right) \exp\left(\sigma\left(x_+^{(A)}\right)^2\right)}{\sum_{r=0}^{|A|} \gamma_r\left(\boldsymbol{\mu}^{(A)} + 2\sigma x_+^{(\bar{A})}\right) \exp\left(\sigma r^2\right)},
\end{aligned}$$

where $x_+^{(A)} = \sum_{i\in A} x_i$ denotes the sum score on the item set $A \subset \Omega$, and r the index of the rest-score that ranges from zero to the size of the set A. Note that we have used a well-known property of elementary symmetric functions (e.g., Verhelst et al. 1984). Namely, that

$$\gamma_s\left(\boldsymbol{\mu} + c\right) = \exp\left(s\,c\right)\gamma_s\left(\boldsymbol{\mu}\right),$$

such that

$$\begin{aligned}
\gamma_r\left(\boldsymbol{\mu}^{(A)} + 2\sigma x_+^{(\bar{A})}\right) &= \sum_{\mathbf{x}^{(A)}:x_+^{(A)}=r} \prod_{i\in A} \exp\left(x_i\left[\mu_i + 2\sigma x_+^{(\bar{A})}\right]\right) \\
&= \exp\left(r\,2\sigma x_+^{(\bar{A})}\right) \sum_{\mathbf{x}^{(A)}:x_+^{(A)}=r} \prod_{i\in A} \exp\left(x_i\mu_i\right)
\end{aligned}$$

is the elementary symmetric function of order r of the sub-vector $\boldsymbol{\mu}^{(A)}$ shifted by the constant $2\sigma x_{+}^{(\bar{A})}$. Since the conditional distribution is a Curie-Weiss model it follows that the model is *closed under conditioning*. That is, the distribution of variables in a Curie-Weiss network conditional upon a subset of the variables in the network results in a Curie-Weiss model. Furthermore, $p\left(\mathbf{X}^{(A)} = \mathbf{x}^{(A)} \mid \mathbf{x}^{(\bar{A})}\right) = p\left(\mathbf{X}^{(A)} = \mathbf{x}^{(A)} \mid x_{+}^{(\bar{A})}\right)$.

Of particular interest is the conditional distribution of one variable x_j conditional upon the remaining variables $\mathbf{x}^{(j)}$,

$$p\left(X_j = x_j \mid \mathbf{x}^{(j)}\right) = \frac{\exp\left(x_j\left[\mu_j + \sigma + 2\sigma x_{+}^{(j)}\right]\right)}{1 + \exp\left(\mu_j + \sigma + 2\sigma x_{+}^{(j)}\right)}, \tag{5.4}$$

which depends on the remaining variables only through the rest score $x_{+}^{(j)} = \sum_{i \neq j} x_i$ and does not depend on the external fields $\boldsymbol{\mu}^{(j)}$ that are associated to the $k-1$ variables we conditioned on. In sum, Eq. (5.4) provides an analytic expression of the item-rest regressions that can be useful for assessing the fit of the Curie-Weiss model. The conditional distribution also reveals how the Curie-Weiss model operates: The field μ_j models the general tendency to respond correctly or incorrectly to item j, and the interaction parameter σ scales the influence of the remaining $k-1$ response variables $\mathbf{x}^{(j)}$ on the response to item j. Observe that the tendency to respond correctly increases with both μ_j and σ.

While the Curie-Weiss model is closed under conditioning, it is not closed under marginalization. To see this, we derive the expression for the marginal distribution of the first $k-1$ variables of a k variable Curie-Weiss network

$$\begin{aligned} p\left(\mathbf{X}^{(k)} = \mathbf{x}^{(k)}\right) &= p\left(x_k = 1, \mathbf{X}^{(k)} = \mathbf{x}^{(k)}\right) + p\left(x_k = 0, \mathbf{X}^{(k)} = \mathbf{x}^{(k)}\right) \\ &= \frac{\left\{\exp\left(\mu_k + \sigma\left[1 + 2x_{+}^{(k)}\right]\right) + 1\right\}\left(\prod_{i=1}^{k-1}\exp\left(x_i\mu_i\right)\right)\exp\left(\sigma\left(x_{+}^{(k)}\right)^2\right)}{\sum_{s=0}^{k}\gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)}. \end{aligned}$$

Using the recursive property of elementary symmetric functions (Fischer 1974, p. 250):

$$\gamma_s\left(\boldsymbol{\mu}\right) = \exp\left(\mu_k\right)\,\gamma_{s-1}\left(\boldsymbol{\mu}^{(k)}\right) + \gamma_s\left(\boldsymbol{\mu}^{(k)}\right)$$

we can simplify this expression to

$$p\left(\mathbf{X}^{(k)} = \mathbf{x}^{(k)}\right) = \frac{\left\{\exp\left(\mu_k + \sigma\left[1 + 2x_{+}^{(k)}\right]\right) + 1\right\}\left(\prod_{i=1}^{k-1}\exp\left(x_i\mu_i\right)\right)\exp\left(\sigma\left(x_{+}^{(k)}\right)^2\right)}{\sum_{r=0}^{k-1}\left\{\exp\left(\mu_k + \sigma\left[1 + 2r\right]\right) + 1\right\}\gamma_r\left(\boldsymbol{\mu}^{(k)}\right)\exp\left(\sigma r^2\right)}, \tag{5.5}$$

Since we cannot factor out the terms that depend on variable i from the sum in the denominator due to its interaction with the rest score $x_+^{(k)}$, we are unable to simplify this expression to the form of a Curie-Weiss model. It follows that the Curie-Weiss model is *not closed under marginalization*.

5.2.2 The Curie-Weiss to Rasch Connection

Equation (5.3) can be written as:

$$\begin{aligned} p\left(\mathbf{X}=\mathbf{x}\right) &= \frac{\left(\prod_{i=1}^{k}\exp\left(x_i\mu_i\right)\right)}{\gamma_{x_+}\left(\boldsymbol{\mu}\right)}\frac{\gamma_{x_+}\left(\boldsymbol{\mu}\right)\exp\left(\sigma x_+^2\right)}{\sum_{s=0}^{k}\gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)} \\ &= p\left(\mathbf{X}=\mathbf{x}\mid X_+=x_+\right)p\left(X_+=x_+\right), \end{aligned} \tag{5.6}$$

where $p\left(\mathbf{X}=\mathbf{x}\mid X_+=x_+\right)$ can be recognized as the conditional likelihood function of the Rasch (1960) model (Andersen 1973) with item difficulties $-\mu_i$. In fact, it is readily seen that our Curie-Weiss model is a special case of the *extended Rasch model* (ERM; Tjur 1982; Cressie and Holland 1983). In general, the ERM is characterized by the following distribution

$$p\left(\mathbf{X}=\mathbf{x}\right) = \frac{\prod_{i=1}^{k}\exp(x_i\mu_i)\,\lambda_{x_+}}{\sum_{s=0}^{k}\gamma_s\left(\boldsymbol{\mu}\right)\,\lambda_s}, \tag{5.7}$$

which equals the Curie-Weiss model when $\lambda_s=\exp\left(\sigma\, s^2\right)$. An empirical illustration of such a quadratic relation can be found in Brinkhuis (in press, Chap. 5).

Importantly, the ERM can be expressed as a *marginal Rasch model* (MRM; Bock and Aitken 1981) iff it's score parameters λ_s form a moment sequence (Cressie and Holland 1983). That our Curie-Weiss model is an MRM can be seen from the original derivation by Kac, who used the following well-known identity (Stratonovich 1957; Hubbard 1959),

$$\exp\left(\sigma s^2\right) = \int_{\mathbb{R}}\frac{1}{\sqrt{\pi}}\exp\left(2\sqrt{\sigma}\,s\,\eta-\eta^2\right)\mathrm{d}\eta,$$

in which we replace the exponential of a square with the integral on the right hand side—the expectation $\mathbb{E}\left(\exp\left(2\sqrt{\sigma}\,s\,H\right)\right)$ of the normal random variable H. Writing the right hand side of this identity for the squared exponential in the numerator of the Curie-Weiss model gives

$$\begin{aligned}
p\left(\mathbf{X}=\mathbf{x}\right) &= \frac{\left(\prod_{i=1}^{k}\exp\left(x_i\mu_i\right)\right)\int_{\mathbb{R}}\frac{1}{\sqrt{\pi}}\exp\left(2\sqrt{\sigma}\,x_+\,\eta-\eta^2\right)\mathrm{d}\eta}{\sum_{s=0}^{k}\gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)}\\
&= \int_{\mathbb{R}}\prod_{i=1}^{k}\exp\left(x_i\left[\mu_i+2\sqrt{\sigma}\eta\right]\right)\frac{\exp\left(-\eta^2\right)}{\sqrt{\pi}\sum_{s=0}^{k}\gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)}\mathrm{d}\eta\\
&= \int_{\mathbb{R}}\prod_{i=1}^{k}\frac{\exp\left(x_i\left[\mu_i+2\sqrt{\sigma}\eta\right]\right)}{1+\exp\left(\mu_i+2\sqrt{\sigma}\eta\right)}\frac{\prod_{i=1}^{k}\left\{1+\exp\left(\mu_i+2\sqrt{\sigma}\eta\right)\right\}\exp\left(-\eta^2\right)}{\sqrt{\pi}\sum_{s=0}^{k}\gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)}\,\mathrm{d}\eta\\
&= \int_{\mathbb{R}}\prod_{i=1}^{k}\frac{\exp\left(x_i\left[\mu_i+\theta\right]\right)}{1+\exp\left(\mu_i+\theta\right)}\frac{\prod_{i=1}^{k}\left\{1+\exp\left(\mu_i+\theta\right)\right\}\exp\left(-\frac{1}{4\sigma}\theta^2\right)}{\sqrt{4\sigma\pi}\sum_{s=0}^{k}\gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)}\,\mathrm{d}\theta\\
&= \int_{\mathbb{R}}\prod_{i=1}^{k}p\left(X_i=x_i\mid\theta\right)f\left(\theta\right)\,\mathrm{d}\theta,
\end{aligned}$$

where we have used the change of variable $\theta=2\sqrt{\sigma}\eta$. Noting that $p\left(X_i=x_i\mid\theta\right)$ is a Rasch model with item difficulties $-\mu_i$, it follows that the Curie-Weiss model corresponds to a marginal Rasch model albeit with a rather peculiar[1] latent variable distribution $f(\theta)$ that depends on the item parameters. This distribution closely resembles a normal distribution or a skewed-normal distribution (Marsman et al. 2018; Haslbeck et al. 2018), depending on the value of the scaling parameter σ.

It is easy to verify that both the ERM and the MRM are closed under marginalization (e.g., Maris et al. 2015) whereas, as we showed earlier, the Curie-Weiss model is not. Specifically, the marginal distribution in Eq. (5.5) is not a Curie-Weiss model, although it is an ERM with

$$\lambda_r=\exp\left(\sigma\,r^2\right)\left[1+\exp\left(\mu_k+\sigma\left(1+2\,r\right)\right)\right],$$

for $r=0,\ \ldots,k-1$. Thus, marginalization gives us an ERM, but not a Curie-Weiss model. The reason that the ERM in Eq. (5.7) is closed under marginalization yet the Curie-Weiss model is not is because the characterization $\lambda_s=\exp\left(\sigma s^2\right)$ introduces a dependency between the score parameters that is not found in Eq. (5.7). The issue with the MRM is slightly different. Observe first that the MRM itself is closed under marginalization (Marsman et al. 2017), such that

$$p\left(\mathbf{X}^{(k)}=\mathbf{x}^{(k)}\right)=\int_{\mathbb{R}}\sum_{x_k}\prod_{i=1}^{k}p\left(X_i=x_i\mid\theta\right)f\left(\theta\right)\,\mathrm{d}\theta=\int_{\mathbb{R}}\prod_{i=1}^{k-1}p\left(X_i=x_i\mid\theta\right)f\left(\theta\right)\,\mathrm{d}\theta,$$

where the latent trait model is the Rasch model. In the Curie-Weiss model, the latent variable distribution $f\ (\theta)$ explicitly depends on the model parameters, including the

[1] A similar construction of the latent variable distribution can also be found in, for instance, Cressie and Holland (1983) and McCullagh (1994), and is used by Marsman et al. (in press) to generalize the Dutch Identity (Holland 1990).

field μ_k of variable k. Even though this ought not be a problem in practical applications with incomplete test designs, it does show why the marginal is not a Curie-Weiss model, since it is clear that the expectations $\mathbf{E}_f\left[p\left(\mathbf{x} \mid \Theta\right)\right]$ and $\mathbf{E}_f\left[p\left(\mathbf{x}^{(k)} \mid \Theta\right)\right]$ differ.

Compared to a regular MRM, the Curie-Weiss model has a number of convenient properties. One is that the origin of the ability distribution is identifiable so that absolute values of the item difficulties can be interpreted. Furthermore, the posterior distribution of the latent variable is available in closed form:

$$f\left(\theta \mid \mathbf{X} = \mathbf{x}\right) = f\left(\theta \mid X_+ = x_+\right) = \frac{1}{\sqrt{4\sigma\pi}} \exp\left(-\frac{1}{4\sigma}\left(\theta - 2\sigma x_+\right)^2\right).$$

This is a normal distribution[2] with mean $2\,\sigma\,x_+$ and variance $2\,\sigma$. Sampling so-called plausible values is thus trivial in the Curie-Weiss model. If we write the association parameter in its original form, $\sigma = J/k$, the posterior is a normal distribution with mean $2\,J\,\bar{x}$ and variance $2\,J/k$:

$$f\left(\theta \mid \mathbf{X} = \mathbf{x}\right) = f\left(\theta \mid X_+ = x_+\right) = \frac{\sqrt{k}}{\sqrt{4J\pi}} \exp\left(-\frac{k}{4J}\left(\theta - 2J\bar{x}\right)^2\right),$$

Note that the posterior standard deviation now shrinks with a rate of $1/\sqrt{k}$. Note further that the posterior variance is the same for all test scores, x_+, which suggests that test information is the same for all values of the latent variable. This constant rate of information is more in line with classical test theory, for example, than with regular IRT models where information is larger for abilities close to the item difficulties.

Another convenient property of the Curie-Weiss model is that it provides analytic expressions for the distribution of observables, and in particular the item-rest regressions—e.g., Eq. (5.4)—and the distribution of test scores—e.g., the second factor in Eq. (5.6). The benefit of having analytic expressions for these distributions is that they can be used to assess the fit of the Curie-Weiss model.

5.3 Maximum Likelihood Estimation of the Curie-Weiss Model

Maximum likelihood (ML) estimation of the Curie-Weiss model has been worked out for the case of a single realization of the Curie-Weiss network with an external field μ that is the same for each variable in the network. It is easily shown that for this constrained $n = 1$ case the two parameters σ and μ cannot be consistently estimated (e.g., Comets and Gidas 1991). In psychometrics and educational measurement, however, we typically have many replications of the Curie-Weiss network, i.e., $n \gg 1$.

[2] Since the Rasch model is in the exponential family, the posterior ability distribution depends on the data only through the sufficient statistic X_+ (Dawid 1979).

We will focus here on the ML procedure for estimating $\boldsymbol{\mu}$ and σ; first for the complete data case, then for the incomplete data case. Sample R-code is provided in an online repository located at https://osf.io/4m3dq/.

5.3.1 Maximum Likelihood in the Complete Data Case

The factorization in Eq. (5.6) shows that the external fields of the Curie-Weiss model can be consistently estimated with conditional maximum likelihood (CML) as used for the regular Rasch model (Andersen 1970, 1973). We may then estimate the association strength conditional upon the CML-estimates of the external fields with little loss of information (Eggen 2000). This is complicated, however, by the fact that the parameters of the Curie-Weiss model are identified, but the parameters of the conditional Rasch model are not.[3] We will therefore focus on joint estimation of the model parameters.

The complete data likelihood for the Curie-Weiss parameters $\boldsymbol{\mu}$ and σ based on the responses of n pupils to k test questions is

$$\begin{aligned} L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots, \mathbf{x}_n\right) &= \prod_{p=1}^{n} \frac{\left(\prod_{i=1}^{k} \exp\left(x_{pi}\mu_i\right)\right) \exp\left(\sigma x_{p+}^2\right)}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right) \exp\left(\sigma s^2\right)} \\ &= \frac{\exp\left(\sum_{i=1}^{k} x_{+i}\mu_i + \sigma \sum_{p=1}^{n} x_{p+}^2\right)}{\left(\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right) \exp\left(\sigma s^2\right)\right)^n}, \end{aligned}$$

and only depends on the (sufficient) statistics x_{+i}, for $i = 1, \ldots, k$, and $\sum_p x_{p+}^2$. We seek values $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}$ that maximize the likelihood function, or, similarly, the roots of its gradient $\nabla \ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots \mathbf{x}_n\right)$. The roots of the gradient can be found using iterative procedures such as the Newton-Raphson (NR) procedure. Unfortunately, NR procedures require the computation of the Hessian matrix of mixed partial (second order) derivatives in every iteration to update parameter values, which can be expensive to compute in practice. We therefore choose a *divide and conquer* strategy and maximize each of the parameters in turn, while ignoring cross-parameter dependency during optimization. Even though this might slow down convergence, it circumvents having to compute (and invert) the complete Hessian matrix in every iteration.

We first investigate the maximization of $\ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots \mathbf{x}_n\right)$ with respect to the external field μ_i of the response variable i, fixing the parameter values of the remaining k parameters to their current estimates. The partial derivative of

[3] This implies that we lose a degree of freedom by factoring the joint distribution and condition on the observed test score. To consistently estimate the parameters of the Curie-Weiss model using a two-step procedure, we therefore need to estimate a shift of the external fields in the test score distributions $p\left(X_+ = x_+\right)$.

$\ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots \mathbf{x}_n\right)$ with respect to μ_i is

$$\frac{\partial}{\partial \mu_i} \ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots \mathbf{x}_n\right) = x_{+i} - n \exp\left(\mu_i\right) \frac{\sum_{s=0}^{k-1} \gamma_s \left(\boldsymbol{\mu}^{(i)}\right) \exp\left(\sigma s^2\right)}{\sum_{s=0}^{k} \gamma_s \, \exp\left(\sigma s^2\right)}.$$

where we have used the following well-known property of the elementary symmetric function,

$$\frac{\partial}{\partial \mu_i} \gamma_s\left(\boldsymbol{\mu}\right) = \exp\left(\mu_i\right) \gamma_{s-1}\left(\boldsymbol{\mu}^{(i)}\right).$$

Setting the derivative to zero we obtain the following closed form expression for the parameter μ_i:

$$\mu_i = \ln\left(\frac{x_{+i}}{n - x_{+i}}\right) + \ln\left(\frac{\sum_{s=0}^{k-1} \gamma_s^{(i)} \exp\left(\sigma s^2\right)}{\sum_{s=0}^{k-1} \gamma_s^{(i)} \, \exp\left(\sigma \, (s+1)^2\right)}\right),$$

where the first term is a function of the sufficient statistics and the second term is a function of the remaining k parameters of the model. In an iteration t of our numerical procedure, we fix the states of these parameters to their current estimates $\hat{\boldsymbol{\mu}}_t^{(i)}$ and $\hat{\sigma}_t$ to compute an updated value for μ_i.

What remains is the maximization of $\ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots \mathbf{x}_n\right)$ with respect to the association strength σ. The partial derivative of $\ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots \mathbf{x}_n\right)$ with respect to σ is

$$\frac{\partial}{\partial \sigma} \ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots \mathbf{x}_n\right) = \sum_{p=1}^{n} x_{p+}^2 - n \frac{\sum_{s=0}^{k} s^2 \gamma_s\left(\boldsymbol{\mu}\right) \, \exp\left(\sigma s^2\right)}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right) \, \exp\left(\sigma s^2\right)}.$$

Setting this derivative to zero does not lead to a closed form solution for the parameter σ. We therefore propose a one-dimensional NR step:

$$\begin{aligned} \sigma &= \hat{\sigma} - \frac{\left.\frac{\partial}{\partial \sigma} \ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots \mathbf{x}_n\right)\right|_{\sigma=\hat{\sigma}}}{\left.\frac{\partial^2}{\partial \sigma^2} \ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots \mathbf{x}_n\right)\right|_{\sigma=\hat{\sigma}}} \\ &= \hat{\sigma} + \frac{\frac{1}{n} \sum_{p=1}^{n} x_{p+}^2 - \frac{\sum_{s=0}^{k} s^2 \, \gamma_s(\boldsymbol{\mu}) \, \exp\left(\hat{\sigma} s^2\right)}{\sum_{s=0}^{k} \gamma_s(\boldsymbol{\mu}) \, \exp\left(\hat{\sigma} s^2\right)}}{\frac{\sum_{s=0}^{k} s^4 \, \gamma_s(\boldsymbol{\mu}) \, \exp\left(\hat{\sigma} s^2\right)}{\sum_{s=0}^{k} \gamma_s(\boldsymbol{\mu}) \, \exp\left(\hat{\sigma} s^2\right)} - \left(\frac{\sum_{s=0}^{k} s^2 \, \gamma_s(\boldsymbol{\mu}) \, \exp\left(\hat{\sigma} s^2\right)}{\sum_{s=0}^{k} \gamma_s(\boldsymbol{\mu}) \, \exp\left(\hat{\sigma} s^2\right)}\right)^2}, \end{aligned}$$

where in an iteration t we evaluate the partial derivatives based on the current estimates $\hat{\boldsymbol{\mu}}_t$ and $\hat{\sigma}_t$ to compute our update of σ.

5.3.1.1 Asymptotic Standard Errors for the Complete Data Case

We estimate the variance-covariance matrix of our estimators for the parameters $\boldsymbol{\mu}$ and σ, by evaluating the inverse of the Fisher information matrix at the estimated values $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}$:

$$\text{Var}\left(\hat{\boldsymbol{\mu}}, \hat{\sigma}\right) \approx \left[\mathcal{I}\left(\hat{\boldsymbol{\mu}}, \hat{\sigma}\right)\right]^{-1},$$

where $\mathcal{I}\left(\hat{\boldsymbol{\mu}}, \hat{\sigma}\right)$ denotes the Fisher Information matrix. To compute the Fisher information matrix we work out (minus) the second order mixed derivatives from the Q-function. As the Curie-Weiss model is a member of the exponential family of distributions, the second derivatives will not depend on data, and we do not have to (numerically) evaluate any expectations in computing the information matrix. In Appendix 1 we present the mixed partial derivatives of the Q-function.

We compute the information matrix in four parts:

$$\mathcal{I}\left(\boldsymbol{\mu}, \sigma\right) = \begin{pmatrix} \mathcal{I}_{\boldsymbol{\mu}\cdot\boldsymbol{\mu}} & \mathcal{I}_{\boldsymbol{\mu}\cdot\sigma} \\ \mathcal{I}_{\sigma\cdot\boldsymbol{\mu}} & \mathcal{I}_{\sigma\cdot\sigma} \end{pmatrix}.$$

The main effects part $\mathcal{I}_{\boldsymbol{\mu}\cdot\boldsymbol{\mu}}$ of the information matrix may be expressed as follows:

$$\mathcal{I}_{\boldsymbol{\mu}\cdot\boldsymbol{\mu}} = \mathbf{A} - \mathbf{v}\mathbf{v}^{\mathsf{T}},$$

where the vector $\mathbf{v}$ has elements

$$v_i = \sqrt{n}\exp\left(\mu_i\right) \frac{\sum_{s=0}^{k-1} \gamma_s\left(\boldsymbol{\mu}^{(i)}\right)\exp\left(\sigma s^2\right)}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma\left(s+1\right)^2\right)},$$

and where $\mathbf{A}$ is a symmetric matrix with off-diagonal elements

$$A_{ij} = n\,\exp\left(\mu_i + \mu_j\right) \frac{\sum_{s=0}^{k-2} \gamma_s\left(\boldsymbol{\mu}^{(i,j)}\right)\exp\left(\sigma\left(s+2\right)^2\right)}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)},$$

and diagonal elements $A_{ii} = \sqrt{n}\, v_i$. The contribution of the scaling parameter $\mathcal{I}_{\sigma\cdot\sigma}$ is the scalar

$$\mathcal{I}_{\sigma\cdot\sigma} = n\frac{\sum_{s=0}^{k} s^4\gamma_s\left(\boldsymbol{\mu}\right)\,\exp\left(\sigma s^2\right)}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\,\exp\left(\sigma s^2\right)} - n\left(\frac{\sum_{s=0}^{k} s^2\gamma_s\left(\boldsymbol{\mu}\right)\,\exp\left(\sigma s^2\right)}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\,\exp\left(\sigma s^2\right)}\right)^2,$$

and we may express the vector $\mathcal{I}_{\boldsymbol{\mu}\cdot\sigma} = \mathcal{I}_{\sigma\cdot\boldsymbol{\mu}}^{\mathsf{T}}$ as

$$\mathcal{I}_{\boldsymbol{\mu}\cdot\sigma} = \mathbf{w} - \sqrt{n}\,\frac{\sum_{s=0}^{k} s^2\gamma_s\left(\boldsymbol{\mu}\right)\,\exp\left(\sigma s^2\right)}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)}\,\mathbf{v},$$

where the vector **w** has elements

$$w_i = n \exp\left(\mu_i\right) \frac{\sum_{s=0}^{k-1} (s+1)^2\, \gamma_s \left(\boldsymbol{\mu}^{(i)}\right) \exp\left(\sigma\, (s+1)^2\right)}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right) \exp\left(\sigma s^2\right)}.$$

Sample R-code for computing this Fisher information matrix is provided at https://osf.io/4m3dq/.

5.3.2 Maximum Likelihood Estimation in the Incomplete Data Case

When test data are collected according to a randomized incomplete test design, the structurally missing data patterns are assumed to be missing (completely) at random (Eggen 1993; Eggen and Verhelst 2011). We may then obtain unbiased estimates of the parameters from the incomplete data likelihood (Eggen 1993; Eggen and Verhelst 2011; Rubin 1976)

$$L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1^O, \ldots, \mathbf{x}_1^O\right) = \prod_{p=1}^{n} p\left(\mathbf{x}_p^O\right) = \prod_{p=1}^{n} \sum_{\mathbf{x}_p^M} p\left(\mathbf{x}_p^O, \mathbf{x}_p^M\right) = \prod_{p=1}^{n} \sum_{\mathbf{x}_p^M} p\left(\mathbf{x}_p\right),$$

where $\mathbf{x}_p^O$ denotes the observed responses for person p, $\mathbf{x}_p^M$ denotes his or her missing responses, and the complete data-likelihood $p\left(\mathbf{x}_p\right)$ is the Curie-Weiss model. When the complete data likelihood $L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1, \ldots, \mathbf{x}_n\right)$ corresponds to the Curie-Weiss model, the incomplete data likelihood $L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1^O, \ldots, \mathbf{x}_n^O\right)$ does not, because the Curie-Weiss model is not closed under marginalization. As a result, estimating the parameters of the model does not simplify to the procedure that we outlined for the complete data case.

To come to a tractable approach for estimating the Curie-Weiss parameters in the incomplete data case, we will use the Expectation-Maximization (EM) algorithm (Dempster et al. 1977), treating the unobserved responses $\mathbf{x}_1^M, \ldots, \mathbf{x}_n^M$ as missing (completely) at random. The EM approach alternates between two steps: an Expectation or E-step in which we compute the expected complete data log-likelihood of the parameters $\boldsymbol{\mu}$ and σ, and a maximization or M-step in which we find the parameter values $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}$ that maximize the expected log-likelihood that was found in the E-step. We will outline a general procedure for which the missing data patterns are unique to the individuals—e.g., as those obtained from computerized adaptive testing (van der Linden and Glas 2002; Eggen 2004; van der Linden and Glas 2010)—but note that the equations and their computations simplify considerably for the simple testing designs that are often encountered, where subsets of the items are allocated to “booklets” and pupils are allocated to one of these booklets (e.g., Eggen 1993).

5.3.2.1 The E-Step

Under the assumption that the responses of individual subjects are independent, we can write the complete data log-likelihood as

$$\ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_1^M, \ldots, \mathbf{x}_n^M, \mathbf{x}_1^O, \ldots, \mathbf{x}_n^O\right) = \sum_{p=1}^{n} \ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_p^M, \mathbf{x}_p^O\right),$$

and we may write the log-likelihood of person p as

$$\begin{aligned}\ln L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{x}_p^M, \mathbf{x}_p^O\right) = \sum_{i=1}^{k} \mu_i \left[x_{pi}^O + x_{pi}^M\right] + \sigma \left[x_{p+}^O + x_{p+}^M\right]^2 \\ - \ln\left(\sum_{s=0}^{k} \gamma_s(\boldsymbol{\mu}) \exp\left(\sigma s^2\right)\right),\end{aligned}$$

where we use

$$x_{pi}^O = \begin{cases} 1 & \text{the observed response of person } p \text{ on item } i \text{ was correct,} \\ 0 & \text{the observed response of person } p \text{ on item } i \text{ was incorrect,} \\ 0 & \text{no response of person } p \text{ on item } i \text{ was observed,} \end{cases}$$

and

$$x_{pi}^M = \begin{cases} 1 & \text{the unobserved response of person } p \text{ on item } i \text{ was correct,} \\ 0 & \text{the unobserved response of person } p \text{ on item } i \text{ was incorrect,} \\ 0 & \text{a response of person } p \text{ on item } i \text{ was observed.} \end{cases}$$

The expected log-likelihood, or Q-function, can now be written as

$$\begin{aligned}Q\left(\boldsymbol{\mu}, \sigma\ ;\ \hat{\boldsymbol{\mu}}, \hat{\sigma}\right) = \sum_{p=1}^{n}\sum_{i=1}^{k} \mu_i \left[x_{pi}^O + \mathbb{E}_{\{\hat{\boldsymbol{\mu}}, \hat{\sigma}\}}\left\{X_{pi}^M \,\middle|\, \mathbf{x}_p^O\right\}\right] \\ + \sigma \sum_{p=1}^{n} \mathbb{E}_{\{\hat{\boldsymbol{\mu}}, \hat{\sigma}\}}\left\{\left[x_{p+}^O + X_{p+}^M\right]^2 \,\middle|\, \mathbf{x}_p^O\right\} \\ - n\ln\left(\sum_{s=0}^{k} \gamma_s(\boldsymbol{\mu}) \exp\left(\sigma s^2\right)\right),\end{aligned}$$

and requires the computation of two expectations of the conditional distribution $p\left(\mathbf{x}^M \mid \mathbf{x}^O\right)$.

Since the Curie-Weiss model is closed under conditioning, we know that the conditional distribution $p_{\{\hat{\boldsymbol{\mu}},\,\hat{\sigma}\}}\left(\mathbf{x}^M \mid \mathbf{x}^O\right)$ is a Curie-Weiss model. Specifically,

$$p_{\{\hat{\boldsymbol{\mu}},\,\hat{\sigma}\}}\left(\mathbf{x}_p^M \mid x_{p+}^O\right) = \frac{\exp\left(\sum_{i=1}^k x_{pi}^M \left[\hat{\mu}_i + 2\hat{\sigma} x_{p+}^O\right] + \hat{\sigma}\left(x_{p+}^M\right)^2\right)}{\sum_{r=0}^{k_p^M} \gamma_r\left(\hat{\boldsymbol{\mu}}^{(o_p)}\right) \exp\left(r\, 2\hat{\sigma} x_{p+}^O\right)\, \exp\left(\hat{\sigma} r^2\right)},$$

$$= \frac{\exp\left(\sum_{i=1}^k x_{pi}^M \left[\hat{\mu}_i + 2\hat{\sigma} x_{p+}^O\right]\right)}{\gamma_{x_{p+}^M}\left(\hat{\boldsymbol{\mu}}^{(o_p)}\right) \exp\left(x_{p+}^M\, 2\hat{\sigma} x_{p+}^O\right)} \frac{\gamma_{x_{p+}^M}\left(\hat{\boldsymbol{\mu}}^{(o_p)}\right) \exp\left(x_{p+}^M\, 2\hat{\sigma} x_{p+}^O\right) \exp\left(\hat{\sigma}\left(x_{p+}^M\right)^2\right)}{\sum_{r=0}^{k_p^M} \gamma_r\left(\hat{\boldsymbol{\mu}}^{(o_p)}\right) \exp\left(r\, 2\hat{\sigma} x_{p+}^O\right)\, \exp\left(\hat{\sigma} r^2\right)}$$

$$= p_{\{\hat{\boldsymbol{\mu}},\,\hat{\sigma}\}}\left(\mathbf{x}_p^M \mid x_{p+}^M, x_{p+}^O\right) p_{\{\hat{\boldsymbol{\mu}},\,\hat{\sigma}\}}\left(x_{p+}^M \mid x_{p+}^O\right)$$

where k_p^M denotes the number of missing responses for pupil p, and $\gamma_r\left(\hat{\boldsymbol{\mu}}^{(o_p)}\right)$ denotes the elementary symmetric function of order r of the subvector $\hat{\boldsymbol{\mu}}^{(o_p)}$, i.e., the external fields corresponding to the missing observations for a person p.

The two expectations can now be found as follows. Assuming that the response X_{pi} of pupil p to question i is missing, we compute its expectation as

$$\mathbb{E}_{\{\hat{\boldsymbol{\mu}},\,\hat{\sigma}\}}\left\{X_{pi}^M \middle| \mathbf{x}_p^O\right\} = \sum_{\mathbf{x}_p^M} x_{pi}^M p_{\{\hat{\boldsymbol{\mu}},\,\hat{\sigma}\}}\left(\mathbf{x}_p^M \mid x_{p+}^O\right)$$

$$= \sum_{x_{pi}^M} x_{pi}^M \sum_{\mathbf{x}_p^{m(i)}} p_{\{\hat{\boldsymbol{\mu}},\,\hat{\sigma}\}}\left(\mathbf{x}_p^M \mid x_{p+}^O\right)$$

$$= \sum_{x_{pi}^M} x_{pi}^M \exp\left(\mu_i\right) \frac{\sum_{r=0}^{k_p^M-1} \gamma_r\left(\hat{\boldsymbol{\mu}}^{(o_p,\,i)}\right) \exp\left((r+1)\ 2\hat{\sigma} x_{p+}^O\right)\, \exp\left(\hat{\sigma}\ (r+1)^2\right)}{\sum_{r=0}^{k_p^M} \gamma_r\left(\hat{\boldsymbol{\mu}}^{(o_p)}\right) \exp\left(r\, 2\hat{\sigma} x_{p+}^O\right)\, \exp\left(\hat{\sigma} r^2\right)}$$

$$= \exp\left(\hat{\mu}_i\right) \frac{\sum_{r=0}^{k_p^M-1} \gamma_r\left(\hat{\boldsymbol{\mu}}^{(o_p,\,i)}\right) \exp\left((r+1)\ 2\hat{\sigma} x_{p+}^O\right)\, \exp\left(\hat{\sigma}\ (r+1)^2\right)}{\sum_{r=0}^{k_p^M} \gamma_r\left(\hat{\boldsymbol{\mu}}^{(o_p)}\right) \exp\left(r\, 2\hat{\sigma} x_{p+}^O\right)\, \exp\left(\hat{\sigma} r^2\right)},$$

otherwise this expectation is set to zero. In the same way, we find

$$\mathbb{E}_{\{\hat{\boldsymbol{\mu}},\,\hat{\sigma}\}}\left\{\left[x_{p+}^O + X_{p+}^M\right]^2 \middle| \mathbf{x}_p^O\right\} = \frac{\sum_{r=0}^{k_p^M} \left[x_{p+}^O + r\right]^2 \gamma_r\left(\hat{\boldsymbol{\mu}}^{(o_p)}\right) \exp\left(r\, 2\hat{\sigma} x_{p+}^O\right)\, \exp\left(\hat{\sigma} r^2\right)}{\sum_{r=0}^{k_p^M} \gamma_r\left(\hat{\boldsymbol{\mu}}^{(o_p)}\right) \exp\left(r\, 2\hat{\sigma} x_{p+}^O\right)\, \exp\left(\hat{\sigma} r^2\right)}.$$

Both expectations are easy to evaluate expressions of the conditional Curie-Weiss model. However, computing the basis functions for the missing item responses per pupil might be expensive. Fortunately, for incomplete test designs these functions often need only be computed per booklet, or cluster of questions, and not per individual.

5.3.3 The M-Step

The E-step effectively completes the incomplete data likelihood, and we now end up with a maximization problem that is comparable to that for the complete data case. We approach the maximization in a similar way as before, and use the same *divide and conquer* strategy to maximize each of the parameters in turn, while ignoring cross-parameter dependency during optimization.

5.3.3.1 M-Step for μ_i

The partial derivative of $Q\left(\boldsymbol{\mu}, \sigma\ ;\ \hat{\boldsymbol{\mu}}, \hat{\sigma}\right)$ with respect to μ_i is

$$\frac{\partial}{\partial \mu_i} Q\left(\boldsymbol{\mu}, \sigma\ ;\ \hat{\boldsymbol{\mu}}, \hat{\sigma}\right) = \left[x^O_{+i} + \sum_{p=1}^{n} \mathbb{E}_{\{\hat{\boldsymbol{\mu}}, \hat{\sigma}\}}\left\{X^M_{pi}\middle|\mathbf{x}^O_p\right\}\right] - n\exp\left(\mu_i\right)\frac{\sum_{s=0}^{k-1}\gamma_s\left(\boldsymbol{\mu}^{(i)}\right)\exp\left(\sigma s^2\right)}{\sum_{s=0}^{k}\gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)}.$$

When we set this derivative to zero we obtain the following closed form solution for the parameter μ_i:

$$\mu_i = \ln\left(\frac{\left[x^O_{+i} + \sum_{p=1}^{n} \mathbb{E}_{\{\hat{\boldsymbol{\mu}}, \hat{\sigma}\}}\left\{X_{pi}\middle|\mathbf{x}^O_p\right\}\right]}{n - \left[x^O_{+i} + \sum_{p=1}^{n} \mathbb{E}_{\{\hat{\boldsymbol{\mu}}, \hat{\sigma}\}}\left\{X_{pi}\middle|\mathbf{x}^O_p\right\}\right]}\right) + \ln\left(\frac{\sum_{s=0}^{k-1}\gamma_s\left(\boldsymbol{\mu}^{(i)}\right)\exp\left(\sigma s^2\right)}{\sum_{s=0}^{k-1}\gamma_s\left(\boldsymbol{\mu}^{(i)}\right)\exp\left(\sigma\left(s+1\right)^2\right)}\right),$$

again fixing the states of the remaining k parameters to their current estimates $\hat{\boldsymbol{\mu}}^{(i)}$ and $\hat{\sigma}$ to compute an updated value for μ_i.

5.3.3.2 M-Step for σ

The partial derivative of $Q\left(\boldsymbol{\mu}, \sigma\ ;\ \hat{\boldsymbol{\mu}}, \hat{\sigma}\right)$ with respect to σ is

$$\frac{\partial}{\partial \sigma} Q\left(\boldsymbol{\mu}, \sigma\ ;\ \hat{\boldsymbol{\mu}}, \hat{\sigma}\right) = \sum_{p=1}^{n} \mathbb{E}_{\{\hat{\boldsymbol{\mu}}, \hat{\sigma}\}}\left\{\left[x^O_{p+} + X^M_{p+}\right]^2\middle|\mathbf{x}^O_p\right\} - n\frac{\sum_{s=0}^{k} s^2\gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)}{\sum_{s=0}^{k}\gamma_s\left(\boldsymbol{\mu}\right)\exp\left(\sigma s^2\right)}.$$

Setting this derivative to zero does not lead to a closed form solution for the parameter σ. We therefore propose a one-dimensional NR step:

$$\begin{aligned}\sigma &= \hat{\sigma} - \frac{\frac{\partial}{\partial\sigma} Q\left(\boldsymbol{\mu}, \sigma\ ;\ \hat{\boldsymbol{\mu}}, \hat{\sigma}\right)\Big|_{\sigma=\hat{\sigma}}}{\frac{\partial^2}{\partial\sigma^2} Q\left(\boldsymbol{\mu}, \sigma\ ;\ \hat{\boldsymbol{\mu}}, \hat{\sigma}\right)\Big|_{\sigma=\hat{\sigma}}}\\ &= \hat{\sigma} + \frac{\frac{1}{n}\sum_{p=1}^{n} \mathbb{E}_{\{\hat{\boldsymbol{\mu}},\, \hat{\sigma}\}}\left\{\left[x_{p+}^{O} + X_{p+}^{M}\right]^2 \middle| \mathbf{x}_p^O\right\} - \frac{\sum_{s=0}^{k} s^2\, \gamma_s\big((\hat{\boldsymbol{\mu}})\big)\, \exp\!\left(\hat{\sigma} s^2\right)}{\sum_{s=0}^{k} \gamma_s\big((\hat{\boldsymbol{\mu}})\big)\, \exp\!\left(\hat{\sigma} s^2\right)}}{\frac{\sum_{s=0}^{k} s^4\, \gamma_s\big((\hat{\boldsymbol{\mu}})\big)\, \exp\!\left(\hat{\sigma} s^2\right)}{\sum_{s=0}^{k} \gamma_s\big((\hat{\boldsymbol{\mu}})\big)\, \exp\!\left(\hat{\sigma} s^2\right)} - \left(\frac{\sum_{s=0}^{k} s^2\, \gamma_s\big((\hat{\boldsymbol{\mu}})\big)\, \exp\!\left(\hat{\sigma} s^2\right)}{\sum_{s=0}^{k} \gamma_s\big((\hat{\boldsymbol{\mu}})\big)\, \exp\!\left(\hat{\sigma} s^2\right)}\right)^2},\end{aligned}$$

where we evaluate the partial derivatives on the current states of all parameters.

5.3.3.3 Asymptotic Standard Errors for the Incomplete Data Case

The *missing information principle* of Louis (1982) states that the observed information is equal to the complete information—the expression that we obtained for the complete data case—minus the missing information (see also Tanner 1996):

$$\mathcal{I}^O\left(\boldsymbol{\mu}, \sigma\right) = \mathcal{I}\left(\boldsymbol{\mu}, \sigma\right) - \mathcal{I}^M\left(\boldsymbol{\mu}, \sigma\right).$$

The observed Fisher information $\mathcal{I}^O$ is computed from mixed partial derivatives of the incomplete data likelihood,

$$L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{X}^O\right) = \frac{\prod_{p=1}^{n} \left(\prod_{i=1}^{k} \mathrm{e}^{x_{pi}^{O}\mu_i}\right) \mathrm{e}^{\sigma\left(x_{p+}^{O}\right)^2} \sum_{\mathbf{x}_p^M} \left(\prod_{i=1}^{k} \mathrm{e}^{x_{pi}^{M}\mu_i}\right) \mathrm{e}^{\sigma\left(\left(x_{p+}^{M}\right)^2 + 2x_{p+}^{M}x_{p+}^{O}\right)}}{\left(\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right) \mathrm{e}^{\sigma\, s^2}\right)^n},$$

and the missing Fisher information $\mathcal{I}^M$ is computed from the conditional distribution of the missing data given the observed data,

$$p\left(\mathbf{X}^M \mid \mathbf{X}^O\right) = \prod_{p=1}^{n} \frac{\exp\left(\sum_{i=1}^{k} x_{pi}^{M}\mu_i + \sigma\left(x_{p+}^{O} + x_{p+}^{M}\right)^2\right)}{\sum_{r=0}^{k_p^M} \gamma_r\left(\boldsymbol{\mu}^{(o_p)}\right) \exp\left(\sigma\left(x_{p+}^{O} + r\right)^2\right)}.$$

This conditional distribution is a member of the exponential family and its mixed second order partial derivatives do not depend on missing data such that we do not have to numerically evaluate any expectations in computing the missing information. We therefore compute the observed information as the difference between the complete information and the missing information. In Appendix 2 we present the mixed partial derivatives of the conditional distribution of the missing data. Sample R-code to compute the observed and missing information is made available at https://osf.io/4m3dq/.

5.4 Numerical Illustrations

In this section we provide two numerical illustrations of our procedures; one based on simulated data, and one based on real data from the 2012 Cito Eindtoets.

5.4.1 Simulated Example

We illustrate that our procedures work using a small simulated example. We have simulated the responses of $n = 10{,}000$ pupils to $k = 20$ test questions.[4] The external fields were sampled uniformly between -1 and $+1$, and the scale parameter σ was set to 0.05 (or $J = k\sigma = 1$). For the incomplete data case, we omitted the responses to the last five questions for 5026 randomly selected pupils and the responses to the first five questions for the 4974 remaining pupils. This mimics a two-booklet test design, where booklet one comprises of questions 1–15 and booklet two comprises of questions 6–20. The parameter estimates for this example are shown in Table 5.1.

From Table 5.1 we observe that the parameter estimates in both the complete data case and in the incomplete data case are close to their true values, and well within the range of their 95% confidence intervals. Thus, we were able to estimate the parameters of the Curie-Weiss model in both the complete and incomplete data case.

We also observe from Table 5.1 that the field estimates of questions for which no data were excluded—questions 6–15—slightly differed from their estimates obtained from the complete data case. At the same time, the (asymptotic) standard errors of all external fields increased in size, even though half of the external field parameters are estimated based on the same amount of observations. This reveals the impact of cross-parameter correlations on the obtained estimates.

5.4.2 The Cito Eindtoets 2012

As an empirical example from educational measurement, we consider an application of the Curie-Weiss model to data from the 2012 Cito Eindtoets. The data consist of the responses of $n = 133{,}768$ Dutch pupils at the end of primary school to $k = 200$ questions on twelve distinct topics from the Dutch primary school curriculum related to mathematics and language education.

The Cito Eindtoets data are typically not analyzed with a Rasch-type model. Since the test comprises of such distinct topics from the educational curriculum, we may *a priori* expect that the Curie-Weiss model fits poorly to the data from this test. However, in an earlier analysis of Cito Eindtoets data using a low-rank Ising model we found that the first principal component explained roughly 99% of the

[4] Data were generated with a Gibbs sampler (Geman and Geman 1984) utilizing the full-conditional distributions in Eq. (5.4).

Table 5.1 Parameter estimates of the Curie-Weiss model based on simulated data from $n = 10{,}000$ pupils responding to $k = 20$ test questions

	True value	Complete data		Incomplete data	
		MLE	(SE)	MLE	(SE)
μ_1	0.309	0.287	(0.068)	0.320	(0.089)
μ_2	−0.581	−0.590	(0.066)	−0.571	(0.085)
μ_3	−0.275	−0.310	(0.067)	−0.336	(0.086)
μ_4	0.550	0.448	(0.069)	0.426	(0.089)
μ_5	−0.726	−0.739	(0.066)	−0.736	(0.085)
μ_6	0.250	0.261	(0.068)	0.271	(0.083)
μ_7	−0.397	−0.428	(0.066)	−0.418	(0.082)
μ_8	−0.736	−0.755	(0.066)	−0.745	(0.082)
μ_9	0.372	0.355	(0.068)	0.365	(0.083)
μ_{10}	−0.551	−0.600	(0.066)	−0.590	(0.082)
μ_{11}	0.770	0.704	(0.070)	0.714	(0.085)
μ_{12}	0.578	0.481	(0.069)	0.491	(0.084)
μ_{13}	0.897	0.915	(0.072)	0.924	(0.086)
μ_{14}	0.646	0.692	(0.070)	0.702	(0.085)
μ_{15}	−0.251	−0.308	(0.067)	−0.298	(0.082)
μ_{16}	0.296	0.289	(0.068)	0.320	(0.088)
μ_{17}	0.541	0.556	(0.069)	0.606	(0.091)
μ_{18}	0.054	0.003	(0.067)	0.003	(0.087)
μ_{19}	0.181	0.163	(0.068)	0.166	(0.088)
μ_{20}	−0.766	−0.755	(0.066)	−0.742	(0.085)
σ	0.050	0.051	(0.002)	0.050	(0.002)

variation in the matrix of observed sufficient statistics (see Marsman et al. 2015). This principal component score correlated highly with the raw test score, and moreover, the estimated elements in the first eigenvector were found to be nearly constant (van den Bergh et al. 2018). Both observations suggest that a one-dimensional model such as a 2PL or Rasch-type model might show a reasonable fit to the observed data.

We assess the fit of the Curie-Weiss model to the 2012 Cito Eindtoets using the item-rest regressions in Eq. (5.4), focusing on the rest-scores with at least 25 observations to obtain stable estimates of the observed proportions. There are 200 item-rest regressions in total, one for each test question. We show the item-rest regressions of four questions (Questions 1, 7, 10, and 12 from the test) in Fig. 5.1, but provide all available item-rest regressions at https://osf.io/4m3dq/. When we investigate the item-rest regressions at https://osf.io/4m3dq/ it is clear that we did not find a good fit of the Curie-Weiss model to all of the 200 questions from the 2012 Cito Eindtoets. An example of an item-rest regression for a poor fitting question is that of Question 1, which is shown in the top-left panel of Fig. 5.1. Even though

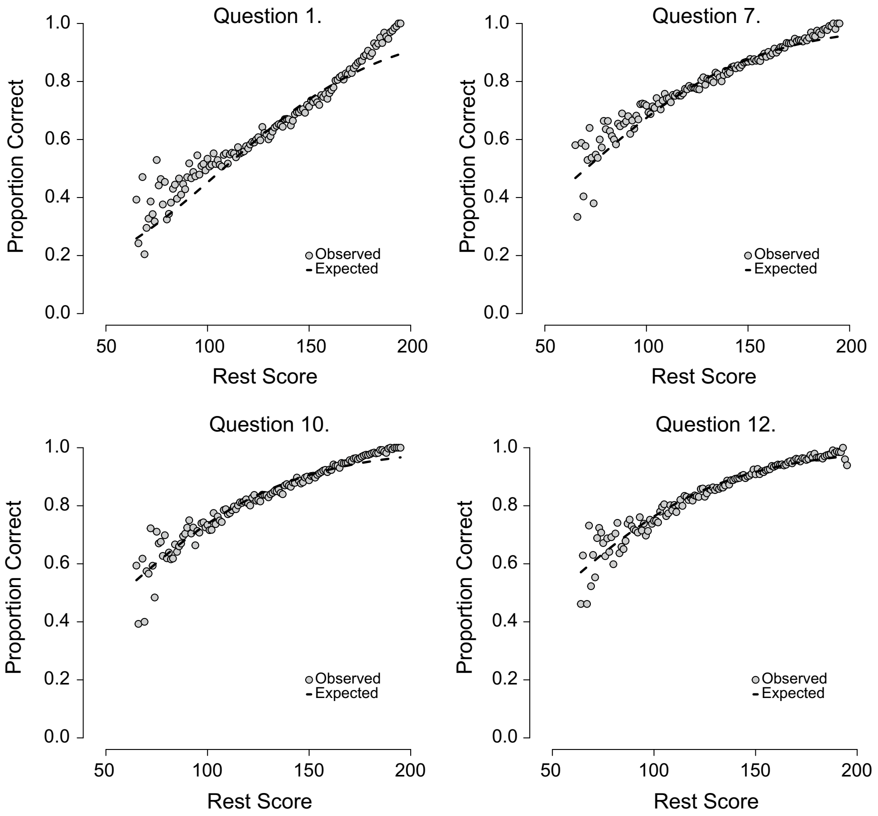


Fig. 5.1 The item-rest regressions for four questions from the Cito Eindtoets 2012

we did not find a good fit of the Curie-Weiss model to all of the questions in the Cito Eindtoets, we did observe many questions for which the estimated Curie-Weiss model did provide an at least reasonable fit. Two examples of questions for which the item-rest regression revealed a reasonable fit to the estimated Curie-Weiss model are Questions 7 and 10 in Fig. 5.1. Question 12 in Fig. 5.1 is an example of an item-rest regression that corresponds to a good fit.

Despite the fact that several of the item-rest regressions indicate a reasonable fit to the data at the item-level, it is evident that the Curie-Weiss model fits poorly to these data on a global level, as it is unable to reproduce the observed score distribution. This is illustrated in Fig. 5.2, where we compare the observed score distribution with the theoretical score distribution, e.g. Eq. (5.6). Even though it is clear that the two score distributions differ from each other, they have the same mean and variance. However, the overdispersion in the observed score distribution suggests that a more complicated model than the Curie-Weiss model, such as the low-rank Ising model that was used in Marsman et al. (2015), is likely more suited for these data.

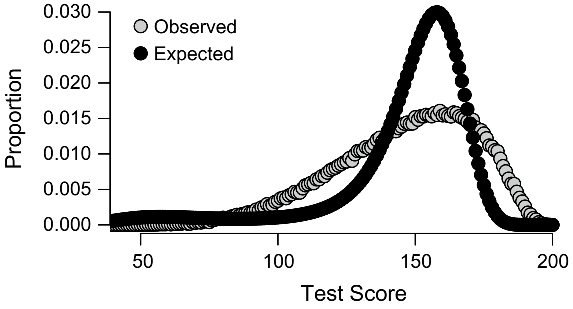


Fig. 5.2 The observed score distribution for the Cito Eindtoets 2012 (gray dots) and the theoretical score distribution as predicted by the estimated Curie-Weiss model (black dots) using Eq. (5.6)

5.5 Discussion

In this chapter we have focused on the statistical analysis of the Curie-Weiss model (Kac 1968) in the context of educational testing. In contrast to other graphical models that are regularly used in the literature—such as the Ising model (Ising 1925; Lenz 1920)—we showed that the Curie-Weiss model is computationally tractable, which makes it a practical tool for the large-scale applications that we often see in educational practice. One of the focal points of our statistical treatment was the analysis of the Curie-Weiss model in the face of data that are missing at random (Rubin 1976), such as the missing data patterns that we often observe in incomplete test designs (Eggen 1993; Eggen and Verhelst 2011). We have developed an approach using the EM algorithm (Dempster et al. 1977) to estimate the Curie-Weiss model on partially observed data. (Sample R-code is made available at https://osf.io/4m3dq/.)

We have provided two illustrations of our work. Simulated data were used to illustrate that we could recover the parameters in both the complete data case and the incomplete data case. The example using the 2012 Cito Eindtoets data was included for two reasons. Firstly, this example allowed us to illustrate the ease with which the fit of the Curie-Weiss model can be investigated using the analytic expressions of the item-rest regressions in Eq. (5.4), and the theoretical score distribution in Eq. (5.6). Secondly, it allowed us to illustrate that a simple model such as the Curie-Weiss model is able to fit complex data, at least at the item-level. The fact that the Curie-Weiss model seems to fit reasonably well to several substantively different questions—ranging from spelling and reading comprehension to working with fractions—definitely warrants further investigation.

The work in this chapter is a first step in the psychometric treatment of the Curie-Weiss model and can be generalized in several ways. For example, the factorization in Eq. (5.6) reminds us of a two-step procedure that is used in the Cito program SAUL (Structural Analysis of a Univariate Latent variable; Verhelst and Eggen 1989; Verhelst and Verstralen 2002). In this program, an IRT model is analyzed and fitted to observed data first, which in the Cito tradition comprises of either a Rasch model or a one parameter logistic model (OPLM; Verhelst and Glas 1995): a tradition that is pursued by the dexter R package (Maris et al. 2018). After a fitting model is obtained, its parameters are fixed, and the influence of background variables on the latent trait is assessed. This suggests two generalizations of the Curie-Weiss model that we have analyzed here. The first is the inclusion of integer weights or discrimination's for each of the variables, as with the OPLM. The advantage of using integer weights is that this ensures that the normalizing constant of the Curie-Weiss model remains tractable. A second generalization would be the assessment of the influence of background variables on the network's structure or score distribution. One way to realize this is by explicitly modeling the scaling parameter, for instance using a regression model to analyze differences in the scaling parameter for different pupils:

$$\sigma_p = \exp\left(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{z}_p\right),$$

where $\mathbf{z}_p$ denotes a vector of covariates that correspond to a pupil p and $\boldsymbol{\beta}$ denotes a set of regression parameters. Both generalizations would likely lead to the improved fit of the Curie-Weiss model in the Cito Eindtoets example, at both the item- and test-level. It is important to observe that the Curie-Weiss model under both generalizations remains entirely tractable.

Even though we have focused on the statistical treatment of the Curie-Weiss model from a classical perspective, it is easy to generalize our work to a Bayesian approach. One way to estimate the model in the complete data case, for example, is by means of the approximate Gibbs sampling approach of Marsman et al. (2015), that was further analyzed by Bechger et al. (2018) and Marsman et al. (2017). Another approach is based on the Gibbs sampler that Maris et al. (2015) developed for estimating the ERM, and extended by Brinkhuis (in press, Chap. 5) to the incomplete data case, where Tanis (2018) has recently shown how to adapt this Gibbs sampler for estimating the Curie-Weiss model. Data augmentation can then be used to handle incomplete data (Tanner and Wong 1987; Tanner 1996). This entails adding an additional step to the Gibbs sampler to impute missing observations based on the observed data using, for example, Eq. (5.4).

Appendix 1: Mixed Partial Derivatives Complete Data Likelihood

To compute the Fisher information matrix, we need the following mixed partial derivatives of the (complete data) likelihood:

$$\frac{\partial^2}{\partial \mu_i \partial \mu_j} L(\boldsymbol{\mu}, \sigma \mid \mathbf{X}) = n\, \mathrm{e}^{\mu_i + \mu_j} \left\{ \frac{\sum_{s=0}^{k-1} \gamma_s \left(\boldsymbol{\mu}^{(i)}\right) \mathrm{e}^{\sigma (s+1)^2}}{\sum_{s=0}^{k} \gamma_s (\boldsymbol{\mu})\, \mathrm{e}^{\sigma s^2}} \frac{\sum_{s=0}^{k-1} \gamma_s \left(\boldsymbol{\mu}^{(j)}\right) \mathrm{e}^{\sigma (s+1)^2}}{\sum_{s=0}^{k} \gamma_s (\boldsymbol{\mu})\, \mathrm{e}^{\sigma s^2}} - \frac{\sum_{s=0}^{k-2} \gamma_s \left(\boldsymbol{\mu}^{(i,j)}\right) \mathrm{e}^{\sigma (s+2)^2}}{\sum_{s=0}^{k} \gamma_s (\boldsymbol{\mu})\, \mathrm{e}^{\sigma s^2}} \right\}$$

$$\begin{aligned}\frac{\partial^2}{\partial \mu_i^2} L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{X}\right) = n\,\mathrm{e}^{\mu_i} \Bigg\{ &\mathrm{e}^{\mu_i} \left(\frac{\sum_{s=0}^{k-1} \gamma_s\left(\boldsymbol{\mu}^{(i)}\right)\, \mathrm{e}^{\sigma(s+1)^2}}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}} \right)^2 \\ &- \frac{\sum_{s=0}^{k-1} \gamma_s\left(\boldsymbol{\mu}^{(i)}\right)\, \mathrm{e}^{\sigma(s+1)^2}}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}} \Bigg\}\end{aligned}$$

$$\frac{\partial^2}{\partial \sigma^2} L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{X}\right) = n \left(\frac{\sum_{s=0}^{k} s^2 \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}} \right)^2 - n\, \frac{\sum_{s=0}^{k} s^4 \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}}$$

$$\begin{aligned}\frac{\partial^2}{\partial \sigma\, \partial \mu_i} L\left(\boldsymbol{\mu}, \sigma \mid \mathbf{X}\right) = n\,\mathrm{e}^{\mu_i} \Bigg\{ &\frac{\sum_{s=0}^{k} s^2 \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}} \, \frac{\sum_{s=0}^{k-1} \gamma_s\left(\boldsymbol{\mu}^{(i)}\right)\, \mathrm{e}^{\sigma(s+1)^2}}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}} \\ &- \frac{\sum_{s=0}^{k-1} (s+1)^2\, \gamma_s\left(\boldsymbol{\mu}^{(i)}\right)\, \mathrm{e}^{\sigma(s+1)^2}}{\sum_{s=0}^{k} \gamma_s\left(\boldsymbol{\mu}\right)\, \mathrm{e}^{\sigma s^2}} \Bigg\}\end{aligned}$$

Appendix 2: Mixed Partial Derivatives of (Conditional) Distribution of the Missing Data

To compute the missing Fisher information, we need the mixed partial derivatives of $p\left(\mathbf{X}^M \mid \mathbf{X}^O\right)$ with respect to $\boldsymbol{\mu}$ and σ. Let m_{pi} be the missing data indicator, such that

$$m_{pi} = \begin{cases} 1 & \text{if the response of pupil } p \text{ to item } i \text{ is missing,} \\ 0 & \text{if the response of pupil } p \text{ to item } i \text{ is observed.} \end{cases}$$

The mixed partial derivatives of $p\left(\mathbf{X}^M \mid \mathbf{X}^O\right)$ are then

$$\begin{aligned}\frac{\partial^2}{\partial \mu_i \partial \mu_j} p\left(\mathbf{X}^M \mid \mathbf{X}^O\right) &= \sum_{p=1}^{n} m_{pi} m_{pj}\, \mathrm{e}^{\mu_i + \mu_j} \\ &\Bigg\{ \frac{\sum_{r=0}^{k_p^M - 1} \gamma_r\left(\boldsymbol{\mu}^{(O_p,\, i)}\right) \mathrm{e}^{\sigma\left(x_p^O + r + 1\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r\left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \, \frac{\sum_{r=0}^{k_p^M - 1} \gamma_r\left(\boldsymbol{\mu}^{(O_p,\, j)}\right) \mathrm{e}^{\sigma\left(x_p^O + r + 1\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r\left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \\ &- \frac{\sum_{r=0}^{k_p^M - 2} \gamma_r\left(\boldsymbol{\mu}^{(O_p,\, i,\, j)}\right) \mathrm{e}^{\sigma\left(x_p^O + r + 2\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r\left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \Bigg\}\end{aligned}$$

$$\frac{\partial^2}{\partial \mu_i^2} p\left(\mathbf{X}^M \mid \mathbf{X}^O\right) = \sum_{p=1}^{n} m_{pi}\, \mathrm{e}^{\mu_i} \left\{ \mathrm{e}^{\mu_i} \left(\frac{\sum_{r=0}^{k_p^M - 1} \gamma_r \left(\boldsymbol{\mu}^{(O_p,\, i)}\right) \mathrm{e}^{\sigma\left(x_p^O + r + 1\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \right)^2 \right.$$
$$\left. - \frac{\sum_{r=0}^{k_p^M - 1} \gamma_r \left(\boldsymbol{\mu}^{(O_p,\, i)}\right) \mathrm{e}^{\sigma\left(x_p^O + r + 1\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \right\}$$

$$\frac{\partial^2}{\partial \sigma^2} p\left(\mathbf{X}^M \mid \mathbf{X}^O\right) = \sum_{p=1}^{n} m_{p+} \left\{ \left(\frac{\sum_{r=0}^{k_p^M} \left(x_p^O + r\right)^2 \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \right)^2 \right.$$
$$\left. - \frac{\sum_{r=0}^{k_p^M} \left(x_p^O + r\right)^4 \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \right\}$$

$$\frac{\partial^2}{\partial \sigma\, \partial \mu_i} p\left(\mathbf{X}^M \mid \mathbf{X}^O\right)$$
$$= \sum_{p=1}^{n} m_{pi}\, \mathrm{e}^{\mu_i} \left\{ \frac{\sum_{r=0}^{k_p^M} \left(x_p^O + r\right)^2 \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \frac{\sum_{r=0}^{k_p^M - 1} \gamma_r \left(\boldsymbol{\mu}^{(O_p,\, i)}\right) \mathrm{e}^{\sigma\left(x_p^O + r + 1\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \right.$$
$$\left. - \frac{\sum_{r=0}^{k_p^M - 1} \left(x_p^O + r + 1\right)^2 \gamma_r \left(\boldsymbol{\mu}^{(O_p,\, i)}\right) \mathrm{e}^{\sigma\left(x_p^O + r + 1\right)^2}}{\sum_{r=0}^{k_p^M} \gamma_r \left(\boldsymbol{\mu}^{(O_p)}\right) \mathrm{e}^{\sigma\left(x_p^O + r\right)^2}} \right\},$$

where terms for which $m_{pi} = 0$ are simply excluded from any computations.

References

Andersen, E. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological), 32*(2), 283–301. Retrieved from https://www.jstor.org/stable/2984535

Andersen, E. (1973). *Conditional inference and models for measuring* (Unpublished doctoral dissertation). Mentalhygiejnisk Forskningsinstitut.

Anderson, C., & Vermunt, J. (2000). Log-multiplicative association models as latent variable models for nominal and/or ordinal data. *Sociological Methodology*, *30*(1), 81–121. https://doi.org/10.1111/0081-1750.00076.

Anderson, C., & Yu, H.-T. (2007). Log-multiplicative association models as item response models. *Psychometrika*, *72*(1), 5–23. https://doi.org/10.1007/s11336-005-1419-2.

Baker, F., & Harwell, M. (1996). Computing elementary symmetric functions and their derivatives: A didactic. *Applied Psychological Measurement*, *20*(2), 169–192. https://doi.org/10.1177/014662169602000206.

Bechger, T. M., Maris, G. K. J., & Marsman, M. (2018). An asymptotically efficient sampler for Bayesian inference in educational measurement. arxiv:1808.03947

Bergsma, W. (1997). *Marginal models for categorical data* (Unpublished doctoral dissertation). Tilburg University.

Bergsma, W., & Rudas, T. (2002). Marginal models for categorical data. *The Annals of Statistics, 30*(1), 140–159. Retrieved from https://www.jstor.org/stable/2700006

Bock, R., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 179–197. https://doi.org/10.1007/BF02293801.

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, *64*(9), 1089–1108. https://doi.org/10.1002/jclp.

Brinkhuis, M.J.S., & Maris, G.K.J. (in press). Dynamic estimation in the extended marginal Rasch model with an application to mathematical computer-adaptive practice. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12157.

Chrétien, M., Gross, E., & Deser, S. (Eds.). (1968). *Statistical physics: Phase transitions and superfluidity, vol. 1, Brandeis University summer institute in theoretical physics*. New York: Gordon and Breach Science Publishers.

Comets, F., & Gidas, B. (1991). Asymptotics of maximum likelihood estimators for the Curie-Weiss model. *The Annals of Statistics, 19*(2), 557–578. Retrieved from https://www.jstor.org/stable/2242074

Cressie, N., & Holland, P. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, *48*(1), 129–141. https://doi.org/10.1007/BF02314681.

Dawid, A. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(1), 1–31.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38. Retrieved from https://www.jstor.org/stable/2984875

Eggen, T. (1993). Psychometrie in de praktijk [psychometrics in practice]. In T. Eggen & P. Sanders (Eds.), (pp. 239–284). Arnhem: Cito.

Eggen, T. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika*, *65*(3), 337–362. https://doi.org/10.1007/BF02296150.

Eggen, T. (2004). *Contributions to the theory and practice of the computerized adaptive testing* (Unpublished doctoral dissertation). University of Twente.

Eggen, T., & Verhelst, N. (2011). Item calibration in incomplete testing designs. *Psicologica: International Journal of Methodology and Experimental Psychology*, *32*(1), 107–132.

Epskamp, S. (2017). *Network psychometrics* (Unpublished doctoral dissertation). University of Amsterdam.

Epskamp, S., Maris, G., Waldorp, L., & Borsboom, D. (2018). Handbook of psychometrics. In P. Irwing, D. Hughes, & T. Booth (Eds.), Network psychometrics. New York, NY: Wiley-Blackwell.

Fischer, G. H. (1974). *Einführung in die theorie psychologischer tests [introduction to the theory of psychological tests]*. Bern: Huber.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596.

Gould, H., & Tobochnik, J. (2010). *Statistical and thermal physics with computer applications*. New Jersey: Princeton University.

Haslbeck, J., Epskamp, E., Marsman, M., & Waldorp, L. (2018). *Interpreting the Ising model: The input matters*. Retrieved from arxiv:1811.02916 (ArXiv preprint.)

Hessen, D. (2012). Fitting and testing conditional multinormal partial credit models. *Psychometrika*, *77*(4), 693–709. https://doi.org/10.1007/s11336-012-9277-1.

Holland, P. (1990). The Dutch Identity: A new tool for the study of item response models. *Psychometrika*, *55*(6), 5–18. https://doi.org/10.1007/BF02294739.

Hubbard, J. (1959). Calculation of partition functions. *Physical Review Letters*, *3*(2), 77–78. https://doi.org/10.1103/PhysRevLett.3.77.

Ip, E. (2002). Locally dependent latent trait model and the Dutch Identity revisited. *Psychometrika*, *67*(3), 367–386. https://doi.org/10.1007/BF02294990.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, *31*(1), 253–258. https://doi.org/10.1007/BF02980577.

Kac, M. (1968). Statistical physics: Phase transitions and superfluidity, vol. 1, Brandeis University Summer Institute in Theoretical Physics. In: M. Chrétien, E. Gross, & S. Deser (Eds.), pp. 241–305). New York: Gordon and Breach Science Publishers.

Kac, M., & Thompson, C. (1966). On the mathematical mechanism of phase transition. *Proceedings of the National Academy of Sciences of the United States of America*, *55*(4), 676–683. https://doi.org/10.1073/pnas.55.4.676.

Kochmański, M., Paszkiewicz, T., & Wolski, S. (2013). Curie-Weiss magnet-a simple model of phase transition. *European Journal of Physics*, *34*(6), 1555–1573.

Lenz, W. (1920). Beiträge zum verständnis der magnetischen eigenschaften in festen körpern. *Physikalische Zeitschrift*, *21*, 613–615.

Louis, T. (1982). Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 44*(2), 226–233. Retrieved from https://www.jstor.org/stable/2345828

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2018). dexter: Data management and analysis of tests [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dexter (R package version 0.8.1)

Maris, G., Bechger, T., & San Martin, E. (2015). A Gibbs sampler for the (extended) marginal Rasch model. *Psychometrika*, *80*(4), 859–879. https://doi.org/10.1007/s11336-015-9479-4.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L., ... Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research*, *53*(1), 15–35. https://doi.org/10.1080/00273171.2017.1379379.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank Ising networks. *Scientific Reports*, *5*(9050). https://doi.org/10.1038/srep09050

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2017). Turning simulation into estimation: Generalized exchange algorithms for exponential family models. *PLoS One*, *12*(1), 1–15 (e0169787). https://doi.org/10.1371/journal.pone.0169787

Marsman, M., Sigurdardóttir, H., Bolsinova, M., & Maris, G.K.J. (in press). Characterizing the manifest probability distributions of three latent trait models for accuracy and response time. *Psychometrika*. https://doi.org/10.1007/s11336-019-09668-3.

Marsman, M., Waldorp, L., & Maris, G. (2017). A note on large-scale logistic prediction: Using an approximate graphical model to deal with collinearity and missing data. *Behaviormetrika*, *44*(2), 513–534. https://doi.org/10.1007/s41237-017-0024-x.

McCullagh, P. (1994). Exponential mixtures and quadratic exponential families. *Biometrika*, *81*(4), 721–729. https://doi.org/10.1093/biomet/81.4.721.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581.

Stratonovich, R. (1957). On a method of calculating quantum distribution functions. *Soviet Physics Doklady*, *2*, 416.

Tanis, C. (2018). *An introduction of the Curie-Weiss model in educational measurement.*

Tanner, M. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4024-2

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540. https://doi.org/10.2307/2289457.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative poisson model. *Scandinavian Journal of Statistics*, *9*(1), 23–30.

van den Bergh, D., Bechger, T., & Marsman, M. (2018). Confirmatory item response theory using contrasts. *Manuscript in preparation*.

van der Linden, W., & Glas, C. (Eds.). (2002). *Computerized adaptive testing: Theory and practice*. New York: Kluwer Academic Publishers. https://doi.org/10.1007/0-306-47531-6.

van der Linden, W., & Glas, C. (Eds.). (2010). *Elements of adaptive testing*. New York: Springer. https://doi.org/10.1007/978-0-387-85461-8.

van der Maas, H., Dolan, C., Grasman, R., Wicherts, J., Huizenga, H., & Raijmakers, M. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842.

Verhelst, N., & Eggen, T. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek [psychometric and statistical aspects of survey research] (PPON-rapport)*. Arnhem: Cito.

Verhelst, N., & Glas, C. (1995). Rasch models. In G. H. Fisher & I. W. Molenaar (Eds.), (pp. 215–237). New York, NY: Springer.

Verhelst, N., Glas, C., & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly*, *1*(3), 245–262.

Verhelst, N., & Verstralen, H. (2002). Structural analysis of a univariate latent variable (SAUL): Theory and a computer program [Computer software manual]. Arnhem.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 6
On the Number of Items in Testing Mastery of Learning Objectives

Anton A. Béguin and J. Hendrik Straat

Abstract In individualized learning trajectories, it could be valuable to administer small tests that focus on a specific learning outcome to determine mastery of the learning objective and to evaluate whether a student can progress to other learning objectives. For this type of application, testing time competes with direct learning time, and a large number of learning objectives could invoke a potentially large burden due to testing. Thus, it is effective to limit the number of items and to reduce testing time as much as possible. However, the number of items is directly related to the accuracy of the mastery decision and the applicability of this type of formative evaluation in practical situations. For the formative evaluation to result in valid inferences, general measurement principles are valuable as well (Bennett in Assess Educ Principles Policy Pract 18:5–25, 2011). In this chapter, we provide techniques to determine the number of items and corresponding cut scores that are necessary to decide on mastery. We apply these techniques in situations with different item characteristics and provide the outcomes for varying test situations, illustrated using a practical example.

6.1 Introduction

The requirements for mastery testing and classification testing have been studied quite extensively (e.g., Wilcox 1976; de Gruijter and Hambleton 1984; van der Linden 1990; Vos 1994; Van Groen 2014). The earlier research focused on the proportion of items mastered in a well-specified content domain, containing all the relevant items in that domain (Hambleton and Novick 1973; de Gruijter and Hambleton 1984). Here, the domain is a hypothetical concept that contains all the possible items in this content domain. This proportion is referred to as π and can be interpreted as the true proportion-correct score of a person on this domain. The standard on the domain is defined as π_0, and if $\pi \geq \pi_0$, the person has mastered the domain. In practice,

A. A. Béguin (✉) · J. H. Straat
Cito, Arnhem, The Netherlands
e-mail: anton.beguin@cito.nl

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_6

only a sample of these items can be administered in a test. The score on the test, x, is evaluated against a cut-score C_x. Theoretically, C_x can be defined as $C_x = n\pi_0$, assuming equal difficulty for all items and perfect reliability, with n the number of items in the test. In reality, the assumptions are never met and misclassifications will occur if $\pi \geq \pi_0$ and $x < C_x$ or if $\pi < \pi_0$ and $x \geq C_x$. It can also be shown that the above definition leads to a non-optimal cut-score when the population mean is higher or lower than the standard, when the reliability is low, and when false positives are valued differently than false negatives (de Gruijter and Hambleton 1984). Alternative approaches to set a cut-score are based on utility theory and Bayesian decision theory (van der Linden 1980).

Wilcox (1976) reported on the appropriate lengths and passing scores based on true score proportions π_0 of 0.7, 0.75, 0.8, and 0.85. To determine the percentage of correct decisions, he defined a zone of indifference around π_0. This zone of indifference varies between $\pi_0 - 0.05$ and $\pi_0 + 0.05$, and in another condition, between $\pi_0 - 0.1$ and $\pi_0 + 0.1$. Individuals with true score probabilities within this interval will not be evaluated as incorrectly classified, independent of whether they score below or above the threshold on a test. He found that 80% or more correct classifications are established for conditions with an indifference zone of $\pi_0 \pm 0.1$ with a 19-item test, $C_x = 14$ and $\pi_0 = 0.7$. With $\pi_0 = 0.8$, this percentage is reached with an 18-item test and $C_x = 15$, while for $\pi_0 = 0.85$, an 11-item test and $C_x = 10$ is sufficient.

Other research related to mastery has focused on the application of item response theory (Rasch 1960; Lord 1980) to scale items in a test form and to determine the accuracy of person-parameter estimates. Item response theory can correct for the differences in difficulty between items and test forms that occur due to sampling from the domain. In line with this, research has been done on mastery and classification decisions, applying adaptive testing procedures (Eggen 1999; Eggen and Straetmans 2000; Van Groen 2014).

A different approach to decide on both test lengths and cut-scores of mastery tests can be based on informative Bayesian hypotheses (Hoijtink et al. 2014). Following their description, mastery is determined based on responses to a set of items. Given the person is a master, the minimum probability of answering each item correctly is determined. This leads to informative hypotheses for each of the items in the set of items. For example, if it is assumed that masters have a probability of 0.8 or more to answer an item i correctly, the following hypothesis is used:

$$H_{i,master} : \pi_i > 0.8.$$

Aggregating over all items $i = 1, \ldots,$ I, and assuming independent non-informative uniform prior distributions (Gelman et al. 2004) on the interval [0–1], the prior for $\pi = [\pi_1, \ldots, \pi_I]$ is:

$$h(\pi) = \prod_{i}^{I} \text{Beta}(\pi_i|1, 1) = 1.$$

The posterior distribution given responses $x = [x_1, \ldots, x_I]$ is:

$$g(\pi|x) \propto h(\pi) \prod_{i}^{I} \pi_i^{x_i} (1 - \pi_i)^{1-x_i}$$

$$\propto \prod_{i}^{I} \mathrm{Beta}(\pi_i|1 + x_i, 1 + (1 - x_i)).$$

The proportion of the posterior in agreement with the mastery hypotheses is:

$$f_i = \int_{\pi \in H_{i,master}} g(\pi|x)\partial\pi,$$

with $i = 1$, …, I.

If we also determine the proportion of the prior in agreement with the mastery hypotheses:

$$c_i = \int_{\pi \in H_{i,master}} h(\pi)\partial\pi,$$

a Bayes factor (Kass and Raftery 1995) can be determined, comparing the informative mastery hypotheses (m) to hypotheses without constraints (u):

$$BF_{mu} = \frac{f_i}{c_i}$$

By the same token, hypotheses can be defined by focusing on the response behavior of non-masters. This can be the complement of the behavior of masters, thus, any response pattern not meeting the criteria for masters, or it can be a set of hypotheses with additional restrictions of their own. For example, a restriction that the probability of answering an item correctly for a non-master is smaller than 0.4 for each of the items is:

$$H_{i,non\text{-}master} : \pi_i < 0.4.$$

Obviously, all kinds of other hypotheses are possible, and also hypotheses that differ per item can be combined. For example, if a researcher adopts the diagnostic perspective as formulated by Hoijtink et al. (2014), one could use latent class analysis (LCA; Lazarsfeld 1950) to define groups of masters and non-masters. More complex constructions of classes can be considered by putting restrictions on the probabilities of answering items correctly, given class membership (e.g., Heinen 1996; Hoijtink 2001; Vermunt 2001). The Bayes factor can then be used to test the most likely class membership, given a specific score pattern.

In the current research, we will apply informative Bayesian hypotheses to evaluate test lengths and cut-scores for items typically used in mastery testing, with a focus on fine-grained learning objectives. Typically, the items in assessments that focus on mastery of a learning objective are constructed in such a way that students who have mastered the learning objective will have a high probability of answering the items correctly. Students who have not mastered the learning objective will have a smaller probability of answering the items correctly. We establish guidelines for test lengths and cut-scores in three studies: a simulation study with homogeneous item characteristics, an empirical example, and a simulation based on the empirical example with heterogeneous item characteristics.

6.2 Method

6.2.1 Simulation Study with Homogeneous Item Characteristics

We evaluated the Bayes factors for number-correct scores on tests with 4–10 items. Mastery on these tests was defined as having a probability higher than 0.8 to answer each of the items correctly. For non-mastery, four different hypotheses were considered. The first hypothesis to define non-mastery was that at least one item should have a probability of being correctly answered lower or equal to 0.8. This is the complement of the definition of mastery given above. The three other hypotheses that defined non-mastery were that the probability of giving a correct answer to an item was smaller than 0.2, 0.4, or 0.6 for all of the items. The Bayes factors for mastery compared to each of these alternatives for non-mastery were calculated using the program BED.exe (Hoijtink et al. 2014).

To interpret the Bayes factors in this study, we followed the proposed guidelines in the literature (Kass and Raftery 1995; Jeffreys 1961) and adopted the rule that Bayes factors over 20 are an indicator of mastery. According to the guidelines, these values are an indication of strong evidence (BF between 20 and 150) or very strong evidence (BF > 150) that the response is based on mastery rather than non-mastery. The rationale behind the somewhat conservative rule and not accepting lower BF values is that in formative evaluations, the cost of false negatives is relatively low, while due to a false positive decision, a student could miss extra education on a topic that needed more attention.

6.2.2 Empirical Example

We applied the Bayesian hypothesis testing method to item response data collected from *Groeimeter* (2017), an evaluation platform containing mastery tests for a large number of mathematics learning objectives. Each learning objective is assessed by a 7-item test and a student takes multiple tests. The data of performances on 25 formative tests were used in Bayesian evaluations of mastery of learning objectives based on inequality constrained hypotheses identified through latent class analyses. Each formative test was evaluated separately using the following two steps:

Step 1. The probabilities of answering the seven items correctly were determined separately for masters and non-masters. In the data, both groups were present, since the formative tests were administered to students who were either masters or used the formative tests for practice. The specific test strategy for a single student was unknown to us; thus, we used latent class analyses (*poLCA;* Linzer and Lewis 2014) to identify the classes of students, which were then interpreted as masters and non-masters. The success probabilities for item i for masters $\pi_{i,masters}$ and non-masters $\pi_{i,\textit{non-master}}$, were used to specify hypotheses in which these probabilities define the borderline case for mastery and non-mastery. This resulted in inequality constrained hypotheses $H_{i,master} : \pi_i \geq \pi_{i,masters}$, and $H_{i,\textit{non-master}} : \pi_i \leq \pi_{i,\textit{non-master}}$.

Step 2. Each of $2^7 = 128$ possible score patterns were evaluated against both sets of hypotheses. If the Bayes factor for mastery against non-mastery exceeded 20, it was concluded that the response pattern corresponded to a student who had mastered the objective. For each learning objective, the Bayes factors were calculated using the program BED.exe (Hoijtink et al. 2014). Subsequently, score patterns resulting in a Bayes factor of 20 or higher were classified as indication for mastery. Since all items differ in the probabilities for mastery and non-mastery the specific score pattern impacted the Bayes factor. Patterns with equal number-correct score but a different score pattern could lead to a different indication for mastery. The minimum number-correct score for mastery was determined based on the proportion of patterns with the same number-correct score leading to a mastery decision.

6.2.3 Simulation Study Based on Empirical Data and Heterogeneous Item Characteristics

The empirical example used the results of 25 * 7 = 175 separate items from 25 different learning objectives. These items psychometrically reflected a wide range of item characteristics that can be found in real data. The relevant item characteristics were the success probabilities for masters and non-masters from the latent class analyses. These probabilities were used to define the inequality constraints for mastery and non-mastery as described in step 1 above. Based on the set of 175 items, new formative tests were assembled with different test lengths.

The required number-correct score for mastery was determined for tests with 4–10 items. For each test length, we simulated 50 replications by drawing from the 175 items without any replacements. We then estimated the Bayes factor for all the possible response patterns for inequality constrained hypotheses for masters and non-masters (similar to Step 2 in the analyses of the original empirical data). This was done to evaluate the effectiveness of different test lengths and different number-correct scores to distinguish between masters and non-masters.

6.2.4 Estimating and Validating a Predictive Model for Bayes Factors

To aggregate the results over the different tests from Groeimeter, a regression model was estimated in which the Bayes factor was predicted based on the response pattern and taking into account item characteristics. Aggregation was necessary since tests for different learning objectives will show variations in item characteristics and consequently in the required number of correct responses to indicate mastery. The dependent variable was the natural logarithm of the Bayes factor, accounting for the non-linear nature of this variable. Four predictors were used: (1) an intercept, (2) the observed proportion correct of the response pattern, (3) the sum of the success probabilities for masters on the incorrect responses, (4) the sum of the success probabilities for non-masters on the correct responses. The last two variables were centralized around the mid-point of the probability scale.

Results from the analysis based on the data from *Groeimeter* were validated with results calculated on the generated samples from the simulation study.

6.3 Results

6.3.1 Simulation Study with Homogeneous Item Characteristics

Results of the simulation study that focused on the number-correct score and test length are given in Table 6.1. The four conditions are indicated in the first column and are a single definition of mastery, with π larger than 0.8 and indicated by (m: > 0.8), crossed with each of the four conditions of non-mastery, ranging from the complement of all π larger than 0.8 (> 0.8) down to all $\pi < 0.2$. Within each condition, Bayes factors are given for test lengths of 4–10 items. Bayes factors 20 or higher are printed in italics. For each test length *n*, all of the possible number-correct scores 0 … *n* were evaluated, but only a limited number of results are reported. Indications of non-mastery and very large Bayes factors are removed from Table 6.1. This includes all factors smaller than 0.2 and larger than 1000.

The Bayes factors in Table 6.1 can be evaluated to find appropriate test lengths and cut-scores for mastery. For example, it can be seen that no Bayes factor was larger than 20 for the 4-item and 5-item tests in condition 1. For tests with lengths of 6–8 items, only a perfect score indicates mastery in condition 1, while a number-correct

Table 6.1 Bayes factor comparing mastery and non-mastery

Condition	Number correct	4	5	6	7	8	9	10
(1) m: > 0.8 nm: > 0.8	n	10.9	19.5	*35.3*	*63.7*	*111.3*	*195.0*	*372.4*
	$n-1$	1.2	2.1	3.9	7.0	12.5	*22.0*	*41.0*
	$n-2$		0.2	0.4	0.8	1.4	2.4	4.6
	$n-3$						0.3	0.5
(2) m: > 0.8 nm: < 0.6	n	*84.6*	*254.3*	*777.4*				
	$n-1$	4.0	12.0	*36.8*	*109.7*	*326.0*	*944.0*	
	$n-2$	0.2	0.6	1.7	5.2	15.6	*45.0*	*142.4*
	$n-3$				0.2	0.7	2.1	6.7
(3) m: > 0.8 nm: < 0.4	n	429.7						
	$n-1$	11.7	*53.2*	*247.4*				
	$n-2$	0.3	1.5	6.8	*30.2*	*134.9*	*577.0*	
	$n-3$				0.8	3.8	16.2	*76.8*
	$n-4$						0.4	2.1
(4) m: > 0.8 nm: < 0.2	n							
	$n-1$	*81.6*	*736.9*					
	$n-2$	1.0	9.0	*84.5*	*742.6*			
	$n-3$			1.0	9.1	*90.4*	*727.8*	
	$n-4$					1.1	9.0	*80.9*
	$n-5$							1.0

score of 8 is also a clear indication of mastery in a 9-item test, and a number-correct score of 9 indicates mastery in a 10-item test.

6.3.2 Empirical Example

Subsequently, the results of the latent class analyses are given to determine success probabilities for masters and non-masters and the resulting Bayes factors for the 25 formative tests.

6.3.2.1 Latent Class Analyses

Figure 6.1 summarizes the results of the latent class analyses for the 25 formative tests sampled from *Groeimeter*. Each plot shows the distributions of 25 * 7 = 175 different items. The three distributions represent (a) the latent class-based estimated success probabilities for the masters, (b) the estimated success probabilities for the non-masters, and (c) the difference between those success probabilities for the masters and the non-masters.

On average, masters had a success probability of 0.84 on the test items, whereas the non-masters had an average success probability of 0.33. The probabilities for masters are close to the generally accepted boundary of 80% correct for mastery, and the success probabilities for the non-masters are low enough to enable a clear distinction between the two groups. The right panel of Fig. 6.1 shows that the difference in success probabilities differs largely across the items; one item even has a higher success probability for the non-masters than for the masters. This is a suitable collection of items to investigate the impact of differences in success probabilities on the resulting Bayes factor.

6.3.2.2 Bayes Factors

We investigated the Bayes factors for all possible response patterns on the seven items for each of the 25 formative tests in *Groeimeter*. We found that no response pattern with zero, one, or two correct responses showed enough evidence for mastery; six and seven correct responses were always congruent with a mastery response pattern. For the other number-correct scores, the cumulative distribution of natural logarithms of the obtained Bayes factors are given in Fig. 6.2. The cut-score to indicate a mastery response pattern on the natural logarithm scale of the Bayes factor is ln(20) = 2.996.

In Fig. 6.2, the distributions for larger number-correct scores shift to the right, indicating that the Bayes factor generally increases with a larger number-correct score. For number-correct scores of 3–6, the percentage of the response patterns congruent with mastery of the learning objective was 2, 35, 91, and 99%, respectively.

To illustrate what conditions lead to more deviant conclusions, Table 6.2 shows two examples of response patterns with corresponding success probabilities for masters and non-masters. Test #3 has incorrect responses for easy items for the mastering group, resulting in a response pattern of five correct items and showing no significant evidence of mastery. In test #40, a response pattern resulting in three correct items was a clear indication of mastery when the correctly answered items had a very small success probability for non-masters (< 0.02).

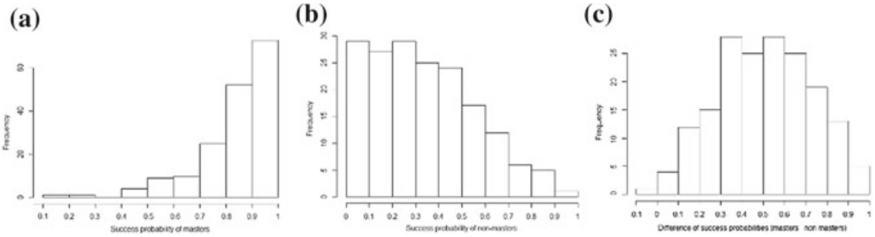


Fig. 6.1 Distributions for latent class-based success probabilities for masters and non-masters, and the difference between these probabilities

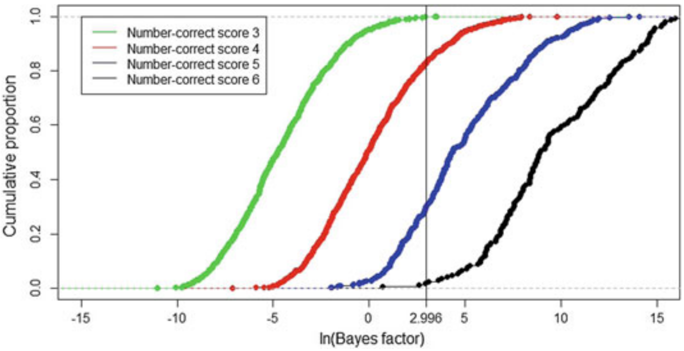


Fig. 6.2 Distribution of the natural logarithm of Bayes factors for number-correct scores 3–6

Table 6.2 Examples of response patterns leading to deviant conclusions

Test 3	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Bayes factor
Success probabilities for masters	>0.980	0.769	0.966	>0.980	0.745	0.574	0.414	
Success probabilities for non-masters	0.400	0.259	0.680	0.448	0.648	0.401	0.378	
Response pattern	0	1	1	0	1	1	1	0.141
Test 40								
Success probabilities for masters	0.529	0.973	0.850	0.895	0.822	0.868	0.719	
Success probabilities for non-masters	0.215	0.246	0.142	<0.020	0.256	<0.020	<0.020	
Response pattern	0	0	0	1	0	1	1	33.821

Table 6.3 Percentage of response patterns congruent with mastering the learning objective for different test lengths and number-correct scores

	Number-correct										
	0	1	2	3	4	5	6	7	8	9	10
4	0%	2%	11%	90%	100%						
5	0%	0%	0%	23%	95%	100%					
6	0%	0%	0%	5%	67%	97%	100%				
7	0%	0%	0%	2%	23%	89%	100%	100%			
8	0%	0%	0%	4%	9%	60%	97%	99%	100%		
9	0%	0%	0%	0%	10%	20%	81%	100%	100%	100%	
10	0%	0%	0%	0%	2%	11%	53%	96%	99%	100%	100%

6.3.3 Simulation Based on the Empirical Data and with Heterogeneous Item Characteristics

Tests were assembled with test lengths ranging from 4 to 10 items, and percentages of response patterns congruent with mastery were calculated for each number-correct score separately. These percentages are given in Table 6.3.

6.3.4 Prediction Model

The estimated regression coefficients for predicting the natural logarithm of the Bayes factors using response patterns and item characteristics are given in Table 6.4.

The effect of the proportion of correct responses can be interpreted as a modification of the intercepts, given a specific number-correct score. The larger the number-correct score, the higher the intercept. The other effects are related to the specific particular response pattern. Generally speaking, high success probabilities on correct responses for non-masters and high success probabilities on incorrect responses for masters resulted in lower Bayes factors.

Table 6.4 Regression coefficients predicting *ln*(*Bayes factor*)

Intercept	−7.112
Proportion of correct responses	15.834
Sum of success probabilities of non-masters for correct responses minus 0.5	−3.890
Sum of success probabilities of masters for incorrect responses minus 0.5	−5.229
R^2	0.953

All coefficients are significant $p < .001$

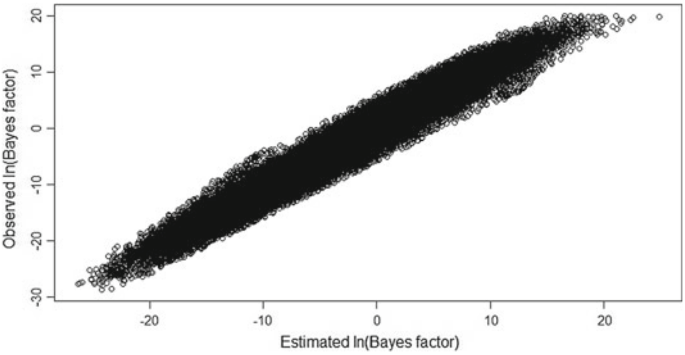


Fig. 6.3 Relationship between observed ln(Bayes factor) and the predicted ln(Bayes factor) by the regression model

6.3.4.1 Validation of the Prediction Model

The application of the regression model, as presented in Table 6.4, to all newly assembled tests in this simulation study (450 tests in total) resulted in a strong correlation (r = 0.975) between observed and predicted Bayes factors.

The relationship for each of the simulated tests is graphically presented in Fig. 6.3. The specificity and sensitivity of classifying response patterns as congruent or conflicting with learning objective mastery are 0.974 and 0.834, respectively.

6.4 Discussion and Conclusions

Bayesian hypotheses tests were used in a number of scenarios to answer the question: “How many correct responses do I need to decide on mastery?” As with all simulation studies and case studies, the results depend on the specific conditions, but some overall trends can be seen in the different studies. In the first theoretical study, inequality hypotheses were compared with equal difference in probability between mastery and non-mastery for all items. The amount of difference varied across conditions and in only one of the conditions the definition of non-mastery was the complement of mastery. In all other cases a comparison was made between two inequality hypotheses that did not cover all theoretically possible outcomes leaving some success probabilities unspecified as indicative for mastery or non-mastery. Probabilities between the upper bound for non-mastery and below the lower bound for mastery could provide alternative hypotheses to predict the data and be better suitable for some response patterns.

In the empirical example, our procedure incorporated results from LCA into inequality constrained hypotheses. The resulting definitions of mastery and non-mastery differed largely in success probabilities. The hypothesis tests based on these success probabilities were extremely powerful in detecting whether or not a response pattern was in line with mastery or non-mastery. In the second simulation study even a test length of just four items provided a significant indication for mastery in 90% of the cases where a student gave three correct answers. This amount of power can be explained by two aspects:

- The LCA indicated a large difference in average success probability for masters and non-masters. This average difference was more than 0.50.
- The success probabilities are used to define two inequality constrained hypotheses that are compared, and all other hypotheses are ignored. The success probabilities are used as lower bound for mastery and as upper bound for non-mastery. Probabilities lower than the lower bound for mastery but higher than the upper bound for non-mastery were not considered as alternative to the mastery and non-mastery hypothesis, while in practice these could give alternative, and potentially more likely, explanations for the response behavior.

As a consequence the items got almost deterministic properties in the second simulation study. If an item was answered incorrectly while the probability of a correct response for masters was very high this probably resulted in a classification as non-master. By the same token, a correct answer on an item with a very low probability for non-masters probably resulted in a classification as master.

In future research, other ways to translate results from LCA into Bayesian hypotheses should be considered. For example, definitions of mastery and non-mastery could be based on mutual exclusive categories (comparable to condition 1 in the first simulation study) or an alternative procedure could be applied in which equality constraints are used to define mastery and non-mastery. Other alternatives are to use inequality constraints on the success probability plus or minus two standard errors for non-mastery and mastery, respectively, and to consider other hypotheses such as indifference about mastery or hypotheses related to specific misconceptions.

The number of items necessary to determine mastery in a test clearly depended on the conditions, the level of certainty of the mastery decision, and the cut-score used. When using a level of certainty of 95%, the difference between heterogeneous item characteristics in the second simulation study and homogeneous item characteristics in condition 3 of the first study did not result in very different outcomes. Both studies indicated mastery for a maximum score on a four item test. With tests containing five and six items a score one point below the maximum was an indication of mastery. The same was found in the heterogenous case for a test with seven items, while a score of five on a seven items test was sufficient in the homogeneous case.

When we want to allow for an incorrect item response, based on the study with homogenous inequality constraints, we need only five items when the definition of non-mastery is based on a success probability for all items of 0.4 or less. Six items is the minimum test length with a non-mastery definition based on a probability of 0.6 or less. When non-mastery is defined as the complement of mastery, at least a 9-item test with eight correct responses is necessary to indicate mastery based on a Bayes factor of 20 or more.

As a general rule, it is reasonable to assume that you at least need six items and a cut-score of 5 to be able to decide on mastery if the test is sufficiently carefully designed to perform as a mastery test. Even in that case, it is necessary to check if all items discriminate between masters and non-masters. If the items are pre-tested and all items are selected to be in line with a mastery decision for difficulty level and discrimination, the test length can be reduced to five items.

As a more general conclusion, this research showed that the evaluation of Bayesian hypotheses can provide practical guidelines for test construction and the evaluation of empirical tests. Extending on the current analyses and in line with the tradition of earlier research into mastery testing, a next step could be to incorporate utility theory (van der Linden 1990; Vos 1994) into the procedure. This can be accomplished using differential weighting of false positive and false negative decisions. Another line of research is to extend the application of the described procedure on response data of formative assessments by identifying classes that indicate particular misconceptions, thereby providing relevant feedback, given a series of responses in the case of non-mastery.

References

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18,* 5–25.

de Gruijter, D. N. M., & Hambleton, R. K. (1984). On problems encountered using decision theory to set cutoff scores. *Applied Psychological Measurement, 8,* 1–8.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with sequential probability ratio tests. *Applied Psychological Measurement, 23,* 249–261.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713–734.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Hambleton, R. K., & Novick, M. R. (1973). Towards an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159–170.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks CA: Sage.

Hoijtink, H. (2001). Confirmatory latent class analysis: Model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research, 36,* 563–588.

Hoijtink, H., Beland, S., & Vermeulen, J. (2014). Cognitive diagnostic assessment via Bayesian evaluation of informative diagnostic hypotheses. *Psychological Methods, 19,* 21–38.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & The interpretation and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. (eds.), *Measurement and Prediction* (pp. 362–472). Princeton, NJ: Princeton University Press.

Linzer, D., & Lewis, J. (2014). *poLCA*: Latent class analysis and latent class regression models for polytomous outcome variables. R package version 1.4.1.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

van der Linden, W. J. (1980). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement, 4,* 469–492.

van der Linden, W. J. (1990). Applications of decision theory to test-based decision making. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological measurement* (pp. 129–156). Boston, MA: Kluwer-Nijhof.

Van Groen, M. M. (2014). *Adaptive testing for making unidimensional and multidimensional classification decisions.*

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement,*

Vos, H. J. (1994). *Simultaneous optimization of test-based decisions in education* (Doctoral dissertation). Enschede: University of Twente.

Wilcox, R. R. (1976). A note on the length and passing score of a mastery tests. *Journal of Educational Statistics, 1,* 359–364.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 7
Exponential Family Models for Continuous Responses

Norman D. Verhelst

In memory of Arie Dirkzwager

Abstract Two models for continuous responses that allow for separation of item and person parameters are explored. One is a newly developed model which can be seen as a Rasch model for continuous responses, the other is a slight generalization of a model proposed by Müller (1987). For both models it is shown that CML-estimation is possible in principle, but practically unfeasible. Estimation of the parameters using only item pairs, a form of pseudo-likelihood estimation, is proposed and detailed expressions for first and second order partial derivatives are given. A comparison of the information function between models for continuous and for discrete observations is discussed. The relation between these models and the probability measurement developed in the 1960s is addressed as well.

7.1 Introduction

In cognitive tests (achievement tests, placement tests) or aptitude tests, as well as in personality tests and attitude tests, the response variables for the items are discrete, having a very limited number of values: often only two in cognitive tests (representing incorrect/correct) or a few as with the use of Likert scales in attitude or personality tests, and described by expressions as 'strongly disagree', 'disagree', 'neutral', 'agree', 'strongly agree'. Note that in the latter case, the categories are considered as ordered, meaning that, e.g., the choice of the category 'strongly agree' is a sign of a more positive attitude (motivation, etc.) than the answer 'agree', and 'agree' is

N. D. Verhelst (✉)
Eurometrics, Tiel, The Netherlands
e-mail: Norman.verhelst@gmail.com

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_7

more positive than ‘neutral’, etc., or the reverse, depending of the wording of the item stem. It is usually assumed that the direction of the order can be derived from a semantic analysis of the item stem, i.e., it is not an outcome of a statistical analysis of the response data.

Measurement models to analyze these highly discrete data are widespread and well known among psychometricians. Models from the logistic family, such as the one-two- or three-parameter logistic models for binary responses (Rasch 1960; Lord and Novick 1968), the (generalized) partial credit model (Masters 1982; Muraki 1997) and the graded response model (Samejima 1974, 1997) and models, known as ‘normal ogive’, originating in biological assay (Finney 1978), but becoming popular in psychometric circles with the seminal paper by Albert (1992) on the use of sampling methods to estimate parameters.

Although certainly less popular than models for discrete data, several interesting models have been proposed and studied for continuous data, almost all in the realm of attitude scaling and personality assessment. Also, different item formats are possible to record continuous responses. One could for example present a piece of line where the endpoints are labeled (e.g., as ‘strongly disagree’ and ‘strongly agree’) and ask the respondent to put a cross at a position that best represents his or her attitude. But there are others as well. The common feature is that all answers are bounded from below and from above and that the possible answers can be regarded as a (close approximation to a) continuous response. The precise item format, although important by itself, is not considered any further in this chapter.

Samejima (1973, 1974) proposed a (family) of continuous IRT models, all derived from the graded response model (Samejima 1969, 1997) as a limiting case when the number of category responses goes to infinity. The model is complex as it assumes a latent response on the interval $(-\infty, +\infty)$ and a transformation to an interval bounded from both sides, see Bejar (1977) for an application. One of the complicating factors in Samejima’s approach is that there are two layers of latent variables: one is the construct (attitude, self-concept) to be measured and the other is a latent response which is an unbounded continuous variable, while the observable response is continuous but bounded from below and from above. To reconcile these incompatible restrictions, a transformation from an unbounded variable to a bounded one is proposed, e.g., the logit transformation. In the model developed by Mellenbergh (1994) such a transformation is not used: the continuous response variable is modeled with a one-factor model:

$$X_{ij} = \mu_j + \lambda_j \theta_i + \varepsilon_{ij},$$

where X_{ij} is the continuous observed response of respondent i to item j, μ_j is the easiness parameter of the item, λ_j the discrimination parameter, θ_i the value of the latent variable for respondent i, and ε_{ij} the continuous residual, which is usually but not necessarily normally distributed with mean zero and variance $\boldsymbol{\sigma^2}$. A comparison between the models, of Samejima and Mellenbergh, can be found in Ferrando (2002).

More recently, an interesting family of models has been proposed where the observed response variable, rescaled to the unit interval, follows a beta distribution, and can therefore be used to model monotone as well as single peaked densities (Noel and Dauvier 2007; Noel 2014). The work of Noel (2017) concentrated on unfolding models for personality traits, emotions and behavioral change.

Neither of the above models, however, are exponential family models, or do allow for separation of person and item parameters. The only published model for continuous responses which is an exponential family model and allows for parameter separation is published by Müller (1987). This model and a newly proposed one are the main topics of this chapter.

Although all the models discussed so far consider the use of continuous response variables, none of them is used for cognitive tests, like achievement tests or placement tests. For these tests considerable attention has been paid to continuous responses, especially for multiple choice tests, not in an IRT framework, but in what was called probability measurement. In this approach, the respondent does not have to pick a single alternative, but has to express his or her degree of belief or subjective probability that each of the response alternatives is the right answer (De Finetti 1965, 1970; Van Naerssen 1961; Toda 1963; Roby 1965). The attention in the sixties was directed to the problem of a good or 'admissible' scoring function, i.e., a scoring rule such that the continuous responses will reflect the true subjective probabilities. See Rippey (1970) for a comparative study of different scoring rules.

An important contribution in this area was made by Shuford et al. (1966) who showed that there was only one scoring rule (i) which maximized the expected score if and only if the continuous answers were equal to the subjective probabilities and (ii) where the score only depends on the response for the correct alternative and not on the distribution across the incorrect ones. This scoring rule is the logarithmic one, and of course has an evident disadvantage: if the continuous response to the correct alternative is zero, the score is minus infinity and can never be repaired by any finite number of (partially) correct responses. Shuford et al. were aware of this anomaly and proposed a slight modification of the logarithmic rule by a truncation on the observable variable: responses at or below 0.01 got a fixed penalty of -2. Dirkzwager (1997, 2001)[1] provided an elegant way to avoid very large penalties and at the same time to have a good approximation to the original logarithmic scoring rule. The approximation is dependent on a tuning parameter. Details on the scoring function and a quite extensive application of the system developed by Dirkzwager can be found in Holmes (2002).

Probably the main factor that hampered the application of these ideas was the poor development of suitable computer systems. Contrary to the situation with scales for personality traits, emotions and attitudes, where a continuous answer can be elicited, for example, by asking to put a mark on a line where the two end points are indicated by labels such as 'strongly disagree' and 'strongly agree' and where the mark expresses best the respondents position, the continuous answers for multiple choice

[1] Many of the writings of the late Arie Dirkzwager are not easy to find. Holmes (2002) gives quite detailed reports of his writings.

questions are multivariate. Nowadays modern computers are widespread and constructing interfaces for administering multiple choice tests with continuous answers is hardly a serious challenge.

In Sect. 7.2 a new model, a simple continuous Rasch model is introduced and parameter estimation is discussed. Section 7.3 treats a slight generalization of Müller's model and gives the details of parameter estimation along the lines sketched, but not elaborated in his 1987 article. Section 7.4 handles the problem of comparisons of information functions for different models and in Sect. 7.5 a discussion is started about the relation between IRT models for continuous responses in multiple choice tests and the scoring functions which were studied in the sixties.

7.2 A Rasch Model for Continuous Responses

7.2.1 *The Model*

Much in the spirit of the probability measurement approaches, let us imagine that a test consisting of k multiple choice items has been answered in a continuous way by assigning a non-zero number to each alternative under the restriction that the sum of these numbers across alternatives equals one. At this moment we leave the precise instructions to the respondents a bit vague, except for the fact that they know that the higher the number assigned to the right answer, the higher their score will be, i.e., there is a monotone increasing relationship between answer and the score. More on this will be said in the discussion section.

In the item response function of the Rasch model for binary data, the denominator $1 + \exp(\theta - \beta_i)$ has the role of a normalizing constant, i.e., the sum of all possible numerators, and guarantees that the sum of the probabilities of all possible answers equals one. If one considers a similar model for continuous responses in the (closed) unit interval, one arrives readily at the conditional probability density function

$$f_i(r_i|\theta) = \frac{\exp[r_i(\theta - \eta_i)]}{\int_0^1 \exp[t(\theta - \eta_i)]\,\mathrm{d}t}, \quad (r_i \in [0, 1]), \tag{7.1}$$

where r_i is a realization of the random variable R_i, the answer given to the correct alternative of item i, θ is the underlying ability, continuous and unbounded and η_i is an item parameter, representing the difficulty as in the Rasch model for binary data[2].

The denominator of (7.1) has a closed form solution. If $\theta = \eta_i$, the integral clearly equals one, as well as the numerator of (7.1). Using α_i as a shorthand for $\theta - \eta_i$, and the result from calculus that

[2] Another symbol for the difficulty parameter is used to avoid the suggestion that the difficulty parameters in the binary and continuous mode should be equal.

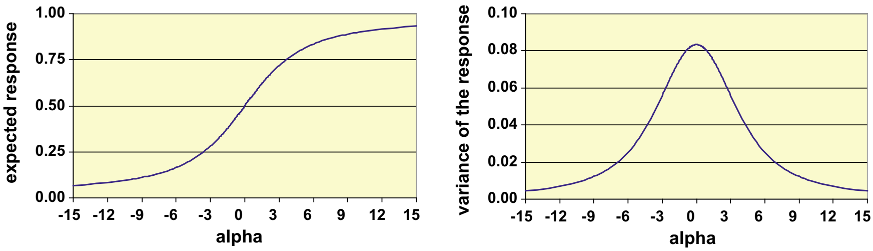


Fig. 7.1 Expected value (left) and variance (right) of R_i as a function of $\alpha_i = \theta - \eta_i$ in the Rasch model for continuous responses

$$\int \exp(t\alpha)dt = \frac{1}{\alpha}\exp(t\alpha),$$

(7.1) can also be written as

$$f_i(r_i|\theta) = \begin{cases} \frac{(\theta-\eta_i)\exp[r_i(\theta-\eta_i)]}{\exp[(\theta-\eta_i)]-1} & \text{if } \theta \neq \eta_i \text{ or } \alpha_i \neq 0, \\ 1 & \text{if } \theta = \eta_i \text{ or } \alpha_i = 0. \end{cases} \tag{7.2}$$

The regression of the response variable R_i on θ is given by

$$E(R_i|\theta) = \frac{\int_0^1 t\ \exp(t\alpha_i)\,dt}{\int_0^1 \exp(t\alpha_i)\,dt} = \begin{cases} \frac{e^{\alpha_i}(\alpha_i-1)+1}{\alpha_i(e^{\alpha_i}-1)} & \text{if } \alpha_i \neq 0, \\ 0.5 & \text{if } \alpha_i = 0, \end{cases} \tag{7.3}$$

where α_i is used as a shorthand for $\theta - \eta_i$. (Note that for the case $\alpha_i = 0$, the two integrals in (7.3) have a trivial solution.) In Fig. 7.1 (left panel) a graphical representation of the regression function is given. Notice that the horizontal axis is the difference between θ and the item parameter, and therefore the regression graph itself (with θ on the horizontal axis) will have the same form as the curve in Fig. 7.1, and for different items the curves will be shifted horizontally with respect to each other, just as the trace lines in the Rasch model for binary items. Without giving formal proofs we state some characteristics of the regression function:

1. $\lim\limits_{\theta\to-\infty} E(R_i|\theta) = 0$,
2. $\lim\limits_{\theta\to+\infty} E(R_i|\theta) = 1$,
3. $E(R_i \mid \theta)$ is monotonically increasing in θ.

One might be wondering about the numbers along the horizontal axis. In the model for binary data the graph of the item response function (which is a regression function) is usually very close to zero or to one if $\mid \theta - \beta_i \mid > 3$, while we see here a substantial difference between the expected value of the response and its lower or higher asymptote for $|\alpha|$ as large as 15. In the section about the information function we will come back to this phenomenon and give a detailed account of it.

It is easy to see from (7.1) or (7.2) that it is an exponential family density with r_i as sufficient statistic for θ. In exponential families the (Fisher) information is the variance of the sufficient statistic, and the expression for this variance is

$$\mathrm{Var}(R_i|\theta) = \begin{cases} \frac{1}{\alpha_i^2} - \frac{e^{\alpha_i}}{(e^{\alpha_i}-1)^2} & \text{if } \alpha_i \neq 0, \\ \frac{1}{12} & \text{if } \alpha_i = 0, \end{cases} \tag{7.4}$$

where the value 1/12 is the limit of the expression above it in (7.4), or the variance of a uniformly distributed variable in the unit interval. In Fig. 7.1 (right panel) a graph of the variance function is displayed.

7.2.2 Parameter Estimation

To make the model complete, we have to add an assumption about the dependence structure between item responses. As is common in most IRT models, we will assume that the responses to the items are conditionally (or locally) independent. With this assumption the model is still an exponential family and the sufficient statistic for the latent variable θ is $R = \Sigma_i R_i$. Therefore conditional maximum likelihood (CML) estimation is in principle possible and in this section the practical feasibility of CML is explored.

The conditional likelihood[3] of the data, as a function of the parameter vector $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_k)$ given the value of the sufficient statistic for θ is given by

$$L(\eta|R = r) = \frac{\exp\left[-\sum r_i\eta_i\right]}{\int_A \prod_i \exp(-t_i\eta_i)dt_i} \tag{7.5}$$

where $\int_\mathrm{A}$ denotes the multiple integral over all score vectors $(\mathrm{t}_1,\ldots,\mathrm{t}_k)$ such that $\Sigma_i t_i = r$. As a general symbol for this multiple integral, we will use $\gamma(r,k)$, where the first argument refers to the score r and the second argument denotes the number of items in the test.

To see the complexity of these γ-functions, consider first the case where $k = 2$. If $r \leq 1$, the score on any of the two items cannot be larger than r, but each score can be zero. So, let t_1 be the score on the first item; then t_1 can run from 0 to r, and of course $t_2 = r - t_1$ is linearly dependent on t_1. If $r > 1$, each score is bounded not to be smaller than $r - 1$, and of course each score cannot be larger than 1. Taking these considerations into account it is easily verified that

$$\gamma(r, 2) = \int\limits_{\max(r-1,0)}^{\min(r,1)} \exp[-(r - t_1)\eta_2] \exp(-t_1\eta_1)\, dt_1 \tag{7.6}$$

[3] For continuous data, the likelihood is proportional to the density of the observed data.

Using similar considerations as in the case with $k = 2$, one can verify that

$$\gamma(r,3) = \int\limits_{\max(r-2,0)}^{\min(r,1)} \exp(-t_2\eta_2) \int\limits_{\max(r-t_2-1,0)}^{\min(r-t_2,1)} \exp\left[(r - t_2 - t_1)\eta_3\right] \exp(-t_1\eta_1) dt_1 dt_2,$$

and in general we can write

$$\gamma(r,k) = \int\limits_{A_{k-1}}^{B_{k-1}} f_{k-1}(t_{k-1}) \ldots \int\limits_{A_i}^{B_i} f_1(t_1) \ldots g(t_1, \ldots, t_{k-1}) dt_1 \ldots dt_{k-1}, \tag{7.7}$$

where

$$\begin{aligned} A_i &= \max\left(0, r - (k-i) - \sum_{j=1}^{i-1} t_{k-j}\right), \\ B_i &= \min\left(1, r - \sum_{j=1}^{i-1} t_{k-j}\right), \\ f_i(t_i) &= \exp(-t_i\eta_i), \\ g(t_1, \ldots, t_{k-1}) &= \exp\left[-\eta_k\left(r - \sum_{j=1}^{k-1} t_j\right)\right]. \end{aligned}$$

It will be clear that evaluation of (7.7), although an explicit solution exists, is very unpractical, since in all integrals a distinction is to be made between two possible minima and two possible maxima in every integration. To illustrate this, consider the solution of (7.6), assuming that $\eta_1 \neq \eta_2$:

$$\gamma(r,2) = \begin{cases} \frac{1}{\eta_2 - \eta_1}\left[\exp(-r\eta_1) - \exp(-r\eta_2)\right], & (0 < r < 1), \\ \frac{1}{\eta_2 - \eta_1}\left[\exp(-\eta_1 - (r-1)\eta_2) - \exp(-\eta_2 - (r-1)\eta_1)\right], & (1 \leq r < 2). \end{cases} \tag{7.8}$$

If $\gamma(r, k)$ is evaluated, k different expressions will result, which are too complicated to work with. Therefore, the maximization of the conditional likelihood function is abandoned; instead recourse is taken to a pseudo-likelihood method, where the product of the conditional likelihood of all pairs of variables is maximized (Arnold and Strauss 1988, 1991; Cox and Reid 2004). This means that the function

$$PL(\eta) = \prod_{i<j} \frac{\exp(-r_i\eta_i - r_j\eta_j)}{\gamma_{r_{ij}}(\eta_i, \eta_j)}$$

will be maximized. The variable r_{ij} is defined by

$$r_{ij} = r_i + r_j, \quad (i \neq j) \tag{7.9}$$

and $\gamma_{rij}(\eta_i,\eta_j)$ is the explicit notation of $\gamma(r_{ij},2)$ with η_i and η_j as arguments.

At this point, it proves useful to reparametrize the model. Define

$$\eta_{ij} = \frac{\eta_i + \eta_j}{2} \tag{7.10}$$

and

$$\varepsilon_{ij} = \eta_{ij} - \eta_i = \frac{\eta_j - \eta_i}{2}. \tag{7.11}$$

It follows immediately that $\varepsilon_{ij} = -\varepsilon_{ji}$. Using definitions (7.10) and (7.11) and Eq. (7.8), the factor of the PL-function referring to the item pair (i,j) can be written as

$$PL(\varepsilon_{ij}) = \begin{cases} \frac{\varepsilon_{ij}\exp[(r_i - r_j)\varepsilon_{ij}]}{\sinh(r_{ij}\varepsilon_{ij})}, & (0 < r_{ij} < 1), \\ \frac{\varepsilon_{ij}\exp[(r_i - r_j)\varepsilon_{ij}]}{\sinh[(2-r_{ij})\varepsilon_{ij}]}, & (1 \leq r_{ij} < 2), \end{cases} \tag{7.12}$$

if $\varepsilon_{ij} \neq 0$, this is, if $\eta_i \neq \eta_j$. If j and i are interchanged (7.12) does not change because $\varepsilon_{ij} = -\varepsilon_{ji}$ and $\sinh(-x) = -\sinh(x)$. If $\eta_i = \eta_j$, the solution can be found directly from (7.5). It is given by

$$PL(0) = \lim_{\varepsilon_{ij} \to 0} PL(\varepsilon_{ij}) = \begin{cases} \frac{1}{r_{ij}}, & (0 < r_{ij} < 1), \\ \frac{1}{2 - r_{ij}}, & (1 \leq r_{ij} < 2). \end{cases} \tag{7.13}$$

In the sequel, only reference will be made to (7.12), but (7.13) has to be used if appropriate.

Although the events $(r_{ij} = 0)$ and $(r_{ij} = 2)$ both have probability zero, they can occur in a data set. The joint conditional density of (r_i, r_j) given $(r_{ij} = 0)$ or $(r_{ij} = 2)$, however, is independent of the item parameters, and can therefore be eliminated from the data set; only values in the open interval (0,2) are to be considered. This is similar to the fact that in the Rasch model for binary data response patterns with all zeros or all ones can be removed from the data without affecting the CML-estimates of the item parameters.

To get rid of the double expression in the right-hand side of (7.12) define the indicator variable A_{ij}, with realizations a_{ij}, as

$$A_{ij} = \begin{cases} 0 \text{ if } 0 \leq r_{ij} < 1, \\ 1 \text{ if } 1 \leq r_{ij} \leq 2, \end{cases}$$

and define the random variable B_{ij}, with realizations b_{ij}, as

$$B_{ij} = R_{ij} + 2A_{ij}(1 - R_{ij}) \tag{7.14}$$

Using (7.14), (7.12) can be rewritten as

$$PL(\varepsilon_{ij}) = \frac{\varepsilon_{ij} \exp\left[(r_i - r_j)\varepsilon_{ij}\right]}{\sinh(b_{ij}\varepsilon_{ij})} \tag{7.15}$$

and (7.13) can be rewritten as

$$PL(0) = \lim_{\varepsilon_{ij} \to 0} PL(\varepsilon_{ij}) = \frac{1}{b_{ij}}$$

Taking the logarithm of (7.15) and differentiating with respect to ε_{ij} yields

$$\frac{d \, \ln[PL(\varepsilon_{ij})]}{d\varepsilon_{ij}} = (r_i - r_j) + \frac{1}{\varepsilon_{ij}} - \frac{b_{ij}}{\tanh(b_{ij}\varepsilon_{ij})}, \tag{7.16}$$

and differentiating a second time gives

$$\frac{d^2 \ln[PL(\varepsilon_{ij})]}{d\varepsilon_{ij}^2} = -\frac{1}{\varepsilon_{ij}^2} + \frac{b_{ij}^2}{\sinh^2(b_{ij}\varepsilon_{ij})}. \tag{7.17}$$

If $\varepsilon_{ij} = 0$, (7.16) and (7.17) are undefined, but can be replaced by their limits:

$$\lim_{\varepsilon_{ij} \to 0} \frac{d \, \ln\left[PL(\varepsilon_{ij})\right]}{d\varepsilon_{ij}} = r_i - r_j$$

and

$$\lim_{\varepsilon_{ij} \to 0} \frac{d^2 \ln[PL(\varepsilon_{ij})]}{d\varepsilon_{ij}^2} = -\frac{b_{ij}^2}{3}.$$

In order to obtain estimates of the original η-parameters, the restrictions, defined by (7.11) have to be taken in account. Define the $k(k-1)/2 \times k$ matrix K by

$$K(ij, \ell) = \begin{cases} -1/2 \text{ if } \ell = i, \\ 1/2 \quad \text{if } \ell = j, \\ 0 \qquad \text{Otherwise} \end{cases} \tag{7.18}$$

where the subscript ij of the rows refers to the item pair (i,j). Define $\varepsilon = (\varepsilon_{12}, \varepsilon_{13}, \ldots, \varepsilon_{ij}, \ldots, \varepsilon_{k-1,k})$ and $\eta = (\eta_1, \ldots, \eta_k)$, then it follows immediately from (7.11) to (7.18) that

$$\varepsilon = K\eta. \tag{7.19}$$

It is immediately clear that

$$\frac{\partial \ln PL(\eta)}{\partial \eta} = K' \frac{\partial \ln PL(\varepsilon)}{\partial \varepsilon}, \tag{7.20}$$

and

$$\frac{\partial^2 \ln PL(\eta)}{\partial \eta \, \partial \eta'} = K' \frac{\partial^2 \ln PL(\varepsilon)}{\partial \varepsilon \, \partial \varepsilon'} K, \tag{7.21}$$

where the elements of the partial derivatives with respect to ε are given by (7.16). The matrix of second partial derivatives in the right-hand side of (7.21) is a diagonal matrix, whose diagonal elements are defined by (7.17). For a sample of n response patterns, the gradient and the matrix of second order partial derivatives of the PL-function is simply the sum over response patterns of expressions given by the right-hand members of Eqs. (7.20) and (7.21), respectively. The model, however, is not identified unless a normalization restriction is imposed on the η-parameters, e.g., $\eta_k = 0$. This amounts to dropping the last element of the gradient vector and the last rows and columns from the matrix of second order partial derivatives.

Initial estimates can be found by equating the right-hand member of (7.16) to zero, and solving as a univariate problem, i.e., ignoring the restrictions (7.19). Applying (7.19) then yields least squares estimates of η, which can be used as initial estimates of the item parameters.

Standard errors can be found by the so-called sandwich method. Define $\boldsymbol{g}_v$ as the vector of first partial derivatives of the pseudo-likelihood function for respondent v and H as the matrix of second partial derivatives (for all respondents jointly). All vectors $\boldsymbol{g}_v$ and the matrix H are evaluated at the value of the pseudo-likelihood estimates of the parameters. Then, the asymptotic variance-covariance matrix can be estimated by (Cox and Read 2004, p. 733)

$$H^{-1}\left[\sum_v \mathbf{g}_v \mathbf{g}_v'\right] H^{-1}.$$

7.3 An Extension of the Müller Model

7.3.1 The Model

As Samejima derived her family of continuous models as limiting cases of the graded response model when the number of possible graded responses goes to infinity, Müller considers a limiting case of Andrich's (1982) rating scale model when the number of thresholds tends to infinity. The category response function of the rating scale model (with possible answers 0, 1,…, m) is given by

$$P(X_i = j|\theta^*) = \frac{\exp[j\alpha_i^* + j(m-j)\delta^*]}{\sum_{h=0}^m \exp[h\alpha_i^* + h(m-h)\delta^*]} \tag{7.22}$$

where $\alpha_i^* = \theta^* - \eta_i^*$, the difference between the latent value and an item specific location parameter, while δ^* is half the (constant) difference between any two consecutive thresholds[4,5] If the answer R_i (with realizations r_i) is elicited by asking the respondent to put a mark on a piece of line with length d(>0) and midpoint c and where the two endpoints are labeled, then the response can be modeled by the density

$$f_i^*(r_i|\theta) = \frac{\exp[r_i\alpha_i + r_i(2c - r_i)\delta]}{\int_{c-d/2}^{c+d/2} \exp[t\alpha_i + t(2c-t)\delta]\,\mathrm{d}t}, \quad \left(r_i \in \left[c - \frac{d}{2}, c + \frac{d}{2}\right]\right), \tag{7.23}$$

with $\alpha_i = \theta - \eta_i$. Moreover, Müller shows that the thresholds are uniformly distributed in the interval $[\eta_i - \delta d, \eta_i + \delta d]$.

Of course, the length of the line in the response format can be expressed in arbitrary units and with an arbitrary reference point, so that we can assume without loss of generality that $c = 0.5$ and $d = 1$. And as a slight extension of Müller's model, it will be assumed that the δ-parameter can vary across items. This gives the density equation we will be using in the present section:

$$f_i(r_i|\theta) = \frac{\exp[r_i\alpha_i + r_i(1-r_i)\delta_i]}{\int_0^1 \exp[t\alpha_i + t(1-t)\delta_i]dt}, \tag{7.24}$$

and by completing the square, the numerator of (7.24) can be written as

$$\exp[r_i\alpha_i + r_i(1-r_i)\delta_i] = \exp\left[-\delta_i\left(r_i - \frac{\alpha_i + \delta_i}{2\delta_i}\right)^2\right] \times \exp\left[\frac{(\alpha_i+\delta_i)^2}{4\delta_i}\right]. \tag{7.25}$$

The second factor in the right-hand side of (7.25) is independent of r_i and will cancel when we substitute the right-hand side of (7.25) in (7.24). Defining

$$\mu_i(\theta) = \frac{\alpha_i + \delta_i}{2\delta_i} \text{ and } \sigma_i^2 = \frac{1}{2\delta_i},$$

(7.24) can be written as

[4] The '*' in Eq. (7.22) is introduced to avoid the suggestion that parameters and variables in the discrete and the continuous model are identical.

[5] In the derivation of the rating scale model, Andrich assumes that a response in category j means that the j most left positioned thresholds out of m have been passed and the $(m - j)$ rightmost ones not. The location parameter η_i is the midpoint of the m thresholds.

$$f_i(r_i|\theta) = \frac{\sigma_i^{-1}\varphi\left[\frac{r-\mu_i(\theta)}{\sigma_i}\right]}{\Phi\left[\frac{1-\mu_i(\theta)}{\sigma_i}\right] - \Phi\left[\frac{-\mu_i(\theta)}{\sigma_i}\right]} \tag{7.26}$$

with $\varphi(.)$ and $\Phi(.)$ denoting the standard normal density and probability functions, respectively. One easily recognizes (7.26) as the probability density function of the truncated normal distribution (Johnson and Kotz 1970). The regression of the response on the latent variable and the variance of the response are given next. Using D_i as shorthand for the denominator of (7.26), i.e.,

$$D_i = \Phi\left[\frac{1-\mu_i(\theta)}{\sigma_i}\right] - \Phi\left[\frac{-\mu_i(\theta)}{\sigma_i}\right],$$

and

$$z_{0i} = \frac{-\mu_i(\theta)}{\sigma_i} \text{ and } z_{1i} = \frac{1-\mu_i(\theta)}{\sigma_i},$$

the regression function is given by

$$E(R_i|\theta) = \mu_i(\theta) + \frac{\varphi(z_{0i}) - \varphi(z_{1i})}{D_i}\sigma_i,$$

and the variance by

$$Var(R_i|\theta) = \sigma_i^2\left[1 + \frac{z_{0i}\varphi(z_{0i}) - z_{1i}\varphi(z_{1i})}{D_i} - \left(\frac{\varphi(z_{0i}) - \varphi(z_{1i})}{D_i}\right)^2\right].$$

The regression function is displayed in the left panel of Fig. 7.2 for three different values of the δ-parameter. Some comments are in order here:

1. The computation of the regression and variance is tricky for extreme values of $z_{0\mathrm{i}}$ and $z_{1\mathrm{i}}$. Using the standard implementation of the normal probability function

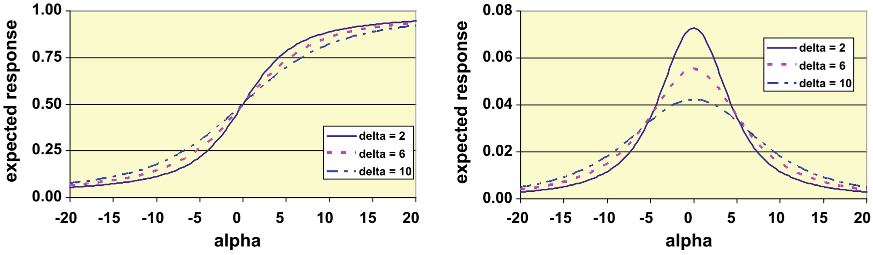


Fig. 7.2 Expected value (left) and variance (right) of R_i as a function of $\alpha_i = \theta - \eta_i$ in Müller's model

in EXCEL or R generates gross errors. Excellent approximations are given in Feller (1957, p. 193).

2. Three values of the δ-parameter have been used, resulting in flatter curves for the higher values. Therefore, the parameter δ can be interpreted as a discrimination parameter, lower values yielding higher discrimination. A discussion on the model properties when $\delta \to 0$ can be found in Müller (1987).
3. Just as in the Rasch model for continuous responses, the numbers along the horizontal axis are quite different from the ones usually displayed for the item response functions of the common Rasch model. Further comments will be given in Sect. 7.4.

The right-hand panel of Fig. 7.2 displays the variance of the response function for the same three values of δ and for the same values of α. The figures are symmetric around the vertical axis at $\alpha = 0$. As this model is an exponential family, the variance of the response is also the Fisher information.

To gain more insight in the information function, one can use a reparameterization of the rating scale model, where the scores are brought back to the unit interval, i.e., by dividing the original integer valued scores by m_i, the original maximum score. Müller shows that this results in the discrete model with category response function[6]

$$f_i^*(r_i|\theta) = \frac{\exp[r_i\alpha_i + r_i(1-r_i)\delta_i]}{\sum\limits_{j=0}^{m}\exp\big[r_j\alpha_i + r_j(1-r_j)\big]}, \quad \left(r_i = 0, \frac{1}{m}, \frac{2}{m}, \dots, 1\right). \tag{7.27}$$

It is easy to see that (7.24) is the limiting case of (7.27) as $m \to \infty$. In Fig. 7.3 the information functions are plotted for some finite values of m and for the limiting case, labeled as ‘continuous’. In the left-hand panel, the discrimination parameter δ_i equals 2 (high discrimination), in the right-hand panel it equals 10 (low discrimination); the value of η_i is zero in both cases such that $\alpha_i = \theta$.

The collection of curves can be described briefly as follows:

1. the maxima of the curves are lower with increasing values of m;
2. the tails are thicker with increasing values of m;

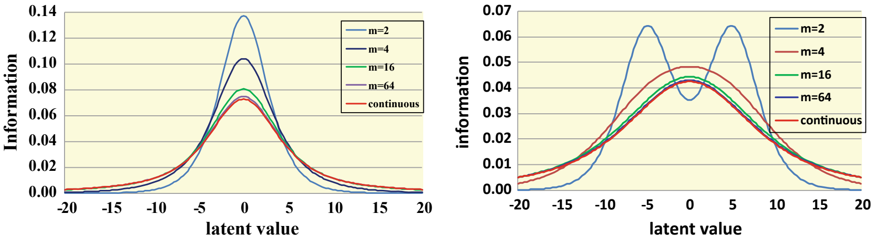


Fig. 7.3 Information functions for different values of m (left: $\delta_i = 2$; right: $\delta_i = 10$)

[6] In Müller’s article the admitted values for r_i are slightly different, but the difference is not important.

3. for low values of m the curves are not necessarily unimodal: see the case for $m = 2$ en $\delta_i = 10$;
4. for $m = 64$ and $m \to \infty$, the curves are barely distinguishable.

One should be careful with conclusions drawn from the graphs in Fig. 7.3: it does not follow, for example, from this figure that, if one switches from a discrete format of an item to a continuous one, that the discrimination or the location parameter will remain invariant. An example will be given in Sect. 7.4.

7.3.2 Parameter Estimation

Müller (1987) proposes the parameter estimation using only pairs of items, but does not give details. In the present section the method is explained; the technical expressions to compute the gradient and the matrix of second partial derivatives are given in the Appendix to this chapter.

As with the continuous Rasch model, definition (7.9), $r_{ij}= r_i+ r_j$ is used and the reparameterization

$$\varepsilon_i = \delta_i - \eta_i .$$

is used. The problems with obtaining CML estimates of the parameters are the same as with the Rasch model for continuous responses (and augmented with numerical problems in evaluating the probability function of the truncated normal distribution for extreme values of the argument). Therefore the pseudo-likelihood function, considering all pairs of items, is studied here.

The conditional likelihood function for one pair of items is given by

$$f_{ij}(r_i, r_j | r_{ij}) = \frac{\exp(r_i\varepsilon_i + r_j\varepsilon_j - r_i^2\delta_i - r_j^2\delta_j)}{\int_{m_{ij}}^{M_{ij}} \exp\left[t\varepsilon_i + (r_{ij} - t)\varepsilon_j - t^2\delta_i - (r_{ij} - t)^2\delta_j)\right] dt} \tag{7.28}$$

where the bounds of the integral in the denominator depend on the value of r_{ij}:

$$\begin{aligned} m_{ij} &= \max(r_{ij} - 1, 0), \\ M_{ij} &= \min(r_{ij}, 1). \end{aligned}$$

Along the same lines of reasoning as followed in Sect. 3.1, (7.28) can also be shown to be a density of the truncated normal distribution, i.e.,

$$f_{ij}(r_i, r_j | r_{ij}) = \frac{\exp\left[-\frac{(r_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right]}{\int_{m_{ij}}^{M_{ij}} \exp\left[-\frac{(t - \mu_{ij})^2}{2\sigma_{ij}^2}\right] dt} = \frac{\sigma_{ij}^{-1}\varphi(x_{ij})}{\Phi(Z_{ij}) - \Phi(z_{ij})}, \tag{7.29}$$

where

$$\begin{aligned}
\sigma_{ij}^2 &= \frac{1}{2(\delta_i + \delta_j)},\\
\mu_{ij} &= \sigma_{ij}^2(\varepsilon_i - \varepsilon_j + 2r_{ij}\delta_j),\\
x_{ij} &= (r_i - \mu_{ij})/\sigma_{ij},\\
Z_{ij} &= (M_{ij} - \mu_{ij})/\sigma_{ij},\\
z_{ij} &= (m_{ij} - \mu_{ij})/\sigma_{ij}.
\end{aligned}$$

Although the function μ_{ij} is not symmetric in i and j, it follows readily that $\mu_{ij}+ \mu_{ji} = r_{ij}= r_{ji}$. Using this, it is not hard to show that the conditional response density (7.29) does not change if i and j are interchanged.

The function to be maximized is the logarithm of the pseudo-likelihood function:

$$\sum_v \ln\, PL_v[\varepsilon, \delta; (r_{v1}, \ldots, r_{vk})] = \sum_v \sum_{i<j} \ln\, f_{ij}\big(r_{vi}, r_{vj} \big| r_{vij}\big), \qquad (7.30)$$

where ε and δ denote the vectors of ε- and δ-parameters, r_{vi} is the response of respondent v to item i and k is the total number of items. Deriving the expressions for the first and second partial derivatives of (7.30) is not difficult, but rather tedious, the expressions are given in the Appendix to this chapter.

7.4 Comparison of Information Functions Across Models

7.4.1 The Unit of the Latent Variable

Müller (1987, p. 173) compares the information functions for the rating scale model and his own model for continuous data, assuming that the latent variable underlying the two models is the same, but this assumption needs not to be correct for two quite different reasons.

1. The first and most important threat to a meaningful comparison is the silent assumption that the same trait is measured because the items are the same; only the item format is different. For example, in a personality inventory, in the discrete case the respondent may be asked to express the applicability of a statement to him/herself on a discrete 5-point scale where each possible answer is labeled by descriptions as ‘certainly not applicable’, ‘not applicable’, ‘neutral’ (whatever this may mean), ‘applicable’ and ‘certainly applicable’, while in the continuous case the respondent is asked to express the applicability of the statement by putting a mark on a piece of line where only the two end points are labeled, as (e.g.,) ‘not applicable at all’ and ‘strongly applicable’. In the discrete case tendencies to the middle may strongly influence to prefer the middle category,

while this tendency may be a less important determinant in the continuous case. In principle there is only one way to have good answers on this important question of construct validity: one has to collect data under both item formats, and estimate the correlation (disattenuated for unreliability) between the traits measured using either of the two formats.

2. But even if one can show that the traits measured using different item formats are the same (i.e., their latent correlation equals one), there is still the problem of the origin and the unit of the scale. It is easily understood that the origin of the scale is arbitrary, as most of the IRT models define their item response functions using a difference between the trait value and a location parameter, such that adding an arbitrary constant to both does not change response probabilities (or densities as is easily seen from (7.1)). But understanding that the unit of the scale is arbitrary as well is not so easy. A historic example of this is the claim in Fischer (1974) that the scale identified by the Rasch model is a 'difference scale', a scale where the origin is arbitrary but the unit fixed, while the same author (Fischer 1995) came, in a complicated chapter, to the conclusion that the scale measured by the Rasch model is an interval scale, with arbitrary origin and unit. In the common Rasch model the unit is chosen by fixing the discrimination parameter (common to all items) at the value of 1, but any other positive value may be chosen. Suppose that in the discrete model, one chooses $c \neq 1$, then one can replace $\theta - \beta_i$ with $c(\theta^* - \beta_i^*)$ where $\theta^* = \theta/c$ and $\beta_i^* = \beta_i/c$ and c the common (arbitrarily chosen) discrimination value.

With the continuous models, however, the choice of the unit of measurement (of the latent variable) is also influenced by the bounds of the integral. We illustrate this with the Rasch model for continuous responses. Suppose data are collected by instructing the respondents to distribute M tokens (with M large enough such that the answer can safely be considered as continuous) among the alternatives of a multiple choice question and let U_i (with realizations u_i) be the number of tokens assigned to the correct alternative of item i. Then, the model that is equivalent to (7.1) is given by

$$f_i(u|\theta^*) = \frac{\exp\left[u(\theta^* - \eta_i^*)\right]}{\int_0^M \exp\left[y(\theta^* - \eta_i^*)\right]\mathrm{d}y}, \quad (u \in [0, M]).$$

If we want to rescale the response variable by changing it to a relative value, i.e., $R_i = U_i/M$, then we find

$$\begin{aligned} f_i(r_i|\theta^*) &= \frac{\exp\left[r_i M(\theta^* - \eta_i^*)\right]}{M\int_0^1 \exp\left[tM(\theta^* - \eta_i^*)\right]\mathrm{d}t} \\ &= \frac{\exp[r_i(\theta - \eta_i)]}{M\int_0^1 \exp[t(\theta - \eta_i)]\mathrm{d}t} \\ &= \frac{1}{M} f_i(y_i|\theta), \quad (y_i = r_i/M \in [0, 1]) \end{aligned}$$

with $\theta = M\theta^*$ and $\eta_i = M\eta_i^*$. So we see that changing the bounds of the integration in (7.1) changes the density by a constant and at the same time the unit of the underlying scale. Changing the density is not important as it will only affect the likelihood by a constant, but at the same time the unit of the underlying scale is changed. Choosing zero and one as the integration bounds in (7.1) is an arbitrary choice, and therefore the unit of measurement is arbitrary as well.

Exactly the same reasoning holds for Müller's model where in Eq. (7.23) the constants c and d (>0) are arbitrary, but will influence the unit of measurement of the latent variable.

7.4.2 An Example

A simple example to see how the information function depends on the unit of measurement is provided by the Rasch model for binary data. Assuming, as in the previous section, that the common discrimination parameter is indicated by c, the information function for a single item is given by

$$I_i(\theta) = c^2 f_i(\theta)[1 - f_i(\theta)],$$

meaning that doubling c will quadruple the information, so that this gives the impression that the information measure is arbitrary. But one should keep in mind that doubling the value of c will at the same time halve the standard deviation (SD) of the distribution of θ or divide its variance by a factor four. This means that the information measure can only have meaning when compared to the variance of the θ-distribution. So, if we want to compare information functions across models, we must make sure that the latent variables measured by the models are congeneric, i.e., their correlation must be one, but at the same time they must be identical, i.e., having the same mean and variance. This is not easy to show empirically, but we can have an acceptable approximation as will be explained by the following example.

At the department of Psychology of the University of Amsterdam, all freshmen participate (compulsorily) in the so-called test week[7]: during one week they fill in a number of tests and questionnaires and take part as subjects in experiments run at the department. One of the questionnaires presented is the Adjective Check List (ACL, Gough & Heilbrunn, 1983; translated into Dutch by Hendriks et al. 1985), where a number of adjectives (363) is presented to the test takers. The task for the student is to indicate the degree to which each adjective applies to him/herself. The standard administration of the test asks the students to indicate the degree of applicability on a five point scale. In 1990, however, two different test forms were administered, each to about half of the students. In the first format, only a binary response was asked for (not applicable/applicable); in the second format, the student was asked to mark the degree of applicability on a line of 5.7 cm, with the left end corresponding to 'not

[7] At least, this was the case until 1990, the year where the continuous data were collected.

applicable at all' and the right end to 'completely applicable'. The score obtained by a student is the number of millimeters (m) of the mark from the left end of the line segment. This number m was transformed into a response (in model terms) by the transformation

$$r = \frac{m + 0.5}{58}.$$

Seven of these adjectives, fitting in the framework of the Big Five, were taken as relevant for the trait 'Extraversion' and these seven are used in the present example. For the polytomous data, the responses collected in the years 1991 through 1993 were pooled, yielding 1064 records. For the continuous data 237 records were available. Dichotomous data are not used here. All these records are completely filled in; some 2% of the original records with one or more omissions have been removed. The polytomous data have been analyzed with the Partial Credit Model (PCM) using CML estimation for the category parameters for ordered categories running from 0 to 4; the continuous data were analyzed with the extension of Müller's model as discussed in Sect. 7.3, using the maximum pseudo-likelihood method. After the estimation of the item parameters, the mean and SD of the θ-distribution were estimated, while keeping the item parameters constant at their previous estimates. For both models a normal distribution of the latent variable was assumed.

Parameter estimates are displayed in Table 7.1. The adjectives were presented to the students in Dutch; the English translation is only put there as information and did not have any influence at all on the actual answers. As there are four parameters per item in the polytomous model, and their estimates are of not much use in judging the item locations, they are not reported in Table 7.1. Instead, the value of the latent variable yielding two (half of the maximum) as expected value is reported, under the symbol β^*.

Table 7.1 Parameter estimates of the 'extraversion' scale

Adjective		Continuous		Polytomous
Dutch	English	η	δ	β^*
Extravert	(Extravert)	1.932	11.224	0.610
gereserveerd[a]	(Reserved)	−1.573	12.007	−0.176
praatgraag	(Talkative)	0.024	8.584	−0.130
terughoudend[a]	(Aloof)	−1.382	12.437	−0.032
verlegen[a]	(Shy)	1.601	7.661	0.298
zwijgzaam[a]	(Taciturn)	−0.734	8.160	−0.298
introvert[a]	(Introvert)	0.132	4.943	0.189
Mean		2.5004		1.1416
SD		5.2180		0.7067

[*] for these adjectives, the complementary scores were used

As is seen from the table the estimated SD for the continuous model is much larger than for the PCM model, and this is consistent with the large numbers along the horizontal axis in Fig. 7.2. The correlation between the η- and β^*-parameters is 0.85.

The latent variable estimated by the PCM will be indicated by θ_p; the one estimated by the continuous model by θ_c. Their means and SDs are indicated by the subscripts 'p' and 'c' as well. The variable θ_p will be linearly transformed such that the transformed variable has the same mean en SD as θ_c, i.e.,

$$T(\theta_p) = B\theta_p + A.$$

It is easy to check that

$$B = \frac{\sigma_c}{\sigma_p} \text{ and } A = \mu_c - B\mu_p$$

In Fig. 7.4, the two information functions for the seven item checklist are given, for values of the latent variable brought to the same scale where the population mean is 2.5 and the SD about 5.22. The vertical dashed line indicates the average of the latent value distribution, and the thick black line indicates the range that encompasses 95% of the distribution. So, for the great majority of the population the information provided by the continuous model is larger than for the discrete PCM.

Of course the procedure we followed can be criticized: no account has been given to the standard errors of the estimates, and the data do not come from a sample that has been tested twice. So, silently, we have assumed that the distribution of the latent trait has not changed (very much) in the period that the data were collected. As in all these cases the population consists of Psychology freshman at the University of Amsterdam, it is fairly unlikely that the distribution has changed in any important way.

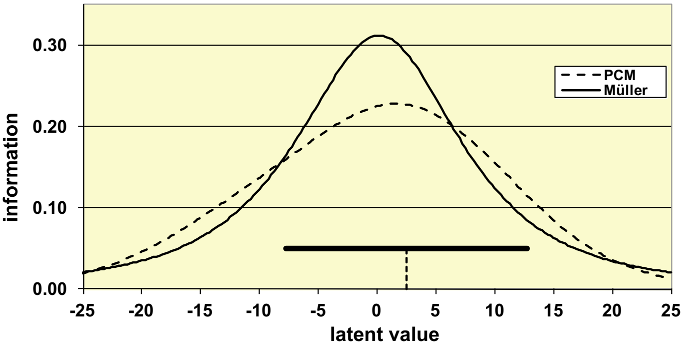


Fig. 7.4 Information functions for the PCM and for Müller's model for the same adjective checklist

Even with all the criticism that may be put forward, it seems quite convincing that the continuous model is more informative than the PCM and therefore certainly deserves more attention than it has received thus far.

7.5 Discussion

Two IRT models for continuous responses that allow for separation between item and person parameters have been studied. Although CML estimation of the item parameters is possible, it leads to unwieldy formulae and only the case for two items has been considered for the two models. This, however, proves to be sufficient to obtain consistent estimates of the item parameters, using the theoretical contributions in the area of pseudo-likelihood estimation.

Using the technical details provided in this chapter, developing appropriate software for parameter estimation is not hard. For the extension of Müller's model, the software developed applies the Newton-Raphson procedure directly on the initial estimates and works very well, although, admittedly, it was only used on a scale with no more than 7 items. For the extension of the Rasch model to continuous responses, software has to be developed yet.

But there are more aspects in the technical realm that have to be given attention to:

1. The estimation of the individual person parameter, not only as a technical problem, but also as a real statistical one: is the maximum likelihood estimator of θ, used with known values of the item parameters, unbiased, and if it is not, can one develop an unbiased estimator along the lines of the Weighted Maximum Likelihood or Warm-estimator?
2. The set-up of goodness-of-fit tests which are informative, e.g., for the contrast between the two models discussed here, but also techniques to detect non-fitting items, all based on sound statistical reasoning are areas which deserve attention.
3. The extension of Müller's model to allow for different discriminations turned out to be rather simple, and as the example in Table 7.1 shows, worthwhile as the estimates differ considerably. Therefore, it also seems worthwhile to think about an extension of the Rasch model for continuous items that allows for different discriminations. A promising route would be to explore the possibilities of Haberman's (2007) interaction model.

There are, however, aspects in testing that cannot be solved by sophisticated techniques but which are fundamental in the discussion of the validity of the test and the conclusion its use leads to. Take the small scale on extraversion which was discussed in Sect. 7.4. If the answers are collected using a Likert scale, there is ample evidence in the literature for certain tendencies (like the tendency to avoid extreme answers) which will create irrelevant variance in the test scores. Suppose we have two respondents who are quite extravert, i.e., who would on average score at or above the middle of the five categories, but one, A, has a strong tendency to avoid extreme answers while the other, B, has a preference for extreme answers, then on the average B will obtain a higher test score than A, while the difference cannot be unambiguously be attributed to a difference in extraversion, i.e., some of the variance of the observed scores is due to a variable which is irrelevant for the trait to be studied, and therefore forms a threat to the correct interpretation of the results, i.e., to the validity.

More in general, there are always determinants of behavior which are responsible for part of the variance in the responses, and which are a kind of a nuisance in the interpretation of the test scores, but which cannot be easily avoided. If these extra variables are associated with the format of the items then they are hard to discover if the use of this format is ubiquitous, like the multiple choice (MC) format. It is highly probable that ‘blind guessing’ as a much used model for explaining the behavior in MC test is highly unrealistic; students use all kinds of strategies to improve the result when they are not sure of the correct response, and some of these strategies will give better results than others, so that the result on an MC test is a contamination of the intended construct and the cleverness in using choice strategies.

As long as there is no variation in the format of the items, this contamination will not do much harm, as the irrelevant construct will be absorbed into the intended construct. But the risk of incorrect interpretations occurs if at some point a drastic change in the format is introduced, like switching from MC to probability measurement.

When Dirkzwager was developing his system of multiple evaluation—the name he gave to probability measurement—he was aware of a serious validity threat: students can show a lack of ‘realism’ when assigning continuous responses to each of the alternatives of a multiple choice question, either by being overconfident and giving too high a weight to one of the alternatives, or by being cautious and tending to a uniform distribution over the alternatives. here is a quotation from Holmes (2002, p. 48): *Considering one item in a test, we can represent the student’s uncertainty as to the correct answer by a probability distribution* $p = (p_1, p_2,\ldots,p_k)$ *over the set* $\{1, 2,\ldots,k\}$ *where* $0 \leq p_j$ *and* $\Sigma p_l = 1$. *The student’s response can be represented by* $r = (r_1, r_2,\ldots r_k)$. *For a perfectly realistic student the response r is equal to p. In effect, such a student is stating: “Given my knowledge of the subject, this item is one of the many items for which my personal probability is* $p = (p_1, p_2,\ldots,p_k)$. *For these items, answer one is correct in a proportion*[8] *of* p_1, *answer 2 in a proportion of* p_2 *etcetera.*

This means that a student is realistic if his continuous responses match his subjective probabilities, which is already elicited by the (approximate) logarithmic scoring function, but for which Holmes developed a measure of realism (see his Chap. 4) and showed that using this measure as feedback was very effective in changing the behavior of ‘unrealistic’ students, which makes the multiple evaluation approach as developed originally by Dirkzwager a good and powerful instrument for formative assessment. Details of the application can be found in Holmes.

Finally, one might ask then why an IRT model is needed if good use can be made of probability measurement. There are, however aspects of formative assessment

[8] The quotation was a bit changed here as Holmes spoke of p_1 cases, clearly confusing frequencies and proportions.

which are very hard to develop at the school level. Progress assessment, for example, is an important one, but also the choice of items whose difficulty matches the level of the students. The construction of an item bank at the national level, for example, and making it available to the schools together with the necessary software, could be a task for a testing agency having the necessary psychometric and IT competences.

Appendix

To estimate the parameters in Müller's model the logarithm of the pseudo-likelihood function is maximized. The function, given in (7.30) is repeated here:

$$\sum_v \ln\, PL_v[\varepsilon, \eta; (r_{v1}, \ldots, r_{vk})] = \sum_v \sum_{i<j} \ln\, f_{ij}(r_{vi}, r_{vj} | r_{vij}),$$

where v indexes the respondent. Concentrating on a single answer pattern, and dropping the index v, we have

$$f_{ij}(r_i, r_j | r_{ij}) = \frac{\sigma_{ij}^{-1} \varphi(x_{ij})}{\Phi(Z_{ij}) - \Phi(z_{ij})} \tag{7.31}$$

with the auxiliary variables

$$\begin{aligned}
\sigma_{ij}^2 &= \frac{1}{2(\delta_i + \delta_j)},\\
\mu_{ij} &= \sigma_{ij}^2(\varepsilon_i - \varepsilon_j + 2r_{ij}\delta_j),\\
x_{ij} &= (r_i - \mu_{ij})/\sigma_{ij},\\
Z_{ij} &= (M_{ij} - \mu_{ij})/\sigma_{ij},\\
z_{ij} &= (m_{ij} - \mu_{ij})/\sigma_{ij},
\end{aligned} \tag{7.32}$$

and the two bounds, M_{ij} and m_{ij}, repeated here:

$$\begin{aligned}
m_{ij} &= \max(r_{ij} - 1, 0),\\
M_{ij} &= \min(r_{ij}, 1).
\end{aligned}$$

Taking the logarithm of (7.31) gives

$$\ln\, f_{ij}(r_i, r_j | r_{ij}) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_{ij}^2) - \frac{1}{2}x_{ij}^2 - \ln\big[\Phi(Z_{ij}) - \Phi(z_{ij})\big]. \tag{7.33}$$

To write down the expressions for the partial derivatives w.r.t. the ε- and δ-parameters, it proves useful to define a sequence of functions A_{pij}, ($p = 0, 1, 2\ldots$):

$$A_{pij} = \frac{Z_{ij}^p \varphi(Z_{ij}) - z_{ij}^p \varphi(z_{ij})}{\Phi(Z_{ij}) - \Phi(z_{ij})}. \tag{7.34}$$

Using the partial derivatives of the auxiliary variables (7.31) and the functions defined by (7.33) the first partial derivatives of (7.32) are given by

$$\begin{aligned}
\frac{\partial\ \ln(f_{ij})}{\partial \varepsilon_i} &= r_i - \mu_{ij} + \sigma_{ij} A_{0ij}, \\
\frac{\partial\ \ln(f_{ij})}{\partial \varepsilon_j} &= r_j - r_{ij} + \mu_{ij} - \sigma_{ij} A_{0ij}, \\
\frac{\partial\ \ln(f_{ij})}{\partial \delta_i} &= r_i^2 + \mu_{ij}^2 + \sigma_{ij}^2(1 - A_{1ij}) - 2\mu_{ij}\sigma_{ij}A_{0ij}, \\
\frac{\partial\ \ln(f_{ij})}{\partial \delta_j} &= -r_j^2 + (r_{ij} - \mu_{ij})^2 + \sigma_{ij}^2(1 - A_{1ij}) + 2(r_{ij} - \mu_{ij})\sigma_{ij}A_{0ij}. \qquad (7.35)
\end{aligned}$$

Notice that $\frac{\partial\ \ln(f_{ij})}{\partial\, \varepsilon_j} = -\frac{\partial\ \ln(f_{ij})}{\partial\, \varepsilon_i}$ and that

$$\frac{\partial\ \ln(f_{ij})}{\partial\, \delta_j} = \frac{\partial\ \ln(f_{ij})}{\partial\, \delta_i} + 2r_{ij}\frac{\partial\ \ln(f_{ij})}{\partial\, \varepsilon_i}. \tag{7.36}$$

For the second derivatives, it turns out that we only need three different expressions; these are given next, but we leave out the double subscript '*ij*' from the right-hand sides.

$$\frac{\partial^2\ \ln(f_{ij})}{\partial\, \varepsilon_i^2} = -\sigma^2\big[1 - A_1 - A_0^2\big], \tag{7.36a}$$

$$\frac{\partial^2\ \ln(f_{ij})}{\partial\, \varepsilon_i\, \partial\, \delta_i} = 2\mu\sigma^2\big[1 - A_1 - A_0^2\big] - \sigma^3[A_0 + A_2 + A_0A_1], \tag{7.36b}$$

$$\begin{aligned}
\frac{\partial^2\ \ln(f_{ij})}{\partial \delta_i^2} &= -\sigma^2[(1 - A_1)(2 + A_1) - A_3] + 4\mu\sigma^3[A_0 + A_2 + A_0A_1] \\
&\quad - 4\mu^2\sigma^2\big[1 - A_1 - A_0^2\big]. \qquad (7.36c)
\end{aligned}$$

It is easily verified that

$$\frac{\partial^2\ \ln(f_{ij})}{\partial\, \varepsilon_i^2} = \frac{\partial^2\ \ln(f_{ij})}{\partial\, \varepsilon_j^2} = -\frac{\partial^2\ \ln(f_{ij})}{\partial\, \varepsilon_i\, \partial\, \varepsilon_j}$$

and using (7.35), it turns out that we only need the three expressions (7.35a), (7.35b) and (7.35c) to define in a simple way the matrix of second partial derivatives. Using the symbols '*a*', '*b*' and '*c*' to denote the value of the right-hand members of (7.35a), (7.35b) and (7.35c), respectively, one obtains the matrix of second partial derivatives as displayed in Table 7.2.

Table 7.2 Symbolic representation of the matrix of second derivatives (r means r_{ij})

	ε_i	ε_j	δ_i	δ_j
ε_i	a	$-a$	b	$b + 2ra$
ε_j	$-a$	a	$-b$	$-b - 2ra$
δ_i	b	$-b$	c	$c + 2rb$
δ_j	$b + 2ra$	$-b - 2ra$	$c + 2rb$	$c + 4rb + 4r^2a$

In the applications that were run for this chapter (see Sect. 7.4), simple initial values for the parameters were computed and immediately used in a Newton-Raphson procedure. The initial values were

$$\delta_i^{[0]} = \frac{0.1}{Var(R_i)} \text{and } \varepsilon_i^{[0]} = 2\delta_i^{[0]}\overline{R}_i$$

where $\overline{R}_i$ and $Var(R_i)$ denote the average and the variance of the observed responses to item i, respectively. To make the model identified, one of the ε-parameters can be fixed to an arbitrary value. To avoid negative estimates of the δ-parameters, it is advisable to reparametrize the model for estimation purposes and to use $\ln(\delta_i)$ instead of δ_i itself.

References

Arnold, B. C., & Strauss, D. (1988). *Pseudolikelihood estimation* (Technical Report 164). Riversdale: University of California.

Arnold, B. C., & Strauss, D. (1991). *Pseudolikelihood estimation: Some examples Sankhya, 53,* 233–243.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics, 17,* 251–269.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47,* 105–113.

Bejar, I. I. (1977). An application of the continuous response model to personality measurement. *Applied Psychological Measurement, 1,* 509–521.

Cox, D. R., & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika, 91,* 729–737.

De Finetti, B. (1965). Methods of discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology, 54,* 87–123.

De Finetti, B. (1970). Logical foundations and the measurement of subjective probabilities. *Acta Psychologica, 34,* 129–145.

Dirkzwager, A. (1997). *A bayesian testing paradigm: Multiple evaluation, a feasible alternative for multiple choice* (Unpublished Report, to be found in Dirkzwager, 2001).

Dirkzwager, A. (2001). *TestBet, learning by testing according to the multiple evaluation paradigm: Program, manual, founding articles* (CD-ROM). ISBN:90-806315-2-3.

Feller, W. (1957). *An introduction to probability theory and its applications* (Vol. 1) New York: Wiley.

Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research, 37,* 521–542.

Finney, D. J. (1978). *Statistical methods in biological assay*. London: Charles Griffin and Co.

Fischer, G. H. (1974). *Einführung in die Theorie Psychologischer Tests*. Bern: Huber.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 15–38). New York: Springer.

Gough, H. G., & Heilbrun, A. B. Jr. (1983). *The adjective checklist manual*. Palo Alto, CA: Consulting Psychologists Press.

Haberman, S. J. (2007). The interaction model. In M. von Davier & C. C. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 201–216). New York: Springer.

Hendriks, C., Meiland, F., Bakker, M., & Loos, I. (1985). *Eenzaamheid en Persoonlijkheidskenmerken* [*Loneliness and Personality Traits*] (Internal publication). Amsterdam: Universiteit van Amsterdam, Faculteit der Psychologie.

Holmes, P. (2002). *Multiple evaluation versus multiple choice as testing paradigm* (Doctoral Thesis). Enschede: University of Twente. Downloaded from: https://ris.utwente.nl/ws/portalfiles/portal/6073340/t0000017.pdf.

Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics. Vol. 2: Continuous univariate distributions-1*, (Chap. 13). New York: Wiley.

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

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29,* 223–236.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika, 52,* 165–181.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–168). New York: Springer.

Noel, Y. (2014). A beta unfolding model for continuous bounded responses. *Psychometrika, 79,* 647–674.

Noel, Y. (2017). *Item response models for continuous bounded responses* (Doctoral Dissertation). Rennes: Université de Bretagne, Rennes 2.

Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement, 31,* 47–73.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rippey, R. M. (1970). A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement, 7,* 165–170.

Roby, T. B. (1965). *Belief states: A preliminary empirical study* (Report ESD-TDR-64-238). Bedford, MA: Decision Sciences Laboratory.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. In *Psychometrika Monograph*, No 17.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203–219.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39,* 111–121.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Shuford, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika, 31,* 125–145.

Toda, M. (1963). *Measurement of subjective probability distributions* (Report ESD-TDR-63-407). Bedford, Mass: Decision Sciences Laboratory.

Van Naerssen, R. F. (1961). A method for the measurement of subjective probability. *Acta Psychologica, 20,* 159–166.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 8
Tracking Ability: Defining Trackers for Measuring Educational Progress

Matthieu J. S. Brinkhuis and Gunter Maris

Abstract In measurements that extend over longer periods of time, we might expect all sorts of changes in model parameters, such as ability and item difficulty. We define trackers as instruments with specific properties to deal with such changing parameters. First, trackers should allow for estimating dynamically changing parameters, adapting to possible changes in ability or item difficulty. Second, if no change occurs for some time, trackers should provide unbiased estimates with a known error variance. Such trackers retain the strengths of both state space models and rating systems, while resolve some of their weaknesses. These properties are especially suitable for educational measurement applications such as tracking individual progress or any aggregate thereof, as in reporting survey research.

8.1 Introduction

In this chapter, we describe trackers and their uses in educational measurement. For now, we loosely define trackers as dynamic parameter estimates, adapting to possible changes in ability or item difficulty. Trackers can be especially useful in measurements that extend over a longer period of time at irregular time intervals, e.g., the continual measurement of abilities in computer adaptive practice (CAP) or computer adaptive learning (CAL) (Brinkhuis et al. 2018; Klinkenberg et al. 2011; Wauters et al. 2010; Veldkamp et al. 2011) or the monitoring of item difficulties in item banks (Brinkhuis et al. 2015). Many actors can be involved in the possible changes of these parameters, including the pupils themselves, their teachers, their parents, educational reforms, etc. Moreover, change in parameters, and the model

M. J. S. Brinkhuis (✉)
Department of Information and Computing Sciences,
Utrecht University, Utrecht, The Netherlands
e-mail: m.j.s.brinkhuis@uu.nl

G. Maris
ACTNext, Iowa City, IA, USA
e-mail: marisg@act.org

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_8

itself, is especially likely if the outcomes of the measurements are used for feedback, as in assessment for learning (Black and Wiliam 2003; Bennett 2011; Wiliam 2011). Since feedback is provided to many actors in education, the result is a complex dynamical system, including all sorts of interactions.

The development of these parameters is not easily modeled due to these changes and feedback loops. Application of latent growth models (McArdle and Epstein 1987; Meredith and Tisak 1990; Hox 2002), change point estimation models (Hinkley 1970; Chib 1998; Visser et al. 2009) or other models that explicitly model the development of parameters is therefore not straightforward. Also state space models such as the Kalman filter (KF) (Kalman 1960; Welch and Bishop 1995; van Rijn 2008), or the more general particle filters (Arulampalam et al. 2002), include an explicit growth model and are therefore not ideal for following educational progress continually, as in CAP systems.

Historically, we see that in other fields where continual progress measurements take place, rating systems emerged. For example, in chess, rating systems were developed for the estimation of continually changing chess playing abilities, such as the widely used Elo rating system (ERS) (Elo 1978; Batchelder and Bershad 1979; Batchelder et al. 1992). Advantages of rating systems such as Elo's are that they are computationally light and do not assume a growth model. After each new measurement, the parameter estimates can be updated using the previous parameter estimate and the new observation, without having to take history into account, i.e., satisfying the Markov property, nor assuming a model for the development of the parameters. The ERS has found many uses, including applications in educational measurement (Klinkenberg et al. 2011; Wauters et al. 2010; Pelánek et al. 2017). Though there are many practical uses of rating systems, they lack certain desirable statistical properties such as convergence and unbiasedness (Brinkhuis and Maris 2009).

In tracking educational progress, we are interested in several properties from both state space models such as KFs and rating systems like the ERS, i.e., we like to dynamically track changes in the abilities of individuals or in item difficulties, without having to assume specific types of development. Moreover we require that, if ability is stable for some time, our tracker provides unbiased estimates with a known error variance. In this chapter, we will describe a tracker with these properties.

8.2 Methods

8.2.1 Formalizing a Tracker

We formalize the representation of a tracker in the scheme in (8.1), where we illustrate the development of someone's unknown true ability θ over time t, where each column represents a consecutive time point:

$$\begin{tikzcd}
\text{ability} & & \theta_1 \arrow[r] \arrow[d] & \theta_2 \arrow[r] \arrow[d] & \theta_3 \arrow[r] \arrow[d] & \cdots \arrow[r] & \theta_t \arrow[d] \\
\text{responses} & & Y_1 \arrow[d] & Y_2 \arrow[d] & Y_3 \arrow[d] & \cdots & Y_t \arrow[d] \\
\text{estimates} & X_0 \arrow[r] & X_1 \arrow[r] & X_2 \arrow[r] & X_3 \arrow[r] & \cdots \arrow[r] & X_t
\end{tikzcd} \tag{8.1}$$

The abilities θ are related by horizontal arrows, since we assume that one's true ability at time point t is related at least to one's ability at time point $t-1$ and likely influenced by many other factors, which we leave out of this scheme. At time point t, scored responses Y_t are obtained using a single item, or a number of items. The ability estimate X_t depends only on the previous state X_{t-1} and the current item response Y_t, therefore satisfying the Markov property. The scheme in (8.1) represents Markov chains in general, including the ERS.

Since we are especially interested in the properties of unbiasedness and convergence, we present a more specific scheme in (8.2). Here, we assume for the moment that someone's ability does *not* change, i.e., $\theta_t = \theta \; \forall \; t$, and we require X_∞ to have a known distribution, for example centered around the true ability θ with normal distributed error $\mathcal{E}$:

$$\begin{tikzcd}
\text{ability} & & \theta \arrow[r] \arrow[d] & \theta \arrow[r] \arrow[d] & \theta \arrow[r] \arrow[d] & \cdots \arrow[r] & \theta \arrow[d] \\
\text{responses} & & Y_1 \arrow[d] & Y_2 \arrow[d] & Y_3 \arrow[d] & \cdots & Y_\infty \arrow[d] \\
\text{estimates} & X_0 \arrow[r] & X_1 \arrow[r] & X_2 \arrow[r] & X_3 \arrow[r] & \cdots \arrow[r] & X_\infty \sim \theta + \mathcal{E}
\end{tikzcd} \tag{8.2}$$

We want to create a tracking algorithm that provides estimates X_t that adapt to changes in θ_t, but has a known distribution if θ_t is invariant for some t, as represented in (8.2). Such a tracking algorithm is similar to KFs in that its states have a known distribution (Arulampalam et al. 2002), and similar to the ERS, specifically Elo's Current Rating Formula (Elo 1978), in that it continually adapts to changes in the underlying parameters without having to specify a growth model. An illustration of a simple tracker that conforms to this definition is given in Sect. 8.2.2, after which a proof of convergence is given in Sect. 8.2.3.

8.2.2 *Example of a Tracker*

We present a simple non-trivial case of a tracker that conforms to scheme (8.1), i.e., it dynamically adapts to change in the model parameters, and converges to a known error distribution if the scheme in (8.2) holds.

8.2.2.1 Coin Tossing Tracker

Consider the following coin tossing example.

$$\Pr(Y_i = 1|\theta) = \theta \tag{8.3}$$

where the probability of tossing head, i.e., $Y_i = 1$, is θ. If we consider a bandwidth of n sequential coin flips, then we simply define the sum score $X_+^{(n)}$ as follows:

$$X_+^{(n)} = \sum_{i=1}^{n} Y_i \sim \text{binom}(n, \theta) \tag{8.4}$$

Since $(Y_1, \ldots, Y_n)$ is independent of $\theta|X_+^{(n)}$, we can define an auxiliary variable Z using the sufficient statistic X_+:

$$\Pr(Z = 1|X_+^{(n)} = x_+) = \frac{x_+}{n} = \Pr(Y_i = 1|X_+^{(n)}, \theta). \tag{8.5}$$

Using the auxiliary variable Z, which is the expected response given the sum score $X_+^{(n)}$, and the observed response Y, we readily find the following sequential update rule for $X_+^{(n)}$, which is denoted with the subscript t as an index for time:

$$X_{t+1}^{(n)} = X_t^{(n)} + Y_t - Z_t \sim \text{binom}(n, \theta) \tag{8.6}$$

which gives us the simplest non-trivial tracker $X_t^{(n)}/n$ for θ_t meeting our definition in (8.2).

We provide some illustrations to demonstrate the workings of the tracker in (8.6), using simulations.[1]

8.2.2.2 Illustration of Convergence

First, we demonstrate the convergence of the sequential estimates of $X_t^{(n)}$ to the invariant distribution, binom(n, θ). As data, 1000 coin tosses are simulated with $\theta = .3$. Using the algorithm in (8.6) with $n = 30$, $X_t^{(n)}$ was sequentially estimated on the data and its distribution plotted in Fig. 8.1. As a reference, the theoretical density of the binomial distribution ($n = 30$, $\theta = .3$) was added. Clearly, this tracker nicely converged to the expected distribution as the two lines in Fig. 8.1 coincide. While the simulation used an invariant probability of the coin falling heads with $\theta = .3$, i.e., conforming to the scheme in (8.2), we can also simulate different changes to θ over time, conforming to the scheme in (8.1).

[1] All simulation in this chapter are performed in R (R Core Team 2015).

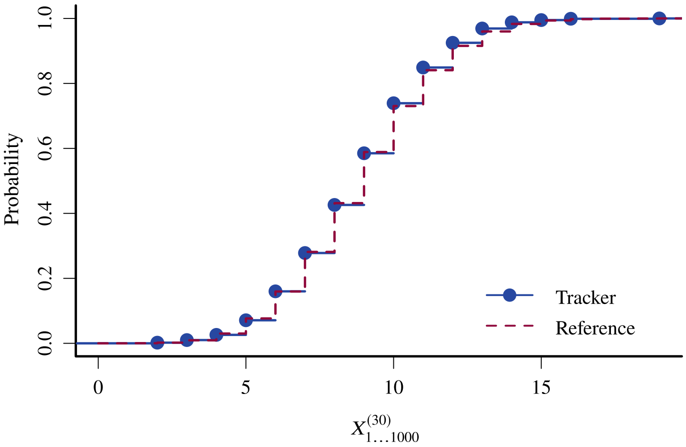


Fig. 8.1 Theoretical and empirical cumulative score distribution

8.2.2.3 Illustration of Tracking Smooth Growth

We simulate a scenario where θ smoothly changes over time t, i.e., we generate 1000 coin tosses with an increasing θ, and evaluate the development of the tracker in Fig. 8.2. Though θ is not stable at any time, it is clear that the tracker follows the development of θ quite closely with little lag. The step size n of the algorithm in (8.6) determines how fast the tracker can adapt to the changes in θ, where for this specific tracker a large n corresponds to a small step size and a small n to a large step size. Since θ is continually changing here, the tracker does not converge, but tracks the change rather well.

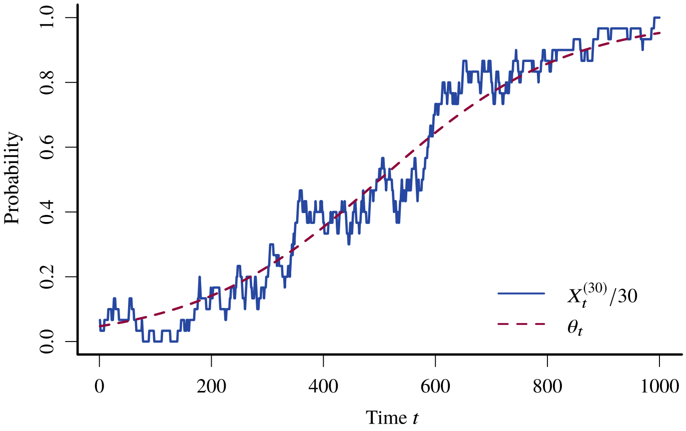


Fig. 8.2 Tracking smooth growth

8.2.2.4 Illustration of Tracking Sudden Changes

Next, we simulate a change point growth model where the probability of the coin falling heads changes from $\theta = .3$ to $\theta = .8$ at $t = 500$. The tracker is plotted in Fig. 8.3. Again, it can be seen that the tracker follows the development of θ closely. The tracker is always lagging, i.e., its development follows the development of θ with some delay depending on the step size n of the algorithm. This lag can be observed after the change point, and its size is related to the size of the change and step size of the algorithm.

8.2.2.5 Illustration of Varying Step Sizes

In Fig. 8.4 we illustrate the effect of varying the step size. A smooth development of θ is simulated, developing from about .1 to just over .8. Three different trackers are simulated using three different step sizes, $n = 10$, $n = 50$, and $n = 100$, where a small n is a large step size. It can be seen that the tracker with the most noise, having the largest step size and therefore the smallest n, adapts quickest to changes in ability θ, where the tracker with the smallest step size shows less noise, and therefore quite some lag. The step sizes are straightforward bias-variance trade-offs. Large step sizes allow for a quick adaption to possibly large changes in θ, at the cost of quite some variance if θ is stable. Small step sizes reduce this variance at the risk of introducing bias under a changing θ.

In these examples, this simplest non-trivial example of a tracker has demonstrated that it tracks the development of θ, cf. the scheme in (8.1), and converges to an invariant distribution if θ is stable, cf. the scheme in (8.2). We like to point out that though this example is simple, i.e., the properties of this tracker in the case of

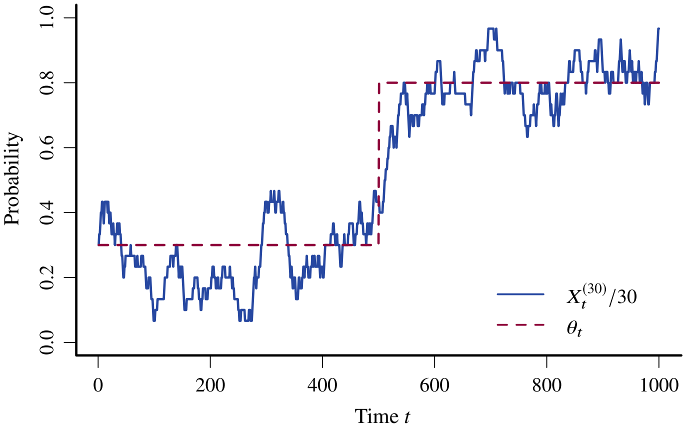


Fig. 8.3 Tracking a change point growth model

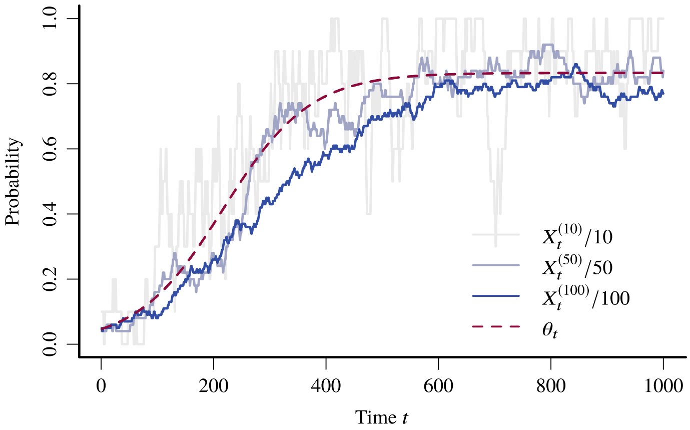


Fig. 8.4 Trackers with step sizes $n = 10$, $n = 50$ and $n = 100$

simple coin flips are the same as using a moving average, it is noteworthy that such trackers differ substantively from both maximum likelihood estimation (MLE) and Bayesian estimation techniques. The tracker estimates are continually adapting to changes in the model parameters, and convergence in distribution takes place while both the model parameter and the transition kernel are unchanging. This property of convergence under the scheme in (8.2) is generalized for all trackers in the following section.

8.2.3 Convergence in Kullback-Leibler Divergence

We provide some general proof of the convergence of Markov chains to an invariant distribution, given this distribution does not change between two time points, t and $t+1$. We use the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951; Eggen 1999) to quantify the divergence between the current distribution f_t and the invariant distribution f_∞.

Theorem 1 (Convergence in KL divergence) *If the invariant distribution* $f_\infty(x)$ *and the transition kernel* $f_\infty(x|y)$ *do not change between* t *and* $t+1$, *the KL divergence between the current distribution and the invariant distribution decreases between two time points:*

$$\int_{\mathcal{R}} \ln\left(\frac{f_\infty(x)}{f_{t+1}(x)}\right) f_\infty(x)dx \leq \int_{\mathcal{R}} \ln\left(\frac{f_\infty(y)}{f_t(y)}\right) f_\infty(y)dy. \tag{8.7}$$

if

$$f_\infty(x) = \int_{\mathcal{R}} f_\infty(x|y) f_\infty(y)dy \tag{8.8}$$

and

$$f_{t+1}(x) = \int_{\mathcal{R}} f_\infty(x|y) f_t(y)dy \tag{8.9}$$

Proof Using Bayes' rule, we can rewrite (8.9) as follows:

$$f_{t+1}(x) = \int_{\mathcal{R}} \frac{f_\infty(y|x)f_\infty(x)}{f_\infty(y)} f_t(y)dy \tag{8.10}$$

and place $f_\infty(x)$ outside the integral:

$$\frac{f_{t+1}(x)}{f_\infty(x)} = \int_{\mathcal{R}} \frac{f_t(y)}{f_\infty(y)} f_\infty(y|x)dy. \tag{8.11}$$

Taking the logarithm and integrating with respect to $f_\infty(x)$ gives:

$$\int_{\mathcal{R}} \ln\left(\frac{f_{t+1}(x)}{f_\infty(x)}\right) f_\infty(x)dx = \int_{\mathcal{R}} \ln\left(\int_{\mathcal{R}} \frac{f_t(y)}{f_\infty(y)} f_\infty(y|x)dy\right) f_\infty(x)dx. \tag{8.12}$$

Using Jensen's inequality, we obtain:

$$\int_{\mathcal{R}} \ln\left(\int_{\mathcal{R}} \frac{f_t(y)}{f_\infty(y)} f_\infty(y|x)dy\right) f_\infty(x)dx \geq \int_{\mathcal{R}}\int_{\mathcal{R}} \ln\left(\frac{f_t(y)}{f_\infty(y)}\right) f_\infty(y|x)f_\infty(x)dydx \tag{8.13}$$

which we can use to simplify (8.12) into:

$$\int_{\mathcal{R}} \ln\left(\frac{f_{t+1}(x)}{f_\infty(x)}\right) f_\infty(x)dx \geq \int_{\mathcal{R}} \ln\left(\frac{f_t(y)}{f_\infty(y)}\right) f_\infty(y)dy. \tag{8.14}$$

Writing (8.14) as a KL divergence, we interchange numerators and denominators and therefore change the sign of the inequality:

$$\int_{\mathcal{R}} \ln\left(\frac{f_\infty(x)}{f_{t+1}(x)}\right) f_\infty(x)dx \leq \int_{\mathcal{R}} \ln\left(\frac{f_\infty(y)}{f_t(y)}\right) f_\infty(y)dy \tag{8.15}$$

which concludes our proof. □

It was proven quite generally that trackers as described by (8.2) possesses an attractive quality. After every item response, the ability distribution of X_t monotonically converges in KL divergence to X_∞ (Kullback and Leibler 1951). The KL divergence is a divergence measure between two distributions, in our case, the theoretical distribution of ability estimates X_t and the invariant distribution of estimates X_∞. If the KL divergence is small, the ability estimates are (almost) converged to the proper invariant distribution. Monotone convergence assures this divergence decreases with every new response under the conditions of (8.2).

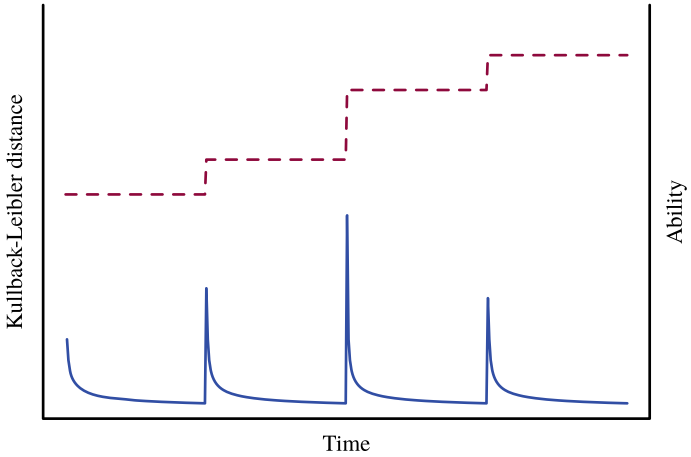


Fig. 8.5 Development of ability (dashed line) and Kullback-Leibler (KL) divergence (solid line) over time

8.2.3.1 Illustration of Development of Kullback-Leibler (KL) Divergence

In Fig. 8.5 we provide an illustration[2] how the KL divergence could develop over time. If ability θ is stable for some time, then changing, and then stable again, we can see how the KL divergence could decrease in times of stability and increase when ability changes.

We believe this illustration shows that the convergence property is suitable for use in the practice of educational measurement, where students mostly respond to sets of items, even if they are assessed frequently (Brinkhuis et al. 2018). The assumption here is that ability is stable during the relatively short time in which a student answers a set of items, and might change between the administrations of sets. Clearly, no convergence takes place if ability is continually changing (Sosnovsky et al. 2018).

8.2.4 Simulating Surveys

Section 8.2.2 provides some simulations to illustrate the invariant distribution of the simple tracker and to demonstrate how the estimates track individual development under several simulation conditions. One goal is to simulate and track the development of groups, as might be done in survey research.

We consider a simplified scenario where an entire population consisting of 100,000 persons would answer just 5 questions in a survey that is administered 4 times, for example to track educational progress of the entire population within

[2] This serves as illustration only, no actual KL divergences were calculated.

a year. Using the algorithm in (8.6), that compares to 5 flips of 100,000 uniquely biased coins for each survey. The probabilities of the coins falling heads changes for every survey, and are sampled from a beta distribution. The parameters of these beta distribution where $a = 5, 6\frac{2}{3}, 8\frac{1}{3}, 10$ and $b = 10$ for the 4 simulated surveys. These 4 beta distributions are plotted in Fig. 8.6 from left to right, using dashed lines. A very large step size ($n = 2$) was used for the algorithm, to allow the estimates $\boldsymbol{X}$ to adapt quickly to the changes in θ. Since θ is sampled from a beta distribution, the estimates $\boldsymbol{X}$ are beta-binomial distributed. Using MLE, the two parameters of the beta distribution of θ were estimated for each of the 4 administrations, and these estimated beta distributions are plotted in Fig. 8.6. The graph demonstrates that it is possible to accurately track the development of an entire population by administrating a limited amount of items to all individuals.

Note that this scenario is quite different from a more traditional sampling approach where many items are administered to complex samples of individuals. If the total number of responses would be kept equal, for tracking the development of the entire population it is beneficial to administer a few questions to many individuals. On the contrary, for tracking individual development, it is beneficial to administer many items to few individuals. For example, while the information in the 5 items per survey described above is too limited for tracking individual growth, especially in considering progress that is made between surveys, it is sufficient for tracking the population parameters. Though trackers can both be used for tracking individual progress or progress of the population, the preferred design of data collection depends on the desired level of inference.

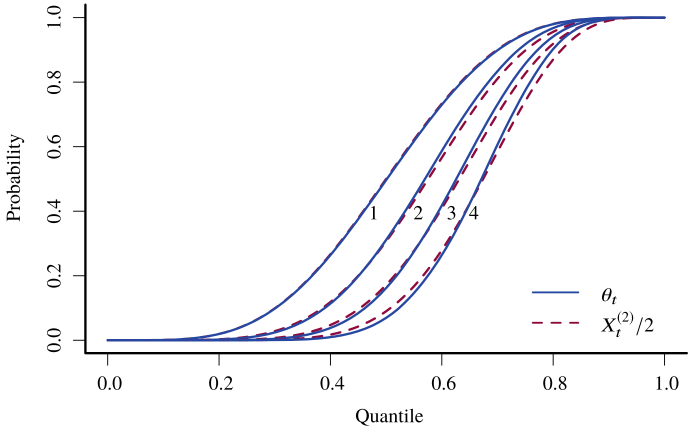


Fig. 8.6 Trackers with step sizes $n = 10$, $n = 50$ and $n = 100$

8.3 Discussion

In this chapter, trackers and their possible uses in educational measurement have been described. Trackers are defined as dynamic parameter estimates with specific properties. These trackers combine some of the properties of the ERS and state space models as KFs, which both have strengths and weaknesses. The ERS is a feasible method for dealing with data with changing model parameter, i.e., ratings. It is simple, provides real-time results, and requires no assumptions on the types of growth. However, it lacks a proper distribution of estimates, and therefore no statistics can be used on the estimates, e.g., to test for change, or to track any aggregate of estimates. KFs, on the other hand, do assume specific distributions of estimates, but need specified growth models, which are not readily available in many educational measurement applications.

Trackers should be able to adapt to changes in both model parameters and the transition kernel, cf. the scheme in (8.1). In addition, we require that the estimates converge in distribution if the model is invariant, cf. scheme (8.2). A simple example of a tracker conforming to this definition has been introduced in (8.6), with a transition kernel that creates a Markov chain with a binomial invariant distribution. A well-known technique for obtaining a transition kernel that creates a Markov chain with a specified distribution is called the Metropolis algorithm (Metropolis et al. 1953; Hastings 1970; Chib and Greenberg 1995). The Metropolis algorithm can be used to create a transition kernel that satisfies (8.2). The general proof that such Markov chains monotonically converge to their invariant distribution has been provided in Theorem 1.

While the simple binomial example might not have much practical use directly, other trackers can be developed to provide estimates with a known error distribution, for example an ability estimate X which is distributed $\mathcal{N}(\theta, \sigma^2)$. Two simple examples of such trackers are presented in Brinkhuis and Maris (2010). Such estimates could directly be used in other statistical analyses since the magnitude of the error does not depend on the ability level itself. These assumptions compare directly to the assumptions of classical test theory, where an observed score equals the sum of the true score and an uncorrelated error. Another simple application of using these estimates directly is to look at empirical cumulative distributions of ability estimates.

Trackers as defined in this chapter retain the strengths of both state space models and ratings systems, and resolve some of their weaknesses. Their properties are suitable for, among others, applications in educational measurement in either tracking individual progress or any aggregate thereof, e.g., classes or schools, or the performance of the entire population as in survey research. The algorithms remain relatively simple and light-weight, and therefore allow to provide real-time results even in large applications. They are unique in that they continually adapt to a new transition kernel and converge in distribution if there is an invariant distribution, which is quite different from both MLE and Bayesian estimation techniques.

Acknowledgements An earlier version of this manuscript is part of Brinkhuis (2014).

References

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, *50*(2), 174–188. https://doi.org/10.1109/78.978374.

Batchelder, W. H., & Bershad, N. J. (1979). The statistical analysis of a Thurstonian model for rating chess players. *Journal of Mathematical Psychology*, *19*(1), 39–60. https://doi.org/10.1016/0022-2496(79)90004-X.

Batchelder, W. H., Bershad, N. J., & Simpson, R. S. (1992). Dynamic paired-comparison scaling. *Journal of Mathematical Psychology*, *36*, 185–212. https://doi.org/10.1016/0022-2496(92)90036-7.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25.

Black, P., & Wiliam, D. (2003). "In praise of educational research": Formative assessment. *British Educational Research Journal*, *29*(5), 623–637. https://doi.org/10.1080/0141192032000133721.

Brinkhuis, M. J. S. (2014). *Tracking educational progress* (Ph.D. Thesis). University of Amsterdam. http://hdl.handle.net/11245/1.433219.

Brinkhuis, M. J. S., Bakker, M., & Maris, G. (2015). Filtering data for detecting differential development. *Journal of Educational Measurement*, *52*(3), 319–338. https://doi.org/10.1111/jedm.12078.

Brinkhuis, M. J. S., & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems* (Measurement and Research Department Reports 09-01). Arnhem: Cito. https://www.researchgate.net/publication/242357963.

Brinkhuis, M. J. S., & Maris, G. (2010). *Adaptive estimation: How to hit a moving target* (Measurement and Research Department Reports 10-01). Arnhem: Cito. https://www.cito.nl/kennis-en-innovatie/kennisbank/p207-adaptive-estimation-how-to-hit-a-moving-target.

Brinkhuis, M. J. S., Savi, A. O., Coomans, F., Hofman, A. D., van der Maas, H. L. J., & Maris, G. (2018). Learning as it happens: A decade of analyzing and shaping a large-scale online learning system. *Journal of Learning Analytics, 5*(2), 29–46. https://doi.org/10.18608/jla.2018.52.3.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, *86*(2), 221–241. https://doi.org/10.1016/S0304-4076(97)00115-2.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*(4), 327–335. https://doi.org/10.2307/2684568.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*(3), 249–261. https://doi.org/10.1177/01466219922031365.

Elo, A. E. (1978). *The rating of chess players, past and present*. London: B. T. Batsford Ltd.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, *57*(1), 1–17. https://doi.org/10.1093/biomet/57.1.1.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. New Jersey: Lawrence Erlbaum Associates Inc.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering, 82*(Series D), 35–45

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86. http://www.jstor.org/stable/2236703.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*(1), 110–133. https://doi.org/10.2307/1130295.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122. https://doi.org/10.1007/BF02294746.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114.

Pelánek, R., Papoušek, J., Řihák, J., Stanislav, V., & Nižnan, J. (2017). Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction*, *27*(1), 89–118. https://doi.org/10.1007/s11257-016-9185-7.

R Core Team. (2015). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing, Vienna, Austria*. http://www.R-project.org/.

Sosnovsky, S., Müter, L., Valkenier, M., Brinkhuis, M., & Hofman, A. (2018). Detection of student modelling anomalies. In V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink, & M. Scheffel (Eds.), *Lifelong technology-enhanced learning* (pp. 531–536). Berlin: Springer. https://doi.org/10.1007/978-3-319-98572-5_41.

van Rijn, P. W. (2008). *Categorical time series in psychological measurement* (Ph.D. Thesis). University of Amsterdam, Amsterdam, Netherlands. http://dare.uva.nl/record/270555.

Veldkamp, B. P., Matteucci, M., & Eggen, T. J. H. M. (2011). Computerized adaptive testing in computer assisted learning? In S. De Wannemacker, G. Clarebout, & P. De Causmaecker (Eds.), *Interdisciplinary approaches to adaptive learning, communications in computer and information science* (Vol. 126, pp. 28–39). Berlin: Springer. https://doi.org/10.1007/978-3-642-20074-8_3.

Visser, I., Raijmakers, M. E. J., & van der Maas, H. L. J. (2009). Hidden Markov models for individual time series. In J. Valsiner, P. C. M. Molenaar, M. C. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (Chap. 13, pp. 269–289). New York : Springer. https://doi.org/10.1007/978-0-387-95922-1_13.

Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*(6), 549–562. https://doi.org/10.1111/j.1365-2729.2010.00368.x.

Welch, G., & Bishop, G. (1995). *An introduction to the Kalman filter* (Technical Report TR 95-041). Chapel Hill, NC, USA: Department of Computer Science, University of North Carolina at Chapel Hill. http://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf. Updated July 24, 2006.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 9
Finding Equivalent Standards in Small Samples

Monika Vaheoja

Abstract The process of resetting performance standards, with small samples, in different forms of an exam is statistically challenging as the estimates are often biased. Empirical information is therefore, often neglected and content experts reset the standard. In the current article, performance standards are set to a new form in by different methods: circle-arc equating and concurrent calibration with OPLM as an IRT model. The responses on the exam forms that are to be equated are simulated in different situations varying in sample size, test length, test difficulty and respondent's abilities. The results demonstrate that even in small samples (50 subjects taking both tests), the IRT-method with OPLM as a model outperforms circle-arc equating when test difficulty and population ability interact.

9.1 Introduction

In computer-administrated tests using item banks, students with different abilities answer different item sets varying in difficulty, discrimination and number of items. When test takers receive a diploma or certificate, such a set of items can be referred to as an exam, through which students demonstrate their mastery of a topic by achieving a certain minimal level. This minimal performance level is reflected in a cut-score which reflects a performance standard for the exam and must be fair: cut-scores set on different exam forms should lead to identical decisions for examinees with the same ability. In other words, the probability to pass a test must be related to the ability of the test taker, not the exam the student has responded to.

Numerous methods are available for setting a performance standard on an exam (for an overview, see Hambleton and Pitoniak 2006) and numerous statistical models are available for test equating (Kolen and Brennan 2004; von Davier 2011). Test equating is a process in which the main goal is to establish, with as near accuracy

M. Vaheoja (✉)
University of Twente OMD-BMS, Enschede, The Netherlands
e-mail: m.vaheoja@utwente.nl

10voordeleraar, The Hague, The Netherlands

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_9

as possible, a valid equivalence between raw scores on multiple forms of a test. For example, to compare examinees' scores on multiple versions or forms (Holland and Rubin 1982). This ensures that test scores from multiple forms can be used interchangeably. However, as the test equating process is a statistical approach, it provides more precise results in larger samples (Kolen and Brennan 2004, p. 307–309). The models and methods that are advised for small samples have been shown to include large equating error and bias (Kim and Livingston 2010). One method that outperformed the advised ones, in the context of small sample testing, was circle-arc equating (Dwyer 2016; Kim and Livingston 2011; LaFlair et al. 2015; Livingston and Kim 2009), an equating method based on classical test theory. This means that circle-arc equating may be preferred for the transfer of performance standards to new forms in cases where there are limited number of examinees.

Dwyer (2016) studied this problem by comparing the estimation precision of the cut-score when content experts set the standard using the Angoff standard-setting approach, to when the cut-score was reset with circle-arc equating. The content experts' estimates were also rescaled by correcting them based on estimates of anchor items. The results showed that circle-arc equating did indeed outperform resetting and rescaling the cut-score in maintaining an equivalent performance standard across exam forms. However, this study had observed score equating as its subject, which introduces equating errors when examinees' ability distributions differ (Lord and Wingersky 1988). When the examinees' abilities differ across forms, item response theory (IRT) is advised.

Item response theory models the probability of a respondent correctly answering an item (Hambleton et al. 1991). A correct score on an item is dependent on the ability of the respondent and on the characteristics of the item. Respondents' ability cannot be observed directly, so it is included in the model as a latent variable (θ). The characteristics of the item can be its difficulty parameter (β), discrimination parameter (α) and guessing parameter (γ). The primary goal of item response models is to estimate this latent variable for each person as precisely as possible. Limited research is available on the minimum sample size required for IRT equating to maintain equivalent performance standards across test forms. Using IRT equating is even discouraged with small samples (Kolen and Brennan 2004, p. 303).

However, the use of IRT may be recommended for transferring a performance standard from one form to the next in cases where exam form consist of highly discriminative items. Because the α and β of items influences the test information function and thereby the test standard error (Yen and Fitzpatrick 2006). And if the exam has been constructed to have its average difficulty nearby the targeted expected cut-point, its standard error will be the smallest around that area.

In IRT, the standard error of measurement differs across ability and gives the smallest error in expected score values where the test gives the most information. This means that where the most information is given, the expected scores for the difficulty of the exam are estimated more accurately (Lord and Novick 2008). With the classical test theory approach, which includes circle-arc equating, the standard error of measurement is equal for all scores. If the cut-score falls into the extremes of the scale score, it may be better to use the circle-arc approach due to the constant standard error across the score scale. However, if the cut-score is in the middle range of the score scale, it may be preferable to use the IRT approach.

Therefore, within this article, both of the above methods are compared in re-estimating the performance standard on a new form in cases where exam forms vary in length, number of examinees, difficulty and ability distribution.

9.2 Method

To maintain an equal performance standard across exam forms and to study its accuracy, we used simulated data. Because our interest is to find the best solution for the practice for the national exam in teacher-training program in the Netherlands, we make use of the characteristics of the student ability distributions, characteristics of the item bank and the process to find a cut-score on a new form that equals with the ability of the cut-score on the reference exam. This means that we transferred the cut-score from the reference exam to a new form with IRT concurrent calibration and circle-arc equating.

The examinees' item responses were simulated for two forms to study the effect of sample size (50, 100, 200, 400 and 800), ability (both populations have equal ability, second population has lower ability, and second population has a higher ability), test length (25, 50 and 100 items), and difficulty (both forms are same difficulty, second form is easier and second form is harder) creating a total of $(5 \times 3 \times 3 \times 3 =)$ 135 facets. The data structure is from computer administrated exams for a maths teacher-training program in a public secondary school in the Netherlands.

Sample size and ability. For sample sizes of 50, 100, 200, 400 and 800 subjects, item responses were simulated for both forms. The ability distribution of the reference population was equal to the mean of the examinees in the maths teacher-trainees population. By subtracting and then adding 0.4 standard deviation to the average population ability, we created populations with both lower and higher abilities (see Table 9.1). This to make sure we created a similar context as it is in the practice.

The exams. Twenty five items were randomly sampled from an item pool that was stratified based on discrimination parameters of the calibrated items. These 25 items defined the reference exam, for which we set a cut-score of 13.456, comparable with the average percentage of the cut-scores that experts had set for all maths exams. The theoretical cut-score on the second form was computed by estimating the expected score that is equal to the ability level on the reference exam.

Table 9.1 Mean and standard deviation of different ability distributions

	Population ability		
	Lower	Reference	Higher
Mean ability	−0.071	0.139	0.349
Standard deviation of ability	0.525	0.525	0.525

Table 9.2 Exam difficulties and corresponding cut-points

	Exam		
	Easier	Reference	Difficult
Mean difficulty	−0.071	0.139	0.349
Mean of anchor items	0.274	0.274	0.274
Mean of other items	−0.157	0.105	0.368
Standard deviation of difficulty	0.437	0.406	0.402
Cut-point	16.530	13.456	9.780

Test length and difficulty. To lengthen the test, these 25 items from the reference exam were doubled for the exam with 50 items and quadrupled to create a form with 100 items, which kept all other test characteristics constant. Five items from the 25 were marked as anchor items, anchoring both forms by 20%. The difficulty of the new form was adjusted by subtracting or adding a constant to the difficulty parameters of non-anchor items. The reasoning behind this being that the difficulty of the total exam should be equal to the mean ability of the different populations. This resulted in both easier and difficult forms (Table 9.2).

General Method. To maintain an equal performance standard across exam forms with the simulated data, we transferred the cut-score from the reference exam to a new form with IRT concurrent calibration and circle-arc equating. This procedure of simulating response data and transferring the cut-score to a new form in each facet was repeated until we had 1000 successful runs. When the IRT model could not be estimated in a run, because the entire examinee cohort had simulated correct or wrong scores, the circle-arc estimators were excluded, and new data was simulated.

Circle-Arc equation. In circle-arc equating, the observed score scales from both forms are equated by assuming a curve-linear relationship between both forms from an arc (see an example in Fig. 9.1; Livingston and Kim 2009). This arc represents an equating function through which the mutually corresponding observed scores on both forms can be related. An arc is mathematically derived and has three points: minimum and maximum possible scores and an empirically estimated midpoint.

In Fig. 9.1, this equating function is illustrated for the new and reference forms. The solid line is the identity line between both forms. In Fig. 9.1, three points are drawn: the lower dot represents the arbitrary chosen minimum possible score, which could be the chance score on both forms. The upper point represents the maximum possible score on both forms. The triangle represents the empirically estimated midpoint, which in the case of non-equivalent anchor test design (NEAT design) is derived from chained linear equating (see Livingston and Kim 2009). If there is no difference between the two forms, then the circle-arc function will be equal to the identity line. If the curve falls to the left, then the reference form is easier. Finally, if the curve falls to the right of the identity line, then the new form is easier.

Concurrent calibration and OPLM. In IRT concurrent calibration, exams that have a common set of items are calibrated jointly (Kolen and Brennan 2004). Then, the

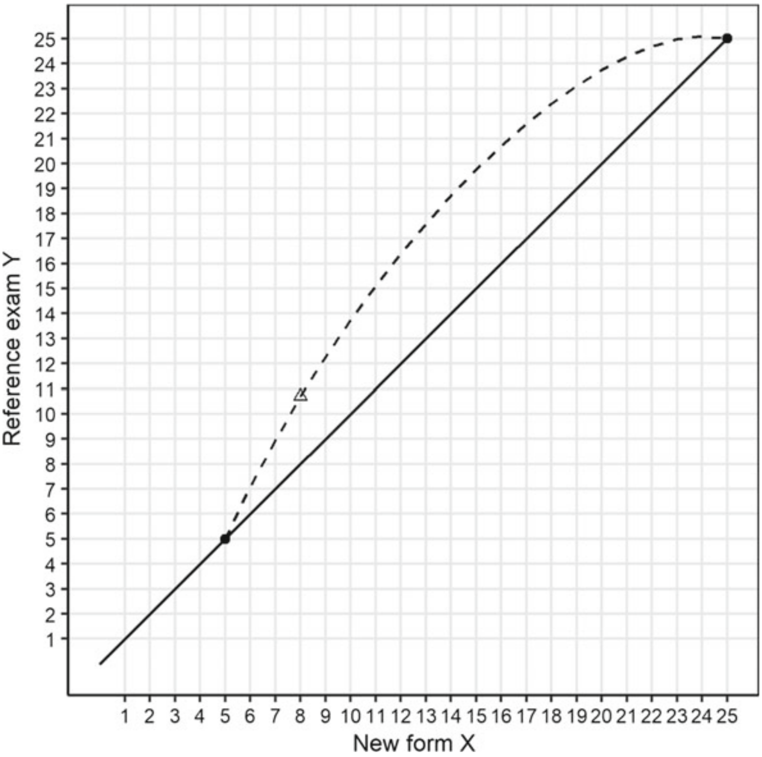


Fig. 9.1 The circle-arc equating function

items are combined into a test function that relates the expected scores on one exam to the ability scale. The test function is the sum of the items from that particular exam. The test function of the second exam is computed in the same manner, which makes it possible to find the expected scores for both exams using a specific ability that is related to the performance standard.

In Fig. 9.2, this process is visualized. The dashed line represents the test function for the reference exam, whilst the solid line represents the new form. In order to find an equivalent cut-score on the second exam, firstly, an ability score that corresponds to the cut-score on the reference exam must be found. In the example in Fig. 9.2, the corresponding ability value equals 0.03. As both forms are calibrated jointly, the same ability score can be used to find an equivalent expected score on the second form. In the lower part of Fig. 9.2, the corresponding expected score is equal to 9.98. From this we can conclude that the second form is more difficult than the reference form as the equivalent cut-score is lower.

The IRT model in the current article is the one parameter logistic model (OPLM), in which elegant features of the Rasch and the Two Parameter Logistic (2PL) models are combined (Verhelst and Glas 1995; Verhelst et al. 1993). The discrimination

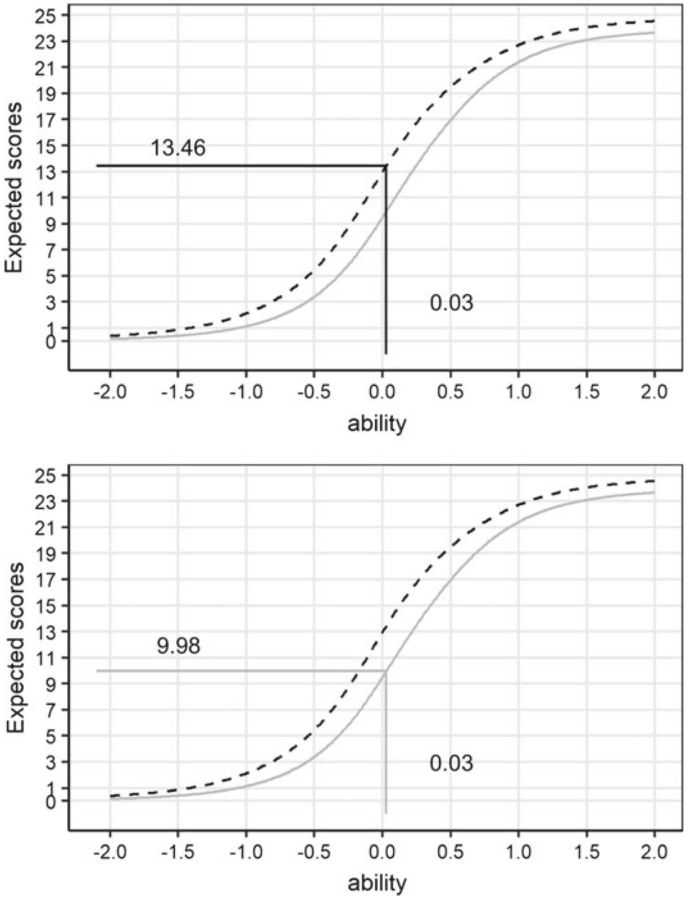


Fig. 9.2 Illustration of IRT concurrent calibration equating to find an equivalent score

parameters (α_i) are imputed into the OPLM as a constant (a_i). These constants are supplied through a hypothesis. This means that in the OPLM, the items discriminate differently, and the difficulty parameters can be estimated by the conditional maximum likelihood.

9.3 Results

Figures 9.3 and 9.4, present the bias and Root Mean Squared Error (RMSE) for each of the facets, per equating method.

Bias and variability in estimators. Bias is the most important measure in our case as when the equating method re-estimates the cut-score and it does not correspond to the same ability, this could have major consequences for the examinees. The corrected bias means that the bias was divided by test length. As there is higher bias for longer tests, more examinees are affected in shorter tests than in longer tests. For example, if forms of 25 and 50 items are administrated to 100 students each, there are relatively more students falling in one score range in shorter exam than for a longer exam. Which means that one score difference in shorter exam affects relatively more students than in a longer exam.

Figure 9.3 shows the average corrected bias for each facet in both methods. A negative bias means that the exam was estimated to be more difficult and a positive bias means that the exam was estimated to be easier. The results demonstrate that the IRT estimators are less biased in each facet than the estimators from the circle-arc method. Therefore, it seems as though the circle-arc method tends to estimate a lower cut-score for an easier exam and a much higher one for a more difficult exam, even in contexts where the participants have equal ability. Nevertheless, the IRT method shows some bias too. The bias in IRT estimators decreases with an increased sample size within, test length and ability change. With circle-arc equating, no difference in bias for sample sizes was found, but bias differed across population ability. Circle-arc estimators have less bias for easier exams in cases where the ability of the second population is higher. Additionally, the estimators show less bias for more difficult exams in cases where the population ability is lower.

In general, the estimators of cut-scores in the new forms vary less with IRT (Fig. 9.4). Only in cases where exams have the same difficulty, circle-arc estimators

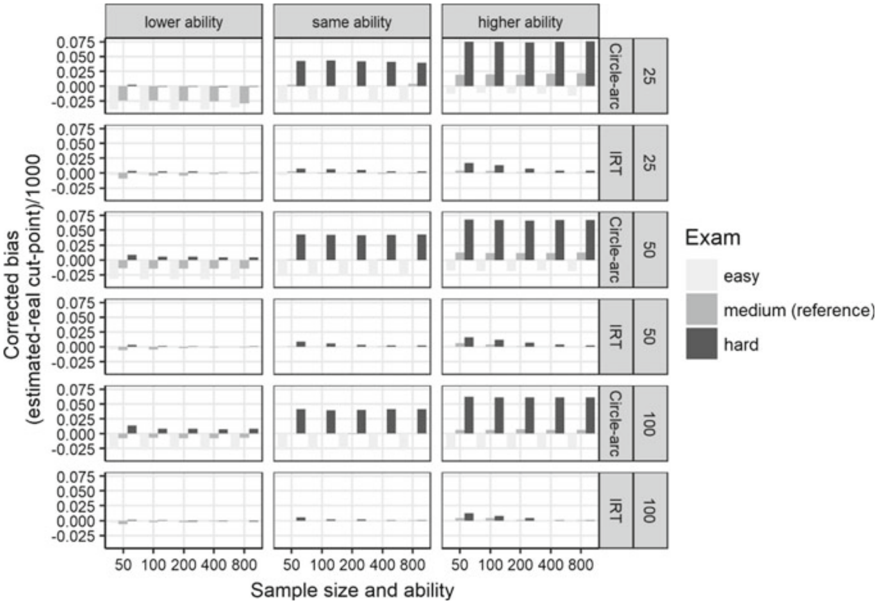


Fig. 9.3 Corrected bias in the estimators from circle-arc equating and IRT

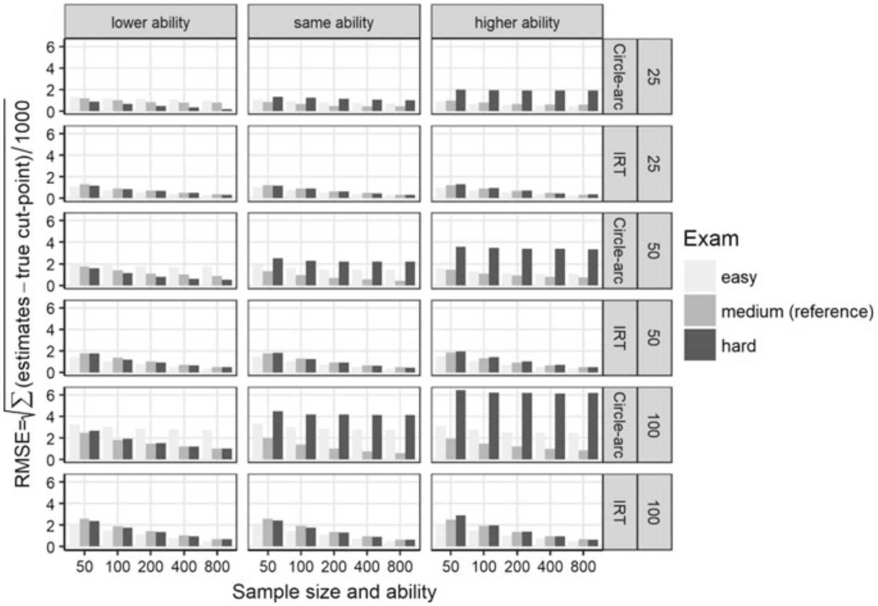


Fig. 9.4 RMSE of the estimators from circle-arc equating and IRT equating

vary less than IRT estimators. The IRT estimators show greater variability for smaller samples, whereas with circle-arc equating, no variability can be seen for different sample sizes. The variability increases with the test length for both methods. Furthermore, the estimators from both methods vary more in estimating the cut-score for difficult exams in contexts where the ability of the examinees is the same or higher. However, in contexts where the examinees from the second population have a lower ability, circle-arc equating varies more in estimating the cut-score for easier exams.

9.4 Conclusion and Discussion

The results generally indicate that IRT estimators were less biased and showed less variability in estimating the cut-score in the new form than the estimators of circle-arc equating. Only in contexts where the examinees had the same ability and both forms had the same difficulty showed the circle-arc estimators less bias and variability. However, in such cases, no equating is needed as both forms are parallel. For both methods, we saw an increase in bias for longer tests. Only for IRT estimators we observed a decrease in corrected bias and variability for large samples. Even for small samples of 50 examinees per form, the IRT estimators were less biased than the circle-arc estimators were.

There are two possible reasons for the success of our study in favour of IRT estimators. Firstly, the stratified random items from the item bank were highly discriminating and difficulty of the items was diverse. Secondly, we used OPLM as our IRT model to calibrate both forms and our focus was only on estimating the cut-score. Cut-scores often fall in the middle range of the score scale where the expected test scores have the smallest standard error.

The use of OPLM to calibrate the exams made it possible to use conditional maximum likelihood to estimate the parameters. This is an advantage when the parameters are estimated with a limited number of examinees per form and the examinees were not a random sample from the population. Their average ability, therefore, does not represent the population and it would be ambitious to assume the ability distribution to be normal.

However, some bias was found in the IRT estimators, particularly within small samples. The bias in the estimators was higher in contexts where the examinees from the second form were in the higher ability group and took the more difficult exam. The estimates were higher, meaning that the difficult exam was estimated as easier than it was supposed to be. This bias might have been caused by the anchor set in our study, because the anchor set was more difficult than the mean difficulty of the reference exam. Which is not advised for equating, in fact, the anchor set should be a miniature version of the exam including its difficulty (Holland and Dorans 2006). This, however, might indicate an important exam construction rule, which seems to be more crucial for small samples than for exams with more examinees.

Investigating the impact of anchor set when transferring the performance standard in small samples, where there is no bias in estimators and the anchor set is a miniature of the exam, could be extremely relevant. Another suggestion would be to use a fixed parameter calibration in which the anchor set parameters, or the parameters of the reference exam items are fixed. Kolen and Brennan (2004; p. 183) briefly addressed this aspect and implied that fixing the parameters in contexts where the ability of the populations differs, might lead to biased estimates. This is because the ability distribution is estimated as a mean of zero and a standard deviation of one. However, this is only the case when the marginal maximum likelihood is used to estimate the parameters, this may not be the case when the conditional maximum likelihood is used, as it is in OPLM.

The bias present in the circle-arc method should not be neglected. Even though, Dwyer (2016) showed promising results in favour of the circle-arc method, this method has some weaknesses. Livingston and Kim (2009) present this equating method as an improvement in the chained linear equating method, as a method to overcome impossible maximum scores because it follows an arc. However, in the case of NEAT design, is the empirically estimated midpoint for the circle-arch method derived from the chained linear equating method. If the new form is difficult and examinees abilities are low, this arc still results in impossible maximum scores. The circle-arc method could be improved by adding a second empirical point which partially restricts the arc. A second weakness of the circle-arc equating method is the definition of the minimum score. Authors tend to leave the decision for the minimum score up to the user. However, the choice of the minimum score affects the length of the circle's radius that is then used to compute the arc and the equating function.

Finally, Hambleton and Jones (1993) observed a limitation of the classical test theory approach; the test scores obtained by classical test theory applications are dependent on the test and biased when examinees' abilities differ. Within this article, this limitation was empirically demonstrated. Additionally, although Kolen and Brennan (2004) did not advise using OPLM as an IRT equating model for small samples, we would urge researchers to consider using OPLM in resetting performance standards due to our promising results.

References

Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement, 53,* 3–22.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement*. American Council on Education and Praeger Publishers.

Hambleton, R. K., Swamminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement*. American Council on Education and Praeger Publishers.

Holland, P. W., & Rubin, B. P. (1982). *Test equating*. New York: Academic Press.

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement, 47*(3), 286–298.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

LaFlair, G. T., Isbell, D., May, L. D. N., Arvizu, M. N. G., & Jamieson, J. (2015). Equating in small-scale language testing programs. *Language Testing, 12*(23), 127–144.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46,* 330–343.

Lord, F. M., & Novic, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Lord, F. M., & Wingersky, M. S. (1988). Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement, 8*(4), 453–461.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models*. New York, NY: Springer.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1993). *OPLM: One parameter logistic model [computer software manual]*. Arnhem.

von Davier, A. A. (2011). *Statistical models for test equating*. New York, NY: Springer.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement*. American Council on Education and Praeger Publishers.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part III
Large Scale Assessments

Chapter 10
Clustering Behavioral Patterns Using Process Data in PIAAC Problem-Solving Items

Qiwei He, Dandan Liao and Hong Jiao

Abstract Technical advances provide the possibility of capturing timing and process data as test takers solve digital problems in computer-based assessments. The data collected in log files, which represent information beyond response data (i.e., correct/incorrect), are particularly valuable when examining interactive problem-solving tasks to identify the step-by-step problem-solving processes used by individual respondents. In this chapter, we present an exploratory study that used cluster analysis to investigate the relationship between behavioral patterns and proficiency estimates as well as employment-based background variables. Specifically, with a focus on the sample from the United States, we drew on a set of background variables related to employment status and process data collected from one problem-solving item in the Programme for the International Assessment of Adult Competencies (PIAAC) to address two research questions: (1) What do respondents in each cluster have in common regarding their behavioral patterns and backgrounds? (2) Is problem-solving proficiency related with respondents' behavioral patterns? Significant differences in problem-solving proficiency were found among clusters based on process data, especially when focusing on the group not solving the problem correctly. The results implied that different problem-solving strategies and behavioral patterns were related to proficiency estimates. What respondents did when not solving digital tasks correct was more influential to their problem-solving proficiency than what they did when getting them correct. These results helped us understand the relationship between sequences of actions and proficiency estimates in large-scale assessments and held the promise of further improving the accuracy of problem-solving proficiency estimates.

Q. He (✉)
Educational Testing Service, Princeton, USA
e-mail: qhe@ets.org

D. Liao
American Institutes for Research, Washington DC, USA

H. Jiao
University of Maryland, College Park, MD, USA

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_10

10.1 Introduction

The use of computers as an assessment delivery platform enables the development of new and innovative item types, such as interactive scenario-based items, and the collection of a broader range of information, including timing data and information about the processes that test takers engage in when completing assessment tasks (He and von Davier 2016). The data collected in log files, which are unique to computer-based assessments, provide information beyond response data (i.e., correct/incorrect) that is usually referred to as *process data*. Such information is particularly valuable when examining interactive problem-solving tasks to identify the step-by-step problem-solving processes used by individual respondents.

10.1.1 Problem-Solving Items in PIAAC

As the largest and most innovative international assessment of adults, the Programme for the International Assessment of Adult Competencies (PIAAC), starting from the first cycle in 2012, has sought to assess computer, digital-learning, and problem-solving skills, which are essential in the 21st century (Organisation for Economic Co-operation and Development [OECD] 2009, 2011, 2012; Schleicher 2008). Of significance here, PIAAC is the first international household survey of skills predominantly collected using information and communication technologies (ICT) in a core assessment domain: Problem Solving in Technology-Rich Environments (PSTRE). This international survey has been conducted in over 40 countries and measures the key cognitive and workplace skills needed for individuals to participate in society and for economies to prosper (OECD 2016). Evidence has shown that process data captured in PSTRE items provide a deeper insight into the cognitive processes used by respondents when they are solving digital tasks (e.g., Goldhammer et al. 2014; Liao et al. 2019; Chen et al. 2019). This additional information helps us understand the strategies that underlie proficient performance and holds the promise of better identifying behavioral patterns by subgroups, thus helping us seek solutions for teaching essential problem-solving skills to adults with particular needs (He and von Davier 2015, 2016).

The PSTRE items that we focused on in this study were used to assess the skills required to solve problems for personal, work, and civic purposes by setting up appropriate goals and plans, as well as how individuals access and make use of information through computers and networks (OECD 2009). This new domain involved more interactive item types and was available only on computers. The construct underlying the PSTRE items describes skillful use of ICT as collecting and evaluating information for communicating and performing practical tasks such as organizing a social activity, deciding between alternative offers, or judging the risks of medical treatments (OECD 2009). To give a response in the simulated computer environments that form the PSTRE tasks, participants were required to click buttons or links, select from dropdown menus, drag and drop, copy and paste, and so on (He and von Davier 2016).

10.1.2 Employability and PSTRE Skills

Employability and the completion of computer-based testing have been shown to be positively correlated in recent research (e.g., OECD 2013b; Vanek 2017; Liao et al. 2019). Although the United States was one of the PIAAC countries with the highest accessibility to computers, internet, and advanced electronic equipment, its performance, especially in the PSTRE domain that focuses on assessing ICT skills, was far lower than expectations. According to a recent report published by the National Center for Education Statistics (Rampey et al. 2016), U.S. test takers scored lower on average than their international peers, ranking in the lowest tier in the PSTRE domain as a country, and having the largest proportion of respondents below Level 1, which is the minimum proficiency level required to complete simple problem-solving tasks in daily life (OECD 2013a). These results raise attention to adults' current PSTRE skills in the U.S. population and their employability, which is highly associated with PSTRE skills.

This chapter presents an exploratory study that used cluster analysis to investigate the relationship between behavioral patterns and proficiency estimates as well as employment-based background variables. Specifically, with a focus on the sample from the United States, we drew on a set of background variables related to employment status and process data collected from one PSTRE item in PIAAC to address two research questions: (1) What do respondents in each cluster have in common regarding their behavioral patterns and backgrounds? (2) Is problem-solving proficiency consistent across clusters, or in other words, is problem-solving proficiency related to respondents' behavioral patterns?

10.2 Method

10.2.1 Sample

The PIAAC sample was representative of the population of adults with an age range of 16–65 years old who had prior experience with computers. Those who had never used computers were excluded from the problem-solving section; the task for this scale was by default (and by definition of the construct) assessed only on a computer-based testing platform (OECD 2010). A total of 1,340 test takers in the U.S. sample who completed the PSTRE items in the second module (PS2)[1] in PIAAC were included in

[1] The PIAAC was designed in a structure of multistage adaptive testing, by routing respondents to different modules in two stages. The PSTRE domain consists of two modules (PS1 and PS2),

the present study. Of them, there were 629 female test takers (46.9%) and 711 male test takers (53.1%). The mean age was 39.2 years (*SD* = 14.0). A majority (680) of members of this sample had an educational level above high school (50.7%), whereas 534 reported completing high school (39.9%), 124 reported less than high school (9.3%), and two cases were recorded as missing (0.1%). Please note that there were 14 cases that could not be matched between the process data and the background variables[2] and thus had to be removed in further data analysis, which resulted in 1,326 test takers in the final sample.

10.2.2 Instrumentation

A total of 14 PSTRE items were administered in the PIAAC main study. We focused on the process data resulting from the task requirements of one item. The "Meeting Room Assignment" item (U02) consisted of three environments: email, web, and word processor. The task required respondents to view a number of emails, identify relevant requests, and submit three meeting room requests using a simulated online reservation site. Meanwhile, a conflict between one request and existing schedule presented an impasse that respondents had to resolve. In the interactive environment, test takers could switch among the three environments, go back and forth to understand the meeting room requests, make reservations or changes, and copy and paste the key information in the word processor environment.[3] An interim score was evaluated based only on the meeting reservation webpage. According to the scoring rule, full credit (3 points) was granted when the respondents correctly submitted all three meeting room requests, and partial credit (1 or 2 points) was given when only one or two requests were submitted successfully. No credit (0 points) was given when none of the requests was correctly fulfilled.

According to the PIAAC technical report (OECD 2016), item U02 was one of the most difficult in the PSTRE domain, with difficulty and discrimination parameter estimates[4] of 0.78 and 1.18, respectively, ranking at difficulty level 3. In the U.S. sample, 932 (70%) test takers received no credit, 294 (22%) received partial credit,

positioned in stage 1 and stage 2, respectively. Each of the modules contains seven items without overlap to each other. The seven items within one module has a fixed position. More details about PIAAC test design refer to PIAAC technical report (OECD 2016).

[2] Process data extracted from the log file and response data from the background questionnaire could be linked with the unique respondent IDs. However, given possible technical issues in data collection, there might exist cases with only process data or only background variables. These cases had to be discarded during analysis as data could not be matched.

[3] Word processor was an optional environment instead of a compulsory one, designed to help the respondents summarize information extracted from the email requests.

[4] Two-parameter-logistic item response modeling was applied in the PIAAC data analysis to estimate the latent trait of test takers' problem-solving skills. The parameter estimates presented here are the common international parameters generally used across countries. For details on data analysis modeling in PIAAC, refer to the PIAAC technical report (OECD 2016).

Table 10.1 Descriptive statistics of number of actions and response time (in minutes) in U02

Features	Mean	SD	Min	Max
Number of actions on U02	34.06	33.89	0.00	194.00
Response time on U02	3.60	3.47	0.09	45.07

and only 114 (9%) received full credit. To explore the difference between test takers who got at least part of the item correct and those who received no credit, the polytomous scores were dichotomized by collapsing partial and full credit in the present study.

Further, the item U02 required a relatively long sequence to solve the task. On average, respondents took 34 actions over 3.6 minutes to complete the task.[5] It is also noted that the distributions of the number of actions and response time were widely spread out. As presented in Table 10.1, the longest sequence in this item used 194 actions over 45 minutes, while the shortest sequence was zero actions and 0.09 minutes: obviously a quick skipping behavior resulting in a missing response. These statistics implied that the behavioral patterns and strategies differed considerably by test takers. This sparse distribution would also impact the feature extraction, which is discussed in detail in the next section.

There are four reasons we selected item U02 as an example. First, as mentioned above, this rather difficult item could potentially provide more information to identify reasons for failure when tracking respondents' process data. Researchers have found that for an item that is difficult but not extremely so, test takers tend to demonstrate more heterogeneous use of strategies, aberrant response behavior, and variant use of response time (e.g., de Klerk et al. 2015; Goldhammer et al. 2014; Vendlinski and Stevens 2002). Second, this item consisted of multiple environments (email, web, and word processor). The action sequences are expected to be more diverse in a problem-solving item with multiple environments than an item with a single environment. Hence, it is possible to extract more information from this item. Third, item U02 had a fixed position in the middle of the PS2. Compared to items at the beginning or the end, items in the middle of a booklet are less likely to demonstrate position effect (e.g., Wollack et al. 2003). Lastly, item U02 shared environments with most items in PS2. This provided the possibility to further investigate the consistency of problem-solving strategies across items for each individual.

10.2.3 Features Extracted from Process Data

10.2.3.1 N-Gram Representation of Sequence Data

The strategy for analyzing item U02 was motivated by the methodologies and applications in natural language processing (NLP) and text mining (e.g., He et al. 2012;

[5] There is no time limitation in the PIAAC cognitive test.

Sukkarieh et al. 2012; He et al. 2017). We chose the n-grams model to disassemble the sequence of data while retaining the sequential order. As He and von Davier (2015, 2016) introduced, unigrams—analogous to the language sequences in NLP—are defined as “bags of actions,” where each single action in a sequence collection represents a distinct feature. An n-gram is defined as a contiguous sequence of *n* words in text mining; similarly, when analyzing action sequences from process data, an n-gram can be defined as a sequence of *n* adjacent actions (Manning and Schütze 1999). Bigrams and trigrams are defined as action vectors that contain either two or three ordered adjacent actions, respectively. For instance, here is a typical sequence for email review actions: “MAIL_VIEWED_4, MAIL_VIEWED_2, MAIL_VIEWED_1”. The unigram is each of the three separate actions (e.g., “MAIL_VIEWED_4”), a bigram is two adjacent actions as one unit, (e.g., “MAIL_VIEWED_2, MAIL_VIEWED_1”), and the trigram is three adjacent actions as one unit (e.g., “MAIL_VIEWED_4, MAIL_VIEWED_2, MAIL_VIEWED_1”). Of note is that the n-gram method was productive in creating features from sequence data without losing too much information in terms of the order in the sequence (He et al. 2018). This approach is a widely accepted tool for feature engineering in fields such as NLP and genomic sequence analysis.

A total of 34 actions (i.e., unigrams) were defined for this item and are listed in Table 10.2. The interpretation describing each action is presented as well. The frequency of sequences that contain the action by each row is shown in the right-hand column.

Besides the unigram features, we also included the total response time and the number of actions as features in the cluster analysis. These two features also showed up in a preliminary principal component analysis as the most influential features with the highest loadings. This resulted in 36 features altogether. Given concerns about the low frequency of bigrams and trigrams, the features from mini sequences were not used in the cluster analysis in this study.

10.2.3.2 Term Weights

Three types of term weights were used in the current study: sampling weights as well as between- and within-individual weights. Between-individual weights highlight how different the frequency of a certain action is among individuals, whereas within-individual weights capture how some actions are used more often than others by an individual. Regarding between-individual differences, a popular weighting method in text mining, inverse document frequency (IDF; Spärck Jones 1972), was renamed as inverse sequence frequency (ISF) and adapted for estimating the weight of each *n-gram*. ISF is defined as $ISF_i = \log(N/sf_i) \geq 0$, where N denotes the total number of sequences in the sample, which is the same as the total number of test takers, and sf_i represents the number of sequences containing action i. A large ISF reflects a rare action in the sample, whereas a small ISF represents a frequent one.

Within-individual differences had to be considered when an individual took some actions more often than others. Although more frequent sequences are more impor-

Table 10.2 Description and frequency of unigrams

No.	Features	Description	Frequency
1	FOLDER_VIEWED	View a folder	5,762
2	ENVIRONMENT_WB	Go to web environment	4,715
3	ENVIRONMENT_MC	Go to email environment	4,317
4	MAIL_VIEWED_1	View 1st email	2,725
5	HISTORY_VIEWCALENDAR	Go to calendar tab in web environment	2,190
6	MAIL_VIEWED_3	View 3rd email	1,968
7	HISTORY_RESERVATION	Go to reservation tab in web environment	1,935
8	COMBOBOX_ROOM	Choose a room when filling out a room request	1,891
9	MAIL_VIEWED_4	View 4th email	1,698
10	MAIL_VIEWED_2	View 2nd email	1,544
11	MAIL_MOVE	Move an email	1,499
12	NEXT_INQUIRY	Go to next item	1,371
13	START	Start item U02	1,326
14	COMBOBOX_START_TIME	Choose start time when filling out a room request	1,312
15	COMBOBOX_END_TIME	Choose end time when filling out a room request	1,304
16	COMBOBOX_DEPT	Choose department when filling out a room request	1,296
17	HISTORY_MEETINGROOMS	Go to meeting room details tab in web environment	1,058
18	ENVIRONMENT_WP	Go to word processor environment	987
19	SUBMIT_RESERVATION_FAILURE	Submit a reservation request unsuccessfully	987
20	SUBMIT_RESERVATION_SUCCESS	Submit a reservation request successfully	971
21	HISTORY_UNFILLED	Go to unfilled tab in the web environment	551
22	SUBMIT_UNFILLED	Submit an unfilled request	414
23	FOLDER	Do folder-related actions (i.e., create/delete a folder)	332
24	HISTORY_HOME	Click on the home button in the web environment	244
25	CHANGE_RESERVATION	Change an existing reservation	227
26	KEYPRESS	Type in word processor environment	152

(continued)

Table 10.2 (continued)

No.	Features	Description	Frequency
27	REPLY	Reply an email	118
28	CANCEL	Click on cancel button	111
29	HELP	Use help function	87
30	COPY	Use copy function	42
31	SEARCH	Use search function	38
32	SORT	Use sort function	21
33	PASTE	Use paste function	15
34	BOOKMARK	Do bookmark-related actions (i.e., add/delete a bookmark)	13

tant than less frequent ones for each individual, the raw frequencies of these action sequences often overestimate their importance (He and von Davier 2015). To account for within-individual differences in the importance of action sequences, a weighting function was employed $f\left(tf_{ij}\right) = 1 + \log\left(tf_{ij}\right)$, where $tf_{ij} > 0$ represents the frequency of action i in each individual sequence j (Manning and Schütze 1999). Combining the between- and within-individual weights, the final action weight can be defined as $weight(i, j) = \left[1 + \log\left(tf_{ij}\right)\right] \log(N/sf_i)$ for $tf_{ij} \geq 1$ (He and von Davier 2015, 2016). Compared to raw frequency, this weighting mechanism was applied for attenuating the effect of actions or action vectors that occurred too often to be meaningful.

The sampling weights were also taken into consideration in this study. In fact, we conducted the cluster analyses both with and without sampling weights, and the differences were marginal. Therefore, we report results only with sampling weights in the next section.

10.2.4 Clustering Sequence Data

Clustering has been widely recognized as a powerful unsupervised data mining approach for grouping similar data points. Unlike supervised learning approaches that typically train a model on known input (data and labels) to predict future outputs, unsupervised learning approaches focus on finding hidden patterns or intrinsic structures in input data (Manning and Schütze 1999). Sequence clustering aims at partitioning sequences into meaningful clusters consisting of similar sequences (Ferreira et al. 2007). It has been applied in various fields, such as gene structure exploration in biology, students' learning progression in education, and pattern recognition in industrial engineering.

To cluster sequence data, it is important to choose a clustering algorithm that is appropriate for the characteristics of the data and sequence features (Dong and Pei

2007). Some popular clustering methods include hierarchical clustering (e.g., Huang et al. 2010; Johnson 1967; Navarro et al. 1997), graph-based clustering (e.g., Kawaji et al. 2001; Felzenszwalb and Huttenlocher 2004), K-means (e.g., Bustamam et al. 2017; Gasch and Eisen 2002; Park et al. 2008), and others. Hierarchical clustering is a method of cluster analysis that, as the name indicates, seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types. They are: (1) agglomerative (Johnson 1967)—a "bottom up" approach in which each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy, and (2) divisive (MacNaughton-Smith et al. 1964)—a "top down" approach in which all observations start in one cluster and splits are performed recursively as one moves down the hierarchy. Graph-based clustering algorithms generally involve two major steps. In the first step, a weighted graph is constructed from the sequences. In the second, the graph is segmented into subgraphs that correspond to the clusters (Dong and Pei 2007). K-means is one of the simplest learning algorithms to solve clustering problems. The procedure follows a straightforward way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea of the K-means algorithm is to discover K (nonoverlapping) clusters by finding K centroids ("central" points) and then assigning each point to the cluster associated with its nearest centroid (Jyoti and Singh 2011).

Our current study adopted the K-means algorithm to cluster the behavioral patterns from one PSTRE item U02 based on features extracted from process data. The reasons for choosing this algorithm can be explained from three aspects: First, K-means is efficient in terms of computational cost even with a large number of variables, which renders wider applications possible in large-scale assessments, especially for complex multidimensional data structures in process data. Second, observations can switch from one cluster to another when the centroids of the clusters are recomputed. This shows that K-means is able to recover from potential mistakes in clustering. However, it also indicates that results from K-means could be strongly influenced by the selection of initial seeds (e.g., Arthur and Vassilvitskii 2007). Therefore, the impact of selecting initial seeds should be carefully examined before interpreting the results, as we did in this study. Third, results of K-means are easily interpretable. Each observation belongs to only one cluster, and the centroids of the clusters are expressed on the scales of the variables. More details about the analytic strategy and algorithms are introduced in the next section.

10.2.5 K-Means Clustering

The K-means algorithm (Lloyd 1982) was adopted for the cluster analysis in the current study. This method starts with k arbitrary centroids and seeks to minimize the squared difference between observations in the same cluster. A cluster centroid is typically the mean or median of the points in its cluster and "nearness" is defined by a distance or similarity function. Ideally the centroids are chosen to minimize the total "error," where the error for each point is given by a function that measures the discrepancy between a point and its cluster centroid, for example, the squared distance. Note that a measure of cluster "goodness" is the error contributed by that cluster (Alphaydin 2009).

The basic K-means algorithm for finding K clusters is as follows:

1. Select K points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids do not change (or change minimally).

The first step is to define k centroids, one for each cluster. These centroids should be placed with careful consideration because different locations cause different results. The best choice is to place them as far away from each other as possible. The next step is to take each point belonging to a given data set and assign it to the nearest centroid. When no point is pending, the first step is completed and an early group membership is done. At this point we need to recalculate k new centroids. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. This generates a loop. As a result of this loop, we could notice that the k centroids may change their location step by step until no more changes occur. In other words, the centroids do not move anymore. Finally, this algorithm aims at minimizing a function of a matrix, for instance, a squared error function (Steinbach et al. 2004).

Unlike the hierarchical algorithms that produce a nested sequence of partitions, K-means is one of the nonhierarchical algorithms that often start out with a partition based on randomly selected seeds, and then refine this initial partition (Manning and Schütze 1999). The initial cluster centers for K-means are usually picked at random. Whether the choice of initial centers is important or not depends on the structure of the set of objects to be clustered (Jyoti and Singh 2011).

In this study, we used 36 features—34 unigrams plus response time and total number of action sequences—extracted from the process data of item U02 to partition test takers into clusters using the K-means clustering method. An appropriate number of clusters, k, was selected based on the change in the total within-cluster sum of squares. As noted previously, one potential uncertainty of K-means is that the clustering results could be strongly influenced by the selection of initial seeds (e.g., Arthur and Vassilvitskii 2007). Therefore, the stability of the cluster membership was examined to maximize the generalizability of the results. Further, clusters were interpreted based on the centroids of the 36 features. We explored the homogeneous characteristics of the clusters, as well as the relationship between cluster membership and proficiency level and/or correctness of U02.

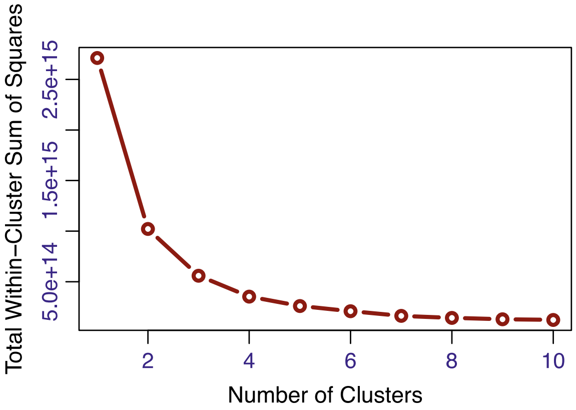


Fig. 10.1 Total within-cluster sum of squares at different number of clusters for the item U02

10.3 Results

10.3.1 Cluster Determination

The basic idea behind cluster partitioning methods, such as K-means clustering, is to define clusters such that the total intra-cluster variation (known as total within-cluster variation or total within-cluster sum of squares) is minimized. We used three methods to determine the optimal number of clusters for the data set, the elbow method (e.g., Ketchen and Shook 1996; Thorndike 1953), the average silhouette method (e.g., Kaufman and Rousseeuw 1990), and the hierarchical clustering method (e.g., Ward 1963). These three methods were chosen to provide insights about the structure of the data through visualization and statistical measures and to mutually validate the results from each.

For the elbow method, substantial drops in total within-cluster sum of squares were present when the number of clusters was set from one to three. After the "elbow point" of three, the changes became marginal despite an increasing number of clusters (see Fig. 10.1).

The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering. The average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k. Given the sample in hand, the highest average silhouette width was shown when two clusters were chosen.

The hierarchical clustering method seeks to form a hierarchy of clusters either through bottom-up or top-down approach. Such an algorithm either starts with all observations in different clusters and merges them into clusters, or starts with all observations in one cluster and gradually partitions into clusters. In the current study, a bottom-up hierarchical clustering method was employed. Given the sample in hand, the results from this method were that the optimal cluster number could be between two and four. Based on mutual validation by these three methods as well as considerations on cluster sample size and interpretability of the results, we chose the three-cluster solution for further investigations. After determining the number of clusters, the clustering analysis was rerun and results were reported based on this.

As mentioned previously, K-means clustering usually starts from a random selection of initial seeds even though using more discriminative seeds that are able to separate the clusters is highly recommended as the initial partition (Arthur and Vassilvitskii 2007). Although many sets are well behaved and most initializations will result in clustering of about the same quality (Manning and Schütze 1999), it would be wise to examine the stability of cluster membership to maximize the generalizability before interpreting the clustering results.

We checked the stability of cluster membership with 100 different initial seeds in the item U02. Among the 1,326 test takers, 1,262 (95%) had no changes in cluster membership in the 100 replications. Only 64 (5%) were assigned to a different cluster in at most 10% of the replications. Overall, only 0.3% of the test-taker-replication combinations demonstrated uncertainty in the cluster membership. This suggested that the clustering results had very little dependence on initial seeds and thus the seeds could be ignored in this study.

We list the centroids of the three-cluster solution in Table 10.3. Note that the term weights and sampling weights were taken into account when the values of centroids were computed. For the 34 unigrams, values presented in Table 10.3 were based on action frequencies weighted by term weights and sampling weights; for the number of actions and response time on U02, the two features were weighted by sampling weights before computing the centroids. In general, Cluster 1 had the lowest weighted frequencies and means in almost all features and Cluster 3 had the highest ones, while Cluster 2 placed between Cluster 1 and Cluster 3. The action unigrams "NEXT_INQUIRY" and "START" had centroids at zero across all three clusters, suggesting that all test takers had taken these two actions, which led to them providing little information in the analysis. When all test takers perform the same action, the ISF of an action would be zero by definition. Thus, these two unigram features did not actually contribute in the clustering because of the zero information. As expected, the number of actions and response time appeared to be the most dominant features in clustering. The reason is probably that these two variables are of a different granularity than the others, as they summarize information for the entire sequence, rather than a partial contribution made by a single action. These two features also showed up in a preliminary principal component analysis as the most influential features with the highest loadings.

The three clusters could be interpreted as test takers with the least, medium, and most effort. The least-action cluster had the largest cluster size with 853 (64%) of the test takers in the analytical sample, the median action cluster had 398 (30%) test takers, and only 75 (6%) were in the most-action cluster (see Table 10.4). This indicated that only a small group of test takers had a great number of actions and spent a long time exploring U02, whereas the majority clustered around fewer actions and a much shorter time.

Table 10.3 Cluster centroids for a three-cluster solution

No.	Features	Clusters		
		1	2	3
1	FOLDER_VIEWED	20,365.6	40,224.5	73,644.8
2	ENVIRONMENT_WB	11,062.0	65,848.4	134,117.1
3	ENVIRONMENT_MC	11,375.0	59,744.3	124,133.6
4	MAIL_VIEWED_1	8,625.7	23,554.9	42,659.6
5	HISTORY_VIEWCALENDAR	8,296.1	61,208.4	130,657.0
6	MAIL_VIEWED_3	7,983.3	41,671.1	84,185.9
7	HISTORY_RESERVATION	6,972.4	53,034.1	117,019.4
8	COMBOBOX_ROOM	6,020.8	54,476.1	110,608.7
9	MAIL_VIEWED_4	8,606.6	35,180.8	67,087.3
10	MAIL_VIEWED_2	7,891.8	33,636.2	65,864.9
11	MAIL_MOVE	18,947.5	42,469.4	87,984.2
12	NEXT_INQUIRY	0.0	0.0	0.0
13	START	0.0	0.0	0.0
14	COMBOBOX_START_TIME	5,498.0	47,928.2	101,684.2
15	COMBOBOX_END_TIME	5,581.8	47,942.1	103,098.3
16	COMBOBOX_DEPT	5,556.0	48,052.1	101,711.1
17	HISTORY_MEETINGROOMS	5,848.3	43,725.6	108,077.0
18	ENVIRONMENT_WP	7,738.8	33,937.1	79,654.0
19	SUBMIT_RESERVATION_FAILURE	4,048.2	46,768.2	109,482.7
20	SUBMIT_RESERVATION_SUCCESS	4,797.0	42,081.0	85,547.9
21	HISTORY_UNFILLED	4,213.2	36,222.2	91,450.9
22	SUBMIT_UNFILLED	3,589.7	34,291.9	69,265.5
23	FOLDER	6,750.6	25,942.1	62,512.5
24	HISTORY_HOME	3,808.0	18,614.7	50,805.3
25	CHANGE_RESERVATION	1,522.0	23,168.2	73,968.0
26	KEYPRESS	2,880.5	12,713.7	65,743.1
27	REPLY	2,936.5	12,319.8	30,153.8
28	CANCEL	3,250.7	13,530.1	37,320.8
29	HELP	3,477.5	10,343.4	17,039.6
30	COPY	897.1	7,628.4	38,517.4
31	SEARCH	2,278.3	3,529.5	18,895.4

(continued)

Table 10.3 (continued)

No.	Features	Clusters		
		1	2	3
32	SORT	949.7	4,759.2	5,540.5
33	PASTE	780.4	1,494.6	33,561.6
34	BOOKMARK	550.6	2,875.4	9,264.9
35	Number of actions on U02	453,665.4	2,241,595.2	5,225,357.2
36	Response time on U02	58,391.4	244,418.8	475,306.6
	Frequency	853	398	75

Note "NEXT_INQUIRY" and "START" show 0 cluster centroids across all three clusters. As all participants used them, they had zero term weights and did not contribute in the clustering analysis

Table 10.4 Cluster size of a three-cluster solution

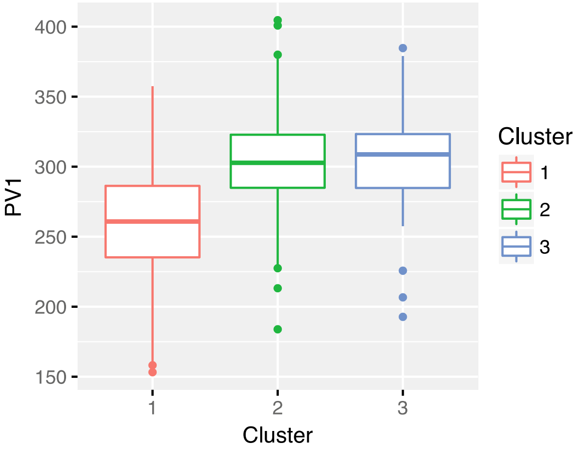
U02score	Clusters		
	1	2	3
0	760	139	23
1	93	259	52
Total	853	398	75

Note The U02score has combined the partial and full credit into "1"

10.3.2 Cluster Membership and Proficiency Level

Based on the clusters derived from process data as described above, we investigated the relationships between cluster membership and PSTRE proficiency level as well as employment-related variables. To increase the accuracy of the cognitive measurement for various subpopulations and the population as a whole, PIAAC uses plausible values—which are multiple imputations—drawn from a posteriori distribution by combining the item response scaling of the cognitive items with a latent regression model using information from the background questionnaire. The "plausible value" methodology correctly accounts for error (or uncertainty) at the individual level by using multiple imputed proficiency values (plausible values) rather than assuming that this type of uncertainty is zero (for details about how the set of plausible values is generated and interpreted, refer to OECD 2016). In PIAAC, 10 plausible values (PV) are generated for each domain as the estimates of scale scores. As all 10 PVs showed similar patterns, we used only the first PV (PV1) as an example for illustration purposes. Figure 10.2 depicted the association between clusters and PSTRE proficiency level (PV1). To explore whether significant differences existed among the clusters regarding PV1, we conducted one-way analysis of variance (ANOVA). Results showed that the three clusters had significantly different proficiency levels as measured by PV1, $F(2{,}1323) = 254.6$, $p < 0.001$.

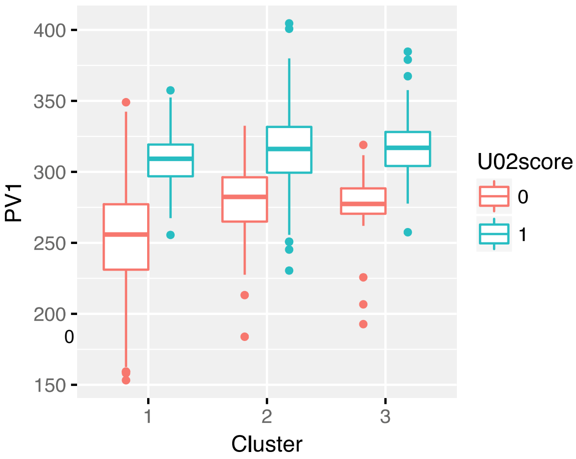
Fig. 10.2 A boxplot of PV1 by cluster membership



Given the results from ANOVA, we further conducted a post hoc pairwise comparison for the three clusters. Given concerns on the unequal sample size by clusters, the pairwise comparison *t*-test method introduced by Benjamini and Hochberg (1995) was employed. This method controls the false discovery rate (the expected proportion of false discoveries) among rejected hypotheses. The false discovery rate is a less stringent condition than the family-wise error-rate-based methods, so this method is more powerful than the others. Notably, a remarkable increase was observed in PV1 from Cluster 1 to Cluster 2, for which the first quartile of Cluster 1 was approximately at the same level as the third quartile of Cluster 2. However, the increase from Cluster 2 to Cluster 3 was marginal. Results showed no significant differences between these two groups. Given the similar proficiency level between clusters 2 and 3, but shorter action sequences and response time in Cluster 2, this might be interpreted as a higher efficiency in Cluster 2 to solve the item U02. That is, both clusters 2 and 3 were more likely to be able to solve the item, but with different strategies and paces.

We further plotted the PV1 distributions by correct and incorrect groups for item U02 for each cluster (see Fig. 10.3). The sample size by each group nested in clusters was reported in Table 10.4. As expected, the majority of those in Cluster 1 did not answer correctly, since only a few actions and a short time were taken. Clusters 2 and 3 tended to have more test takers who were successful in solving U02. In general, across the three clusters, the PV1 of test takers who responded correctly to item U02 was consistently higher than those who responded incorrectly, although the mean difference among the correct groups in pairwise comparisons was not statistically significant. This suggested that the actions or response time did not make a significant impact on how respondents correctly solved the item. Besides, the correct group in Cluster 1 actually could be interpreted as the most efficient group in finding the correct answer since they used the shortest action sequences and response times

Fig. 10.3 A boxplot of PV1 by U02score nested in clusters



across all correct groups by clusters.[6] Comparatively, a significant difference was found among the incorrect groups in the one-way ANOVA, resulting in $F(2, 919) = 55.2$, $p < 0.001$. Similar to the general pattern found in Fig. 10.3, substantial differences were found between Cluster 1 and the other two clusters, whereas little difference was found between Cluster 2 and Cluster 3.

These findings suggested that the correct group applied various problem-solving strategies to obtain a correct answer, and the choice of strategy was not necessarily associated with PSTRE skills in the correct group. As noted above, a small group of test takers in Cluster 1 was able to use only a few actions to solve U02 in a short time, and that group's PSTRE scores were similar to those who applied many more actions. While adopting more actions might be an indication of high motivation to extensively explore the item, it could also signify that the test taker used less efficient strategies when those actions became excessive. For the incorrect group, however, the number of actions and time spent on the item could be informative regarding a test taker's PSTRE skills. A test taker who put more effort into solving U02, even though he or she failed, was more likely to have higher PSTRE skills.

10.3.3 Cluster Membership and Employment-Based Background Variables

To understand the profiles for each cluster and the common characteristics that might be shared within the cluster, we further explored the relationship between problem-solving strategies and background variables. In particular, we focused

[6] In the PIAAC complex problem-solving items, multiple choice items were seldom employed. Item types such as short responses, drag-and-place, and web navigations were used to keep the guessing effect as low as possible.

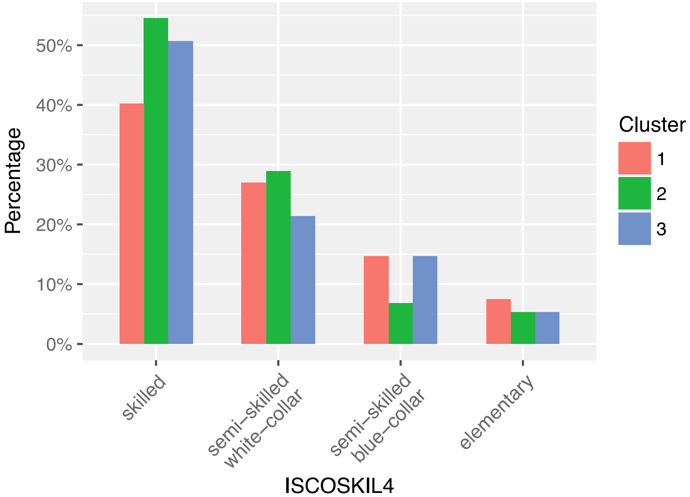


Fig. 10.4 Distribution of ISCOSKIL4 in the three clusters (Percentages in plots represent the percentages of a certain employment status/skill-use level in a cluster.)

on employment-related characteristics: variables related to occupational category, monthly income, work-related skills, age, and education.

Figure 10.4 shows the distribution of the degree of skills (ISCOSKIL4) in all three clusters, which indicates the occupational classification of the test taker's last or current job. Cluster 2 appeared to have the largest proportion of those in skilled and semi-skilled white-collar occupations, while Cluster 1 had the smallest proportion. In contrast, clusters 1 and 3 both had the largest proportion in semi-skilled blue-collar occupations, while Cluster 2 had the smallest. As expected, Cluster 1 had the largest proportion in elementary-level occupations, while the other two clusters shared equally low proportions.

As for the monthly earnings variable (EARNMTHALLDCL) in Fig. 10.5, Cluster 2 and Cluster 3 showed a substantial proportion in the highest earning deciles, from 7th to 10th, whereas Cluster 1 tended to have higher percentages in the lower earning deciles. Two exceptions were found in the first and fourth deciles, in which most test takers were grouped in Cluster 2 and Cluster 3. Despite the general pattern that earnings were mainly positively related to the number of actions and response time spent on the item, some test takers who were younger or at the early stage of their careers may have had lower salaries but higher problem-solving capacity.

Variables regarding work-related skill use also demonstrated similar patterns. Figure 10.6 depicts the distribution of variables for skills related to work: ICT (ICTWORK), numeracy (NUMWORK), reading (READWORK), and writing (WRITWORK). These skills were each divided into five levels in intervals of 20 percentage points, plus a separate category for nonresponses. Cluster 1 was more likely to include test takers in the lower skill-use levels (<40%), while more test takers with high skill use levels (>40%) were in Cluster 2 and Cluster 3. Notably, even though those in

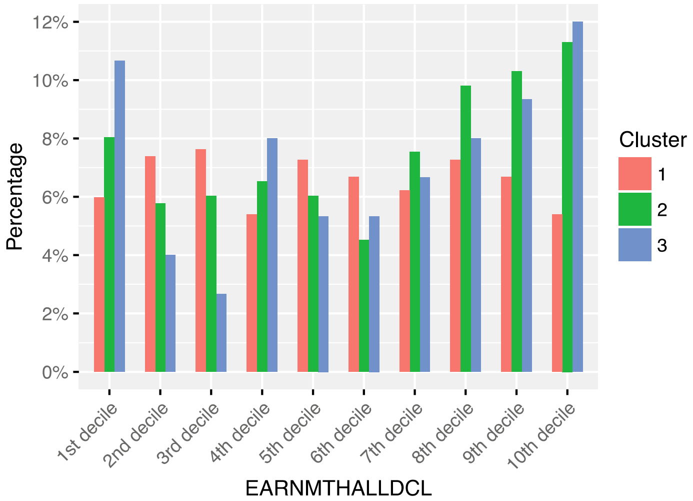


Fig. 10.5 Distribution of EARNMTHALLDCL in the three clusters

Cluster 3 had the largest number of actions and spent the longest time on the item, this group consisted of more test takers in the lower skill levels than Cluster 2. This is also consistent with the finding in the occupational classification that Cluster 3 included a large proportion of test takers in semi-skilled blue-collar occupations, which did not necessarily require higher levels of ICT, numeric, reading, or writing skill use. In addition, considering the item context related to a practical working environment—a meeting room reservation, which is more or less an assistant-style task—this item might not have required as high a level of skills as did other more complex items.

Figure 10.7 exhibits the distribution of the three clusters by five age groups: 24 or less, 25–34, 35–44, 45–54, and 55 plus. Over 30% of test takers in Cluster 3 were younger than 24 years old, representing the highest proportion for this age group. Cluster 2 had the highest proportion in the 25–34 age group, while Cluster 1 had the largest proportion in the oldest group (over 55). This finding provided another perspective for interpreting the pattern observed from process data. Since a large proportion of test takers in Cluster 3 were younger than test takers in the other two clusters, different behaviors could be expected. Compared to the older test takers, younger test takers tended to be more active in human-computer interactions, more familiar with manipulating computer systems, and learned faster when encountering new interfaces. Furthermore, they were expected to exhibit more curiosity about exploring the item, which could increase the number of actions and response time.

Lastly, we took educational backgrounds into consideration. Figure 10.8 presents the distribution of six education levels (EDCAT6) for each cluster. Cluster 1 had the highest percentage for those with a lower/upper secondary level or less of education, whereas Cluster 3 had the highest percentages in the postsecondary and tertiary

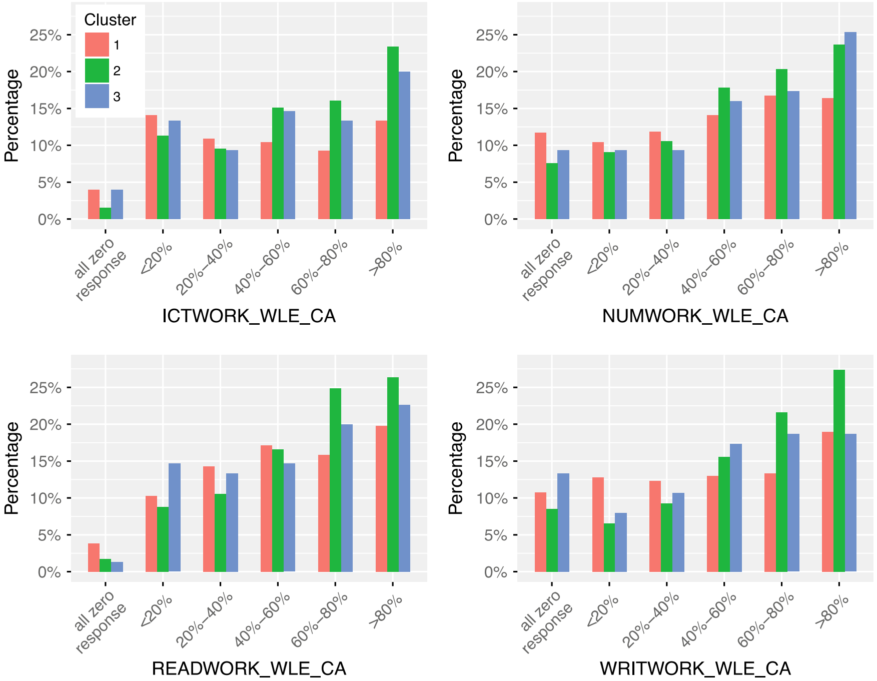


Fig. 10.6 Distribution of ICTWORK_WLE_CA, NUMWORK_WLE_CA, READWORK_WLE_CA, and WRITWORK_WLE_CA in the three clusters

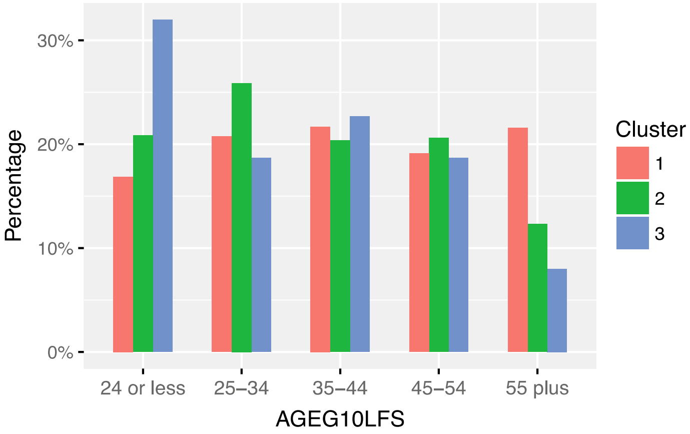


Fig. 10.7 Distribution of AGEG10LFS in the three clusters

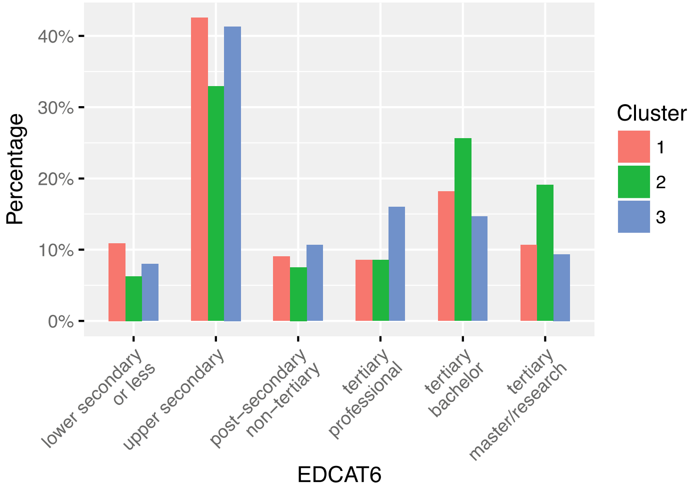


Fig. 10.8 Distribution of EDCAT6 in the three clusters

professional degree levels. Cluster 2 had the greatest proportions of test takers with a bachelor's degree or higher. Thus, test takers in Cluster 1 were the lowest performing group in PSTRE, with the lowest education level overall. Although Cluster 3 provided a slightly higher median PV1, the percentages in the bachelor's degree or higher categories were the lowest. It turned out that test takers in Cluster 3 might not possess the highest education level, but the openness to experience enabled them to score well in PSTRE.

10.4 Discussion

The current study shows an example of clustering students' action sequences and response times, which reveals how these data can provide important information beyond students' item responses. Such information has the potential to help educators understand student motivation and the specific struggles that students might have during a task, and could shed light on the effectiveness of a particular teaching practice.

To summarize, we grouped test takers into three clusters based on 36 features extracted from process data. We found that more actions and longer response time in general were associated with higher PSTRE scores, but such a pattern was more evident when the test takers did not answer U02 correctly. In other words, it was possible to obtain a correct answer with various strategies, but when test takers answered incorrectly, process data could be informative about the extent to which interventions would be needed. In fact, this finding was reiterated when we conducted the same clustering method on other problem-solving items. In the examination of process data patterns with background variables, it was found that test takers who did not put much effort into solving the item tended to work in semi-skilled or elementary occupations, have lower monthly income and lower work-related skill use, be of a higher age, and have lower education. This group of test takers might be in need of further education or intervention.

An interesting finding was that the group with the highest action frequencies, response time, and PSTRE scores did not necessarily possess the highest income, work-related skill use, or education level. The youngest group with longer response time and action sequences was distinct from other test takers in that these individuals were the most explorative or adventurous test takers and were willing to engage in a large number of different actions in solving a problem. This characteristic was likely to relate to higher PSTRE skills.

Besides the merits of this study, some limitations are also worth discussing. First, the response time and number of actions seemed to play a dominant role in the clustering in the current study. It would be worthwhile to try the standardized variables of response time and number of actions in the future study to check whether different results may occur.

Second, the information contributed from the key actions (unigrams) might be difficult to distinguish and not show up very clearly in this clustering analysis. Previous studies have shown that the mini sequences of bigrams and trigrams are more informative than unigrams and were robust classifiers in distinguishing subgroups (He and von Davier 2015). We also extracted bigrams and trigrams for item U02 in this study. However, because of the sparse distribution of action combinations, which resulted in over 40,000 n-grams altogether, it would be very challenging to use this large number of features in its entirety in the cluster analysis. Meanwhile, given the low frequency (lower than five times) of the majority of bigrams and trigrams, we had to exclude them from further cluster analysis to ensure the reliability of calculation. A substantial increase in sample size would help enhance the frequency of mini sequences in future studies in clustering.

Third, we conducted 100 replications with different initial seeds to determine the cluster membership in this study, but need to make cross-validation to examine the clustering performance in further studies. As in this study we mainly explored the relationship between behavioral patterns and proficiency level, a formal classification based on clustering results is not that essential. However, clustering is usually regarded as a “prelude” to classification (Dong and Pei 2007). It would be more appropriate to include a further classification or other means of validation to better evaluate the cluster results.

Fourth, the current study focused on only one PSTRE item. It is not clear yet whether the respondent may choose consistent strategies in solving other items with similar environments. It would be interesting to further examine the consistency of each individual across different items to better generalize the findings from the current study. Some explorations that have been done in this direction may benefit the further analysis in consistency investigation. For instance, He et al. (2019) used the longest common subsequence (LCS) method, a sequence-mining technique commonly used in natural language processing and biostatistics to compare the action sequences followed by PIAAC respondents to a set of "optimal" predefined sequences identified by test developers and subject matter experts, which allows studying problem solving behaviors across multiple assessment items.

Finally, this exploratory study was conducted only based on the U.S. sample in the PSTRE domain of PIAAC. It would be desirable to include multiple countries in a future study to examine the cross-country differences in a general way. Further, it would also be interesting to use process data to explore the relationship between problem-solving skills and numeracy and literacy to better understand the consistency of test takers' behavioral patterns in different domains.

In conclusion, from this study, we have learned that different problem-solving strategies and behavioral patterns may influence proficiency estimates, and are more impactful in the groups that fail to give correct responses than the groups that succeed in answering correctly. Additionally, groups with different backgrounds may show different problem-solving patterns. This suggested that various solutions would need to be properly adapted to different groups to improve their problem-solving skills. In future studies, we recommend researchers further explore the methods to better model the relationship between behavioral patterns and proficiency estimates in large-scale assessments, and challenge other researchers to develop models in estimating problem-solving proficiency more accurately by possibly integrating the new data source from process data.

References

Alpaydin, E. (2009). *Introduction to machine learning*. Cambridge, MA: MIT Press.

Arthur, D., & Vassilvitskii, S. (2007). K–means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57,* 289–300.

Bustamam, A., Tasman, H., Yuniarti, N., Frisca, & Mursidah, I. (2017). Application of K-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV). In *AIP Conference Proceedings* (Vol. 1862, No. 1, p. 030134). AIP Publishing.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem–solving process data: An event history analysis approach. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00486.

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education, 85,* 23–34.

Dong, G., & Pei, J. (2007). *Sequence data mining* (Vol. 33). Berlin: Springer Science & Business Media.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision, 59*(2), 167–181.

Ferreira, D., Zacarias, M., Malheiros, M., & Ferreira, P. (2007). Approaching process mining with sequence clustering: Experiments and findings. In *International conference on business process management* (pp. 360–374). Berlin, Germany: Springer.

Gasch, A. P., & Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy K-means clustering. *Genome Biology, 3*(11), research0059-1.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608.

He, Q., Borgonovi, F., & Paccagnella, M. (2019, forthcoming). *Using process data to understand adults' problem-solving behaviours in PIAAC: Identifying generalised patterns across multiple tasks with sequence mining. OECD Research Paper.*

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In A. van der Ark, D. Bolt, S. Chow, J. Douglas, & W. Wang (Eds.), *Quantitative psychology research: Proceedings of the 79th annual meeting of the psychometric society* (pp. 173–190). New York, NY: Springer.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749–776). Hershey, PA: Information Science Reference.

He, Q., von Davier, M., & Han, Z. (2018). Exploring process data in computer-based international large-scale assessments. In H. Jiao, R. Lissitz, & A. van Wie (Eds.), *Data analytics and psychometrics: Informing assessment practices* (pp. 53–76). Charlotte, NC: Information Age Publishing.

He, Q., Veldkamp, B. P., & de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Research, 198*(3), 441–447.

He, Q., Veldkamp, B. P., Glas, C. A. W., & de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment, 24*(2), 157–172.

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics, 26*(5), 680–682.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241–254.

Jyoti, K., & Singh, S. (2011). Data clustering approach to industrial process monitoring, fault detection and isolation. *International Journal of Computer Applications, 17*(2), 41–45.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: John Wiley and Sons.

Kawaji, H., Yamaguchi, Y., Matsuda, H., & Hashimoto, A. (2001). A graph-based clustering method for a large set of sequences using a graph partitioning algorithm. *Genome Informatics, 12,* 93–102.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal, 17*(6), 441–458.

Liao, D., He, Q., & Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: An investigation of us adults' employment status in PIAAC. *Frontiers in Psychology, 10*, 646. https://doi.org/10.3389/fpsyg.2019.00646.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

MacNaughton-Smith, P., Williams, W. T., Dale, M. B., & Mockett, L. G. (1964). Dissimilarity analysis: A new technique of hierarchical sub-division. *Nature, 202*(4936), 1034.

Navarro, J. F., Frenk, C. S., & White, S. D. (1997). A universal density profile from hierarchical clustering. *Astrophysical Journal, 490*(2), 493.

Organization for Economic Co-operation and Development. (2009). *PIAAC problem solving in technology-rich environments: A conceptual framework* (OECD Education Working Paper No. 36). Paris, France: Author.

Organisation for Economic Co-operation and Development. (2010). *New millennium learners project: Challenging our views on ICT and learning*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2011). *PISA 2009 results: Students on line: Digital technologies and performance* (Vol. VI.) http://dx.doi.org/10.1787/9789264112995-en.

Organisation for Economic Co-operation and Development. (2012). *Survey of adult skills (PIAAC)*. Available at http://www.oecd.org/skills/piaac/.

Organisation for Economic Co-operation and Development. (2013a). *Technical report of the survey of adult skills (PIAAC)*. Retrieved from http://www.oecd.org/skills/piaac/_technical%20report_17oct13.pdf.

Organisation for Economic Co-operation and Development. (2013b). *Time for the U.S. to reskill?* Paris, France: OECD Publishing. https://doi.org/10.1787/9789264204904-en.

Organisation for Economic Co-operation and Development. (2016). *Skills matter: Further results from the survey of adult skills*. http://dx.doi.org/10.1787/9789264258051-en. https://www.oecd.org/skills/piaac/Skills_Matter_Further_Results_from_the_Survey_of_Adult_Skills.pdf.

Park, S., Suresh, N. C., & Jeong, B. K. (2008). Sequence-based clustering for web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. *Data & Knowledge Engineering, 65*(3), 512–543.

Rampey, B. D., Finnegan, R., Goodman, M., Mohadjer, L., Krenzke, T., Hogan, J., & Provasnik, S. (2016). *Skills of U.S. unemployed, young, and older adults in sharper focus: Results from the program for the international assessment of adult competencies (PIAAC) 2012/2014: First look* (NCES Report No. 2016–039). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2016/2016039.pdf.

Schleicher, A. (2008). PIAAC: A new strategy for assessing adult competencies. *International Review of Education, 54,* 627–650.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. *New directions in statistical physics* (pp. 273–309). Berlin, Germany: Springer.

Sukkarieh, J. Z., von Davier, M., & Yamamoto, K. (2012). *From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks* (Research Report No. RR-12-25). Princeton, NJ: Educational Testing Service.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika, 18*(4), 267–276.

Vanek, J. (2017). *Using the PIAAC framework for problem solving in technology-rich environments to guide instruction: An introduction for adult educators*. Retrieved from https://piaac.squarespace.com/s/PSTRE_Guide_Vanek_2017.pdf.

Vendlinski, T., & Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *Journal of Technology, Learning and Assessment, 1*(3).

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*(301), 236–244.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40*(4), 307–330.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 11
Reliability Issues in High-Stakes Educational Tests

Cees A. W. Glas

Abstract High-stakes tests and examinations often give rise to rather specific measurement problems. Though nowadays item response theory (IRT) has become the standard theoretical framework for educational measurement, in practice, number-correct scores are still prominent in the definition of standards and norms. Therefore, in this chapter methods are developed for relating standards on the number-correct scale to standards on the latent IRT scale. Further, this chapter focuses on two related issues. The first issue is estimating the size of standard errors when equating older versions of a test to the current version. The second issue is estimating the local reliability of number-correct scores and the extra error variance introduced through number-correct scoring rather than using IRT proficiency estimates. It is shown that the first issue can be solved in the framework of maximum a posteriori (MAP) estimation, while the second issue can be solved in the framework of expected a posteriori (EAP) estimation. The examples that are given are derived from simulations studies carried out for linking the nation-wide tests at the end of primary education in the Netherlands.

11.1 Outline of the Problem

The problem addressed here is that the standard scoring rule in much educational measurement, that is, the number-correct score, is not the same one as the optimal scoring rule that is derived from the IRT model that fits the data. In this chapter, a method is outlined for how to evaluate the consequences of this discrepancy for an important inference that is often made using IRT, that is, the consequences for test equating. To explain this further, we first introduce an IRT model and outline the principle of test equating.

The IRT models used in this chapter are the one-, two- and three-parameter Logistic models. The data are responses of students labeled with an index $n = 1, \ldots, N$ to

C. A. W. Glas (✉)
University of Twente, Enschede, The Netherlands
e-mail: c.a.w.glas@utwente.nl

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_11

items labeled with an index $i = 1, \ldots, K$. To indicate whether a response is available, we define a variable

$$d_{ni} = \begin{cases} 1 \text{ if a response of student } n \text{ to item } i \text{ is available} \\ 0 \text{ if this is not the case.} \end{cases} \tag{11.1}$$

The responses will be coded by a stochastic variable Y_{ni}. In the sequel, upper-case characters will denote stochastic variables and lower-case characters will denote realizations. In the present case, there are two possible realizations, defined by

$$y_{ni} = \begin{cases} 1 \text{ if } d_{ni} = 1 \text{ and student } n \text{ gave a correct response to item } i \\ 0 \text{ if } d_{ni} = 1 \text{ and student } n \text{ did not give a correct response to item } i \\ c \text{ if } d_{ni} = 0, \text{ where } c \text{ is an arbitrary constant unequal 0 or 1.} \end{cases} \tag{11.2}$$

Define the logistic function $\Psi(.)$ as:

$$\Psi(x) = \frac{\exp(x)}{1 + \exp(x)}.$$

In the 3-parameter logistic model (3PLM, Birnbaum 1968) the probability of a correct response depends on three item parameters, a_i, b_i and c_i which are called the discrimination, difficulty and guessing parameter, respectively. The parameter θ_n is the latent proficiency parameter of student n. The model is given by

$$\begin{aligned} P_i(\theta_n) &= c_i + (1 - c_i) + \Psi(a_i(\theta_n - b_i)) \\ &= c_i + (1 - c_i)\frac{\exp(a_i(\theta_n - b_i))}{1 + \exp(a_i(\theta_n - b_i))}. \end{aligned} \tag{11.3}$$

The 2-parameter logistic model (2PLM, Birnbaum 1968) follows by setting the guessing parameter equal to zero, so by introducing the constraint $c_i = 0$. The 1-parameter logistic model (1PLM, Rasch 1960) follows by introducing the additional constraint $a_i = 1$.

Note that in the application of the models in high-stakes situations, the number of proficiency parameters θ_n can become very large. Besides the practical problem of computing estimates of all model parameters concurrently, this also leads to theoretical problems related to the consistency of the estimates (see, Neyman and Scott 1948; Kiefer and Wolfowitz 1956). Therefore, it is usually assumed that the proficiency parameters are drawn from one or more normal proficiency distributions, indexed $g = 1, \ldots, G$, which are often also referred to as population distributions. That is, θ_n has the density function

$$g(\theta_n; \mu_{g(n)}, \sigma^2_{g(n)}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\theta_n - \mu_{n(g)})^2}{\sigma^2}\right), \tag{11.4}$$

Table 11.1 Example of proficiency estimates and their standard errors on two linked tests

Score	Test A				Test B			
	Freq	Prob	θ	$Se(\theta)$	Freq	Prob	θ	$Se(\theta)$
0	156	0.03	−1.91	0.54	6	0.00	−2.40	0.36
1	504	0.13	−1.34	0.50	16	0.00	−2.03	0.34
2	1055	0.34	−0.81	0.47	52	0.02	−1.68	0.33
3	1077	0.56	−0.38	0.45	122	0.04	−1.31	0.33
4	839	0.73	0.02	0.42	261	0.09	−0.96	0.34
5	658	0.86	0.41	0.40	516	0.20	−0.57	0.36
6	367	0.93	0.78	0.37	956	0.39	−0.17	0.37
7	194	0.97	1.15	0.37	1194	0.63	0.25	0.38
8	102	0.99	1.51	0.38	978	0.82	0.71	0.40
9	42	1.00	1.87	0.39	638	0.95	1.19	0.42
10	6	1.00	2.22	0.41	261	1.00	1.73	0.46

where *g(n)* is the population to which student *n* belongs.

Test equating relates the scores on one test to the scores on another test. Consider a simulated example based on the estimates displayed in Table 11.1. The estimates emanate from two tests. A sample of 5000 students of a population A was given a test A consisting of the items $i = 1,\ldots,10$, while a sample of 5000 other students of a population B was given a test B consisting of the items $i = 6,\ldots,15$. So the anchor between the two tests, that is, the overlap between the two tests, consists of 5 items. The anchor supports the creation of a common scale for all parameter estimates. The responses were generated with the 2PLM. The difficulties of the two tests differed: test A had a mean difficulty parameter, $\overline{b}_A$, of 0.68, while the difficulty level of test B, $\overline{b}_B$, was equal to −0.92. The mean of the proficiency parameters θ_n of sample A, μ_A was equal to -0.25, while the mean of the proficiency parameters of sample B, μ_B was equal to 0.25. The variances of the proficiency parameters and the mean of the discrimination parameters were all equal to one.

Suppose that test A has a cutoff score of 4, where 4 is the highest number-correct score that results in failing the test. In the fourth column of Table 11.1, the column labeled θ, it can be seen that the associated estimate on the latent θ-scale is 0.02. We chose this point as a latent-cutoff point, that is, $\theta_0 = 0.02$. If the Rasch model would hold for these data, the number-correct score would be the sufficient statistic for θ. In the 2PLM, the relation between a number-correct score and a θ-estimate is more complicated; this will be returned to below. Through searching for number-correct scores on Test B with θ-estimates closest to the latent cutoff point, we find that a cutoff score 6 on Test B best matches a cutoff score 4 on Test A. This conclusion is consistent with the fact the average difficulty of test A was higher than the average difficulty of test B. On the other hand, the sample administered test B was more proficient than the sample of test A. The columns labeled "Freq" and "Prob" give the frequency distributions of the number-correct scores and the associated cumulative proportions, respectively. Note that 73% of sample A failed their test, while 39% of sample B failed theirs. Again, this is as expected.

The next question of interest is the reliability of the equating procedure. This can be translated into the question how precise the two cutoff scores can be distinguished. If we denote the cutoff scores by S_A and S_B, and denote the estimates of the positions on the latent scale associated with these two cutoff points by $\hat{\theta}_{SA}$ and $\hat{\theta}_{SB}$, then $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$ can be used as a measure of the precision with which we can distinguish the two scores. The estimates $\hat{\theta}_{SA}$ and $\hat{\theta}_{SB}$ are not independent. Firstly, they both depend on the same linked data set and, secondly, they both depend on a concurrent estimate of all item-parameters, a_i, b_i and c_i, and (functions of) all latent proficiency parameters θ_n. Therefore, the standard error $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$ cannot be merely computed as the square root of $Var(\hat{\theta}_{SA} - \hat{\theta}_{SB}) = Var(\hat{\theta}_{SA}) + Var(\hat{\theta}_{SB})$, but the covariance of the estimates must be taken into account also. The method to achieve this is outlined below, after the outline of a statistical framework and considering the problem of test scoring with number-correct scores when these are not sufficient statistics.

11.2 Preliminaries

Nowadays, marginal maximum likelihood (MML, see, Bock and Aitkin 1981) and fully Bayesian estimation (Albert 1992; Johnson and Albert 1999) are the prominent frameworks for estimating IRT models. Mislevy (1986, also see, Glas 1999) point out that they are closely related, because MML estimation is easily generalized to Bayes modal estimation, an estimation method that seeks the mode of the posterior distribution rather than the mode of the likelihood function. In this chapter, we adopt the MML and Bayes modal framework. In this framework, it is assumed that the θ-parameters are drawn from a common distribution, say, a population proficiency distribution as defined in Formula (11.4). Estimates of the item parameters and the parameters of the population proficiency distribution are obtained by maximizing a likelihood function that is marginalized with respect to the θ-parameters.

An important tool for deriving the estimation equations is Fisher's identity (Efron 1977; Louis 1982). For this identity, we distinguish N independent observations y_n and unobserved data z_n. The identity states that the first order derivatives of the parameters of interest δ with respect to the log-likelihood function $L(.)$ are given by

$$\frac{\partial L(\delta)}{\partial \delta} = \sum_{n=1}^{N} E_{z|y}(\nabla_n(\delta)|\, y_n) = \sum_{n=1}^{N} \int, \ldots, \int \left[\frac{\log p(y_n, z_n; \delta)}{\partial \delta}\right] p(z_n|y_n; \delta) dz_n, \tag{11.5}$$

where $p(y_n, z_n; \delta)$ is the likelihood if z_n would be observed, $\nabla_n(\delta)$ is the first-order derivative of its logarithm, and $p(z_n|y_n; \delta)$ is the posterior distribution of the unobserved data given the observations.

Bock and Aitkin (1981) consider the θ-parameters as unobserved data and use the EM-algorithm (Dempster et al. 1977) for maximum likelihood estimation from incomplete data to obtain estimates of the item and population parameters. In this framework, Glas (1999, 2016) uses Fisher's identity to derive estimation and testing procedures for a broad class of IRT models.

Standard errors can be obtained as the square roots of the covariance matrix of the estimates $Cov(\hat{\delta}, \hat{\delta})$ which can be obtained by inverting the observed Fisher information matrix, say, $Cov(\hat{\delta}, \hat{\delta}) = I(\hat{\delta}, \hat{\delta})^{-1}$. Louis (1982) shows that this matrix is given by

$$I(\delta, \delta) = -\frac{\partial^2 L(\delta)}{\partial \delta\, \partial \delta^t} = -\sum_{n=1}^{N} E_{z|y}\big(\nabla_n(\delta, \delta^t)\big|\, y_n\big) - Cov_{z|y}\big(\nabla_n(\delta)\nabla_n(\delta)^t\big|\, y_n\big), \tag{11.6}$$

where $\nabla_n(\delta, \delta^t)$ stands for the second-order derivatives of $\log p(y_n, z_n; \delta)$ with respect to δ. Evaluated at the MML estimates, the information matrix can be approximated by

$$I(\hat{\delta}, \hat{\delta}) \approx \sum_{n=1}^{N} E_{z|y}\big(\nabla_n(\delta)\nabla_n(\delta)^t\big|\, y_n\big) \tag{11.7}$$

(see Mislevy 1986). In the next sections, this framework will be applied to the issues addressed in this chapter: the reliability of tests scored with number-correct scores and to equating errors.

11.3 MAP Proficiency Estimates Based on Number-Correct Scores

Glas (1999, 2016) shows how the estimation equations for the item and population parameters of a broad class of IRT models can be derived using Fisher's identity. This identity can also be applied to derive an estimation equation for a proficiency estimate based on a number-correct score

$$s = \sum_{i=1}^{k} d_i y_i, \tag{11.8}$$

with d_i and y_i as defined in (11.1) and (11.2) dropping the subscript n. The application of Fisher's identity is based on viewing a response pattern as unobserved and the number-correct score as observed. Define $L_s(\theta)$ as the product of the normal prior distribution $g(\theta; \lambda)$ with $\lambda = (\mu, \sigma^2)$ and the probability of a number-correct score s given θ. Define $\{y|s\}$ as the set of all response patterns resulting in a number correct score s. Then the probability of a number-correct score s given θ is equal to the sum over $\{y|s\}$ of the probabilities of response patters $P(y|\theta, \beta)$ given item parameters β and proficiency parameters θ. Application of Fisher's identity results in a first order derivative

$$\frac{\partial L_s(\theta)}{\partial \theta} = E_{y|s}(\nabla(\theta)|\, s, \beta) = \frac{\sum_{\{y|s\}} \left[\frac{\partial \log P(y,\theta;\beta,\lambda)}{\partial \theta}\right] P(y|\theta, \beta)}{\sum_{\{y|s\}} P(y|\theta, \beta)}. \tag{11.9}$$

Equating this expression to zero gives the expression for the MAP estimate. Computation of the summation over $\{y|s\}$ can be done using the recursive algorithm by Lord and Wingersky (1984). The algorithm is also used by Orlando and Thissen (2000) for the computation of expected a-posteriori estimates of θ given a number-correct score s.

Note that in expression (11.9), the prior $g(\theta; \lambda)$ cancels in the posterior, so $p(y|s; \theta, \beta, \lambda) \equiv p(y|s; \theta, \beta)$.

As an example, consider the 2PLM, given by expression (11.3) with $c_i = 0$. The probability of a response pattern becomes

$$\begin{aligned} L_s(\theta) = \sum_{\{y|s\}} \log P(y, \theta; \beta, \lambda) &= \log g(\theta; \mu, \sigma^2) \\ &+ \sum_{\{y|s\}} \sum_{i=1}^{K} \log\big(P_i(\theta)^{d_i y_i}(1 - P_i(\theta))^{d_i(1-y_i)}\big), \end{aligned} \tag{11.10}$$

and

$$\frac{\partial L_s(\theta)}{\partial \theta} = \frac{\mu - \theta}{\sigma^2} + \sum_{\{y|s\}} \sum_{i=1}^{K} (d_i a_i (y_i - P_i(\theta))) p(y|s; \theta, \beta)\,. \tag{11.11}$$

The estimation equation can be solved by either the Newton-Raphson algorithm, or by the EM algorithm. Standard errors can be based on observed information as defined in expression (11.7). One way of estimating θ and computing the standard errors is to impute the item parameters as known constants. However, when we want to compare the estimated proficiencies obtained for two tests through their difference, say, $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$, we explicitly need to take the precision of the estimates of all item and population parameters into account. How this is accomplished is outlined in the next section.

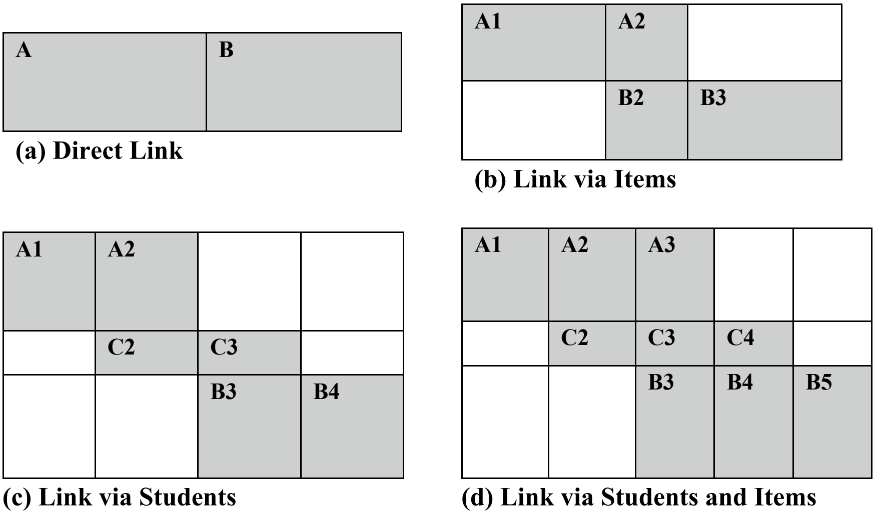
11.4 Equating Error

Suppose θ_0 is a cutoff point on the latent scale and we want to impose this cutoff point on several test versions. Further, we want to estimate the reliability of the created link. Three procedures for the computation of equating errors will be discussed, using some possible data collection designs displayed in Fig. 11.1.

To introduce the first method, consider the design displayed in Fig. 11.1a. In this design, students were administered both test versions, that is, Version A and Version B. The first measure for the strength of the link is based on the standard error of the difference between the average difficulties of the two versions, say, $Se(\overline{b}_A - \overline{b}_B)$, where $\overline{b}_A$ is the estimate of the mean difficulty of Version A and $\overline{b}_B$ the estimate of the mean difficulty of Version B. The strength of the link is mainly determined by the number of students, but also by the number of item parameters making up the two means. Since the estimates are on a latent scale that is subject to linear transformations, we standardize the standard error with the standard deviation of the proficiency distribution. This leads to the definition of the index

$$\text{Equating Error} = \frac{Se(\overline{b}_A - \overline{b}_B)}{Sd(\theta)}. \qquad (11.12)$$

The standard error can be computed as the square root of $Var(\overline{b}_A - \overline{b}_B)$, which can be computed by pre- and post-multiplying the covariance matrix by a vector of weights, that is, $\boldsymbol{w}^t Cov(\hat{\delta}, \hat{\delta})\boldsymbol{w}$,



(a) Direct Link

(b) Link via Items

(c) Link via Students

(d) Link via Students and Items

Fig. 11.1 Four designs for test equating

$$\text{where } \boldsymbol{w} \text{ has elements} \quad \mathrm{w}_j = \begin{cases} \frac{d_{iA}}{\Sigma_i d_{iA}} - \frac{d_{iB}}{\Sigma_i d_{iB}} & \text{if } j \text{ is related to } Cov(\hat{b}_i, \hat{b}_i) \\ 0 & \text{if this is not the case,} \end{cases} \tag{11.13}$$

where d_{iA} and d_{iB} are defined by expression (11.1), for a student administered test A and a student administered test B, respectively.

Figure 11.1b gives an example of equating two tests via common items (the so-called anchor). The test consisting of the items A1 and A2 is linked to the test consisting of the items B2 and B3, because A2 and B2 consist of the same items. The larger the anchor, the stronger the link. In this design it is usually assumed that the means of the two proficiency distributions are different. This leads to a second definition of an index for equating error, that is:

$$\text{Equating Error} = \frac{Se(\hat{\mu}_A - \hat{\mu}_B)}{Sd(\theta)}, \tag{11.14}$$

where $Sd(\theta)$ is a pooled estimate of the standard deviations of the proficiency distributions of the two populations. In Fig. 11.1c, the test consisting of parts A1 and A2 and the test consisting of the parts B3 and B4 have no items in common, but a link is forged by the students administered C2 and C3.

Again, the standard error can be computed as the square root of the associated variance, which can be computed by pre- and post-multiplying the covariance matrix of the parameter estimates by a vector of weights, that is, $\boldsymbol{w}^t Cov(\hat{\delta}, \hat{\delta})\boldsymbol{w}$, where $\boldsymbol{w}$ has elements

$$w_j = \begin{cases} 1 & \text{if } j \text{ is related to } Cov(\hat{\mu}_A, \hat{\mu}_A) \\ -1 & \text{if } j \text{ is related to } Cov(\hat{\mu}_B, \hat{\mu}_B) \\ 0 & \text{if this is not the case.} \end{cases} \tag{11.15}$$

A third method to assess a equating error is based on the position of the cutoff point on the latent scale. This approach gives a more precise estimate of the equating error of the cutoff point, but below it becomes clear that it is somewhat more complicated to compute. Suppose θ_0 is the cutoff point on the latent scale. On both tests, we choose an observed cutoff score, say S_A and S_B, that are associated with the same (mean) proficiency level θ_0. Then an equating error index can be defined as

$$\text{Equating Error} = \frac{Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})}{Sd(\theta)} \tag{11.16}$$

where $\hat{\theta}_{SA}$ and $\hat{\theta}_{SB}$ are the estimates of the positions on the latent scale with the two observed cutoffs.

To define this standard error, we augment the log-likelihood given the observed data with two observations, one for each of the sum scores S_A and S_B. So the complete likelihood becomes $L(\delta, \theta) = L(\delta) + L_s(\theta)$, and the information matrix becomes

$$I(\delta,\theta) \approx E_\theta \left(\begin{matrix} \nabla(\delta)\nabla(\delta)^t & \nabla(\delta)d(\theta_{SA})^t & \nabla(\delta)d(\theta_{SB})^t \\ \nabla(\theta_{SA})\nabla(\delta)^t & \nabla(\theta_{SA})\nabla(\theta_{SA})^t & 0 \\ \nabla(\theta_{SB})\nabla(\delta)^t & 0 & \nabla(\theta_{SB})\nabla(\theta_{SB})^t \end{matrix} \middle| \, y \right). \tag{11.17}$$

As above, the standard error of the difference between $\hat{\theta}_{SA}$ and $\hat{\theta}_{SB}$ can be computed as the square root of the associated variance, which can be computed by pre- and post-multiplying the covariance matrix by a vector of weights, that is, $\boldsymbol{w}^t Cov(\hat{\delta}, \hat{\delta})\boldsymbol{w}$. In this case, the vector $\boldsymbol{w}$ has elements

$$w_j = \begin{cases} 1 & \text{if } j \text{ is related to } Cov(\hat{\theta}_{SA}, \hat{\theta}_{SA}) \\ -1 & \text{if } j \text{ is related to } Cov(\hat{\theta}_{SB}, \hat{\theta}_{SB}) \\ 0 & \text{if this is not the case.} \end{cases} \tag{11.18}$$

Examples will be given below.

EAP estimates and another approach to the reliability of number-correct scores.

In test theory we distinguish between global reliability and local reliability. Global reliability is related to the precision with which we can distinguish two randomly drawn students from some well-defined population, while local reliability relates to the precision given a specific test score. We discuss these two concepts in the framework of IRT in turn.

One of the ways in which global reliability can be defined is as the ratio of the true variance relative to the total variance. For the framework of IRT, consider the variance decomposition

$$var(\theta) = var[E(\theta|\boldsymbol{y})] + E[var(\theta|\boldsymbol{y})], \tag{11.19}$$

where $\boldsymbol{y}$ is an observed response pattern, $var(\theta)$ is the population variance of the latent variable, $var[E(\theta|\boldsymbol{y})]$ is the posterior variance of the expected person parameters (say, the EAP estimates of θ). So this EAP estimate is the error variance averaged over the values that can be observed weighted with their probability of their occurrence under the model. Further, $E[var(\theta|\boldsymbol{y})]$ is the expected posterior variance of the EAP estimate. Then reliability is given by the ratio

$$\rho = \frac{var[E(\theta|\boldsymbol{y})]}{var(\theta)} = 1 - \frac{E[var(\theta|\boldsymbol{y})]}{var(\theta)} \tag{11.20}$$

(See, Bechger et al. 2003). The middle expression in (11.20) is the variance of the estimates of the person parameters relative to the 'true' variance, and the right-hand expression in (11.15) is one minus the average variance of the estimates of the student parameters, say, the error variance, relative to the 'true' variance.

The generalization to number-correct scores s is straightforward. If the observations are restricted from $\boldsymbol{y}$ to s, a student's proficiency can be estimated by the EAP $E(\theta|s)$, that is, the posterior expectation of θ given s, and the precision of the estimate is given by the posterior variance $var(\theta|s)$. Then global reliability generalizes to

$$\rho_s = \frac{var[E(\theta|s)]}{var(\theta)} = \frac{\text{var}(\theta) - E[var(\theta|s)]}{var(\theta)}. \tag{11.21}$$

If the 1PLM holds, s is a sufficient statistic for θ. Therefore, it is easily verified that $E(\theta|s) \equiv E(\theta|\boldsymbol{y})$ and the expressions (11.20) and (11.21) are equivalent. In all other cases, computation of the posterior distribution involves a summation over all possible response patterns resulting in a number-correct score s, and, as already noticed above, this can be done using the recursive algorithm by Lord and Wingersky (1984).

If the 1PLM does not hold, there is variance in $E(\theta|\boldsymbol{y})$ conditional on s. This leads to the interesting question how much extra error variance is created by using s as the basis for estimating θ. That is, we are interested in the contribution of $Var(E(\theta|\boldsymbol{y})|s)$ to the total error variance, that is, to the posterior variance $Var(\theta|s)$. This contribution can be worked out by using an identity analogous to Expression (11.21), that is,

$$Var(\theta|s) \ = E(Var(\theta|\boldsymbol{y})|s) + Var(E(\theta|\boldsymbol{y})|s)). \tag{11.22}$$

Note that $E(Var(\theta|\boldsymbol{y})|s)$ is the squared measurement error given $\boldsymbol{y}$ averaged over the distribution of $\boldsymbol{y}$ given s, and $Var(E(\theta|\boldsymbol{y})|s))$ is the variance of the EAP estimates, also over the distribution of $\boldsymbol{y}$ given s. In the next section, examples of local reliability estimates will be given.

Examples of Reliability Estimates

In this section, two simulated examples are presented to show the kind of results that the local reliability indices presented above produce.

The first example is created by simulating 1000 response patterns on a 20-item test. The data were created with the 2PLM, with the θ-values drawn from a standard normal distribution. The 20 item parameters were the product of a set of four discrimination parameters $a = \{0.8, 0.9, 1.10, 1.20\}$ and five difficulty parameters $b = \{-1.0, -0.5, 0.0, 0.5, 1.0\}$. MML estimates (i.e., Bayes modal estimates) were computed with a standard normal distribution for the θ-values. The results are displayed in Table 11.2.

Note that the MAP estimates and the EAP estimates are very similar, as are their standard deviations displayed in the columns labeled $Sd_{MAP}(\theta|s)$ and $Sd_{EAP}(\theta|s)$. The last three columns give the variance decomposition as defined in Expression (11.22). It can be seen that $Var(E(\theta|\boldsymbol{y})|s)$ is relatively small compared to $E(Var(\theta|\boldsymbol{y})|s))$. So the potential bias in a student's proficiency estimate when using number-correct scores is much less than the inflation of the precision of the estimate. A final observation that can be made from this simulation study is that the global reliability when switching from scoring using the complete response patters to using the number-correct scores dropped from 0.788 to 0.786. So the loss in global reliability was negligible.

It is expected that if the variability of the discrimination parameters is enlarged, $Var(E(\theta|\boldsymbol{y})|s)$ increases. The reason is that if the discrimination parameters are considered known, the weighted sum score $\Sigma_i d_i a_i y_i$ is a sufficient statistic for θ. If

Table 11.2 MAP and EAP estimates and their local reliability

Score	Freq	MAP (θ)	$Sd_{MAP}(\theta\|s)$	EAP(θ)	$Sd_{EAP}(\theta\|s)$	$Var(\theta\|s)$	$Var(E(\theta\|\mathbf{y})\|s)$	$E(Var(\theta\|\mathbf{y})\|s)$
0	1	−2.24	0.59	−2.26	0.60	0.36	0.00	0.36
1	15	−1.92	0.55	−1.93	0.56	0.32	0.00	0.31
2	17	−1.64	0.52	−1.67	0.53	0.29	0.00	0.28
3	32	−1.38	0.50	−1.37	0.51	0.26	0.01	0.25
4	46	−1.14	0.48	−1.17	0.49	0.24	0.00	0.23
5	51	−0.92	0.46	−0.93	0.47	0.22	0.00	0.22
6	64	−0.71	0.45	−0.72	0.46	0.21	0.00	0.21
7	59	−0.51	0.44	−0.53	0.45	0.20	0.01	0.20
8	79	−0.32	0.44	−0.34	0.44	0.19	0.01	0.19
9	85	−0.13	0.43	−0.14	0.44	0.19	0.00	0.19
10	89	0.06	0.44	0.07	0.44	0.20	0.01	0.19
11	73	0.25	0.44	0.25	0.44	0.19	0.00	0.19
12	88	0.44	0.44	0.44	0.44	0.19	0.01	0.19
13	61	0.63	0.44	0.65	0.45	0.20	0.01	0.20
14	73	0.83	0.45	0.84	0.46	0.21	0.00	0.21
15	64	1.04	0.46	1.06	0.47	0.22	0.01	0.22
16	37	1.26	0.48	1.28	0.49	0.24	0.00	0.23
17	30	1.50	0.50	1.50	0.51	0.26	0.01	0.25
18	19	1.75	0.52	1.78	0.53	0.28	0.00	0.28
19	12	2.04	0.55	2.08	0.57	0.32	0.00	0.32
20	5	2.36	0.59	2.38	0.60	0.36	0.00	0.36

all discrimination parameters are equal to 1.0, the 2PLM becomes the 1PLM, and then the number-correct score becomes a sufficient statistic. So the more variance in the discrimination parameters, the greater the violation of the 1PLM and the depreciation of the appropriateness of the scoring rule.

To investigate this effect, the discrimination parameters of the simulation were changed to parameters $a = \{0.40, 0.60, 1.40, 1.60\}$. The results are displayed in Table 11.3. It can be seen that the standard deviations in the columns labeled $Sd_{MAP}(\theta|s)$ and $Sd_{EAP}(\theta|s)$ blew up a bit, but the effect was not very large. Further, in the column labeled $Var(E(\theta|\mathbf{y})|s)$ the values clearly increased, while this is less the case in the column labeled $E(Var(\theta|\mathbf{y})|s))$. For instance, if we consider a number-correct score 10, we observe that the initial values 0.01 and 0.19 changed to 0.04 and 0.17. The net effect was a change in $Var(\theta|s)$ from 0.20 to 0.21. So the increase in variance of θ-estimates (that is, of expectations $E(\theta|\mathbf{y})$) was counterbalanced by an increase of the overall precision $Var(\theta|\mathbf{y})$.

11.5 Simulation Study of Equating Errors

In this section, two sets of simulation studies will be presented. The first study was based on the design displayed in Panel b of Fig. 11.1, which displays a design with a link via common items. The simulation was carried out to study the effect of the size of the anchor. The second set of simulations was based on the design of Panel c of Fig. 11.1, which displays a design with common students. These simulations were carried out to study the effect of the number of students in the anchor.

The studies were carried out using the 2PLM. To create realistic data, the item parameters were sampled from the pool of item parameters used in the final tests in primary education in the Netherlands. Also the means of proficiency distributions and cutoff scores were chosen to create a realistic representation of the targeted application, that entailed equating several versions and cycles of the tests.

For the first set of simulations, two tests were simulated with 2000 students each. The proficiency parameters for the first sample of students were drawn from a standard normal distribution, while the proficiency parameters for the second sample of students were drawn from a normal distribution that was either standard normal or normal with a mean 0.5 and a variance equal to 1.0. Cutoff points were varied as $\theta_0 = -0.5$ or $\theta_0 = 0.0$. The results are displayed in Table 11.4. The first column gives the length of the two tests; the tests were of equal size. 50 items is considered realistic for a high-stakes test, tests of 20 and 10 items were simulated to investigate the effects of decreasing the test length.

The second column gives the size of the anchor. The total number of items in the design displayed in the third column follows from the length of the two tests and the size of the anchor. 100 replications were made for every one of the 24 conditions. For every replication, the item parameters were redrawn from the complete pool of all item parameters of all (five) test providers. The complete pool consisted of approximately 2000 items. The last three columns give the three equating errors

Table 11.3 MAP and EAP estimates and their local reliability when the variance of the discrimination parameter is increased

Score	Freq	MAP(θ)	$Sd_{MAP}(\theta\|s)$	EAP(θ)	$Sd_{EAP}(\theta\|s)$	$Var(\theta\|s)$	$E(Var(\theta\|\mathbf{y})\|s)$	$E(Var(\theta\|\mathbf{y})\|s)$
0	2	−2.26	0.57	−2.21	0.64	0.41	0.00	0.41
1	5	−1.96	0.54	−1.99	0.61	0.37	0.00	0.37
2	23	−1.69	0.51	−1.73	0.60	0.36	0.02	0.34
3	35	−1.44	0.49	−1.48	0.57	0.33	0.04	0.29
4	47	−1.21	0.47	−1.22	0.54	0.29	0.05	0.24
5	58	−1.00	0.46	−1.00	0.53	0.28	0.06	0.23
6	67	−0.80	0.45	−0.83	0.48	0.23	0.03	0.20
7	71	−0.60	0.44	−0.61	0.47	0.22	0.04	0.19
8	63	−0.41	0.43	−0.42	0.45	0.20	0.03	0.17
9	86	−0.23	0.43	−0.21	0.45	0.21	0.04	0.17
10	99	−0.04	0.43	−0.05	0.46	0.21	0.04	0.17
11	81	0.14	0.43	0.15	0.45	0.20	0.03	0.17
12	87	0.33	0.43	0.36	0.46	0.21	0.04	0.17
13	63	0.52	0.44	0.51	0.47	0.22	0.04	0.18
14	60	0.71	0.45	0.77	0.49	0.24	0.04	0.20
15	48	0.91	0.45	0.98	0.51	0.26	0.04	0.22
16	44	1.13	0.47	1.19	0.54	0.29	0.04	0.25
17	33	1.35	0.49	1.43	0.57	0.33	0.05	0.28
18	19	1.60	0.51	1.64	0.60	0.36	0.05	0.31
19	8	1.87	0.53	1.89	0.62	0.39	0.02	0.36
20	1	2.17	0.57	2.13	0.64	0.40	0.00	0.40

Table 11.4 Simulation of equating via common items

Number of items							
Examination	Anchor	Total	θ_0	μ_B	$Se(\overline{b}_A - \overline{b}_B)$	$Se(\hat{\mu}_A - \hat{\mu}_B)$	$Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$
50	30	70	0.00	0.00	0.050	0.010	0.441
				0.50	0.053	0.010	0.441
			−0.50	0.00	0.050	0.010	0.445
				0.50	0.052	0.010	0.446
	20	80	0.00	0.00	0.054	0.014	0.441
				0.50	0.055	0.015	0.441
			−0.50	0.00	0.055	0.015	0.447
				0.50	0.056	0.016	0.447
	10	90	0.00	0.00	0.062	0.022	0.434
				0.50	0.059	0.023	0.434
			−0.50	0.00	0.059	0.022	0.441
				0.50	0.061	0.023	0.441
20	10	30	0.00	0.00	0.053	0.018	0.651
				0.50	0.054	0.020	0.651
			−0.50	0.00	0.052	0.018	0.677
				0.00	0.054	0.018	0.651
	5	35	0.00	0.00	0.082	0.028	0.666
				0.50	0.080	0.031	0.666
			−0.50	0.00	0.086	0.029	0.682
				0.50	0.079	0.031	0.683
10	5	15	0.00	0.00	0.100	0.024	0.889
				0.50	0.086	0.026	0.889
			−0.50	0.00	0.097	0.024	0.937
				0.50	0.087	0.026	0.937

defined above. Note that $Sd(\theta)$ was always equal to 1.0, so the equating errors were equal to the analogous standard errors.

The results are generally as expected. Note first that there was always a substantial main effect of the test length for all three indices. For a test length of 50 items, decreasing the size of the anchor increased the equating errors for the average item difficulties $Se(\overline{b}_A - \overline{b}_B)$ and the proficiency means $Se(\hat{\mu}_A - \hat{\mu}_B)$. The effect on $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$ was small. This pattern was sustained for a test length of 20 items, but in that case also $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$ increased slightly when the anchor was decreased from 10 to 5. Finally, there were no marked effects of varying the position of the cutoff points and the differences between the two proficiency distributions.

The second set of simulations was based on the design of panel c of Fig. 11.1, the design with common students. The general setup of the study was analogous to the first one, with some exceptions. All samples of students were drawn from standard normal distributions and the cutoff point was always equal to $\theta_0 = 0.0$. There were three tests in the design: two tests to be equated and a test given to the linking group. As can be seen in the first column of Table 11.5, the tests to be equated had either 40 or 20 items. In the second column, it can be seen that the linking groups were either administered tests of 20, 10, or 4 items. These linking tests always comprised of an equal number of items from the two tests to be equated. The third column shows how the size of the sample of the linking group was varied. The two tests to be equated were always administered to 2000 students. In general, the results are much worse than those displayed in Table 11.4. In fact, only the combination of two tests of 40 items with a linking group of 1600 students administered a test of 20 items comes close to the results displayed in Table 11.4. Note that linking tests of 40 items with linking groups administered 4 items completely breaks down, especially the results for $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$ with 100, 400 or 800 students in the linking groups become extremely poor.

11.6 Conclusion

Transparency of scoring is one of the major requirements for the acceptance of an assessment by stakeholders such as students, teachers and parents. This is probably the reason why number-correct scores are still prominent in education. The logic of such scoring is evident: the higher the number of correct responses, the higher the student's proficiency. The alternative of using the proficiency estimates emanating from an IRT model as test scores is more complicated to explain. In some settings, such as in the setting of computerized adaptive testing, it can be made acceptable that students that respond to more difficult items get a higher proficiency estimate than students with an analogous score on more easy items. However, explaining the dependence of proficiency estimates on item-discrimination parameters is more cumbersome.

A potential solution to the problem is using the 1PLM model, where all items are assumed to have the same discrimination index, and the proficiency estimate only depends on the number of correct responses to the items. However, the 1PLM seldom fits educational test data and using the 1PLM to utilize all the advantages of IRT leads to notable loss of precision. Therefore, the 2PLM and 3PLM have become the standard models for analyzing educational test data. In this chapter, a method to combine number-correct scoring with the 2PLM and 3PLM was suggested and methods for relating standards on the number-correct scale to standards on the latent IRT scale were outlined. Indices for both the global and local reliability of number-correct scores were introduced. It was shown that the error variance for number-correct scoring can be decomposed into two components. The first component is the variance of the proficiency estimates given the response patterns conditional on

Table 11.5 Simulation of equating via common students

Number of items		Number of students			
Examination	Linking group	Linking group	$Se(\overline{b}_A - \overline{b}_B)$	$Se(\hat{\mu}_A - \hat{\mu}_B)$	$Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$
40	20	100	0.092	0.098	0.505
		400	0.088	0.045	0.495
		800	0.081	0.036	0.494
		1600	0.069	0.032	0.494
40	10	100	0.468	0.179	0.574
		400	0.201	0.062	0.499
		800	0.139	0.051	0.496
		1600	0.101	0.045	0.495
40	4	100	0.416	2.043	3.310
		400	0.209	0.487	3.290
		800	0.129	0.342	2.240
		1600	0.096	0.222	0.587
20	10	100	0.128	0.118	0.702
		400	0.116	0.060	0.692
		800	0.107	0.050	0.692
		1600	0.089	0.044	0.691
20	4	100	0.618	0.167	0.724
		400	0.289	0.107	0.705
		800	0.204	0.098	0.703
		1600	0.130	0.092	0.702

number-correct scores. This component can be viewed as a measure for the bias introduced by using number-correct scores as estimates for proficiency rather than estimating the proficiency under the 2PLM or 3PLM based on a student's complete response pattern. The second component can be interpreted as the average error variance when using the number-correct score. The presented simulation studies indicate that, relative to the second component, the first component is small.

When equating two tests, say an older version and a newer version, it is not only the standard error of the proficiency estimates on the two tests which is important, but also the standard error of differences between proficiency estimates on the two tests. To obtain a realistic estimate of the standard errors of these differences, the whole covariance matrix of the estimates of all item and population parameters in the model must be taken into account. The size of these standard errors depends on the strength of the link between the two tests, that is, on the number of items and students in the design and the sizes of the overlap between, respectively, items and students. The simulation studies presented in this chapter give an indication of the standard errors of these differences for various possible designs.

The procedure for number-correct scoring was presented in the framework of unidimensional IRT models for dichotomously scored items. It can be generalized in various directions. First of all, a sum score can also be defined for a test with polytomously scored items by adding the scores on the individual items in the test. These sum scores can then be related to a unidimensional IRT model for polytomously scored items such as the generalized partial credit model (Muraki 1992), the graded response model (Samejima 1969) or the sequential model (Tutz 1990) in an manner that is analogous to the procedure presented above. Also multidimensional versions of these models (Reckase 1985) present no fundamental problems: the proficiency distributions and response probabilities introduced above just become multivariate distributions in multivariate θ parameters. For the generalized definitions of reliabilities refer to van Lier et al. (2018).

A final remark concerns the statistical framework of this chapter, which was the related Bayes modal and marginal maximum likelihood framework. In the preliminaries section of this chapter, it was already mentioned that this framework has an alternative in the framework of fully Bayesian estimation supported by Markov chain Monte Carlo computational methods (Albert 1992; Johnson and Albert 1999). Besides with dedicated samplers, the IRT models discussed here can also be estimated using general purpose samplers such as Bugs (Lunn et al. 2009) and JAGS (Plummer 2003). But details of the generalizations to other models and another computational framework remain points for further study.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics, 17,* 251–269.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27,* 319–334.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46,* 443–459.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society: Series B, 39,* 1–38.

Efron, B. (1977). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1–26.

Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika, 64,* 273–294.

Glas, C. A. W. (2016). Maximum-likelihood estimation. In W.J. van der Linden (ed.). *Handbook of Item Response Theory: Vol. 2. Statistical tools* (pp. 197–216). Boca Raton, FL: Chapman and Hall/CRC.

Johnson, V., & Albert, J. H. (1999). *Ordinal data modeling*. New York, NJ: Springer.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887–903.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8,* 453–461.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B, 44,* 226–233.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28,* 3049–3067.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Neyman, J., & Scott, E. L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica, 16,* 1–32.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria. ISSN 1609-395X.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmark Paedagogiske Institute.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph, 17,* 1–100.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39–55.

van Lier, H. G., Siemons, L., van der Laar, M. A. F. J., & Glas, C. A. W. (2018). Estimating optimal weights for compound scores: A multidimensional IRT approach. *Multivariate Behavioral Research.* Published Online: https://www.tandfonline.com/doi/full/10.1080/00273171.2018.1478712.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 12
Differential Item Functioning in PISA Due to Mode Effects

Remco Feskens, Jean-Paul Fox and Robert Zwitser

Abstract One of the most important goals of the Programme for International Student Assessment (PISA) is assessing national changes in educational performance over time. These so-called trend results inform policy makers about the development of ability of 15-year-old students within a specific country. The validity of those trend results prescribes invariant test conditions. In the 2015 PISA survey, several alterations to the test administration were implemented, including a switch from paper-based assessments to computer-based assessments for most countries (OECD 2016a). This alteration of the assessment mode is examined by evaluating if the items used to assess trends are subject to differential item functioning across PISA surveys (2012 vs. 2015). Furthermore, the impact on the trend results due to the change in assessment mode of the Netherlands is assessed. The results show that the decrease reported for mathematics in the Netherlands is smaller when results are based upon a separate national calibration.

R. Feskens (✉)
Cito, Arnhem, The Netherlands
e-mail: remco.feskens@cito.nl

J.-P. Fox
University of Twente, Enschede, The Netherlands

R. Zwitser
University of Amsterdam, Amsterdam, The Netherlands

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_12

12.1 Introduction

The Programme for International Student Assessment (PISA) is an international comparative research project investigating the knowledge and skills of 15-year old students in science, reading literacy, and mathematics (OECD 2016a). Since 2000, PISA has been conducted every three years under the auspices of the Organization for Economic Cooperation and Development (OECD) with each administration featuring a different core component. The main aim of PISA is to provide participating countries with comparable information which can be used to evaluate and improve their educational policies. The evaluation of existing educational policy is mainly based upon the assessment of the functional literacy skills of students as measured by standardized cognitive tests. Country-specific performance can be compared to other countries within one administration, but country performances can also be compared between administrations. Comparisons based upon the latter inform countries if performance in one of the three subjects has changed.

The trend results for 2015 showed that many countries decreased in their overall performance when compared to the previous administration in 2012. This holds in particular for the core domain, science, but many countries also experienced a decrease in their mean mathematics performance. For example in the Netherlands, the mean country estimate for mathematics in 2015 decreased when compared to 2012, dropping 13 points (Feskens et al. 2016). This decrease is substantial, given that changes in country means measured over a short period of time are usually small (Mazzeo and von Davier 2008; cf. Robitzsch et al. 2017). This could imply that many of the 2015 country cohorts are less proficient when compared to their 2012 counterparts and that consequently, the overall educational level has dropped in these countries.

However, several changes were implemented in the 2015 PISA administration. The main changes in the 2015 methodology involved: (1) a change from traditional paper-based assessment (PBA) to a computer-based assessment (CBA) in most of the participating countries; (2) a different assessment of item nonresponse; (3) the use of a two-parameter logistic model (Birnbaum 1968; Muraki 1992) instead of a one-parameter logistic model (Rasch 1960; Masters 1982) for item response modelling; and (4) a change in the treatment of differential item functioning (DIF) across countries (OECD 2016b). The trend result decreases could also be due to these changes rather than just reflecting a decrease in academic performance.

Changing the mode of assessment administration can affect student performance and induce DIF (Kolen and Brennan 2014). Earlier studies have already discussed the validity of the trend results from PISA 2015. Robitzsch et al. (2017) concluded that an observed decrease in mathematics and science might also be due to unfamiliarity with computer use in classroom settings among German students. Robitzsch and colleagues (2017) came to this conclusion by re-analysing the PISA data and estimating the mean German performances comparing several calibration designs. Another recent study conducted by Jerrim et al. (2018) also questions the comparability of PBA and CBA results. Both studies are among others based upon a reanalysis of the

2014 PISA field trial study data available for some countries. This field trial served to ensure the comparability of test results across paper and computer modes (OECD 2016b). The OECD (2016b) concluded "there are very few samples where the main effect of the mode of delivery reaches statistical significance" (p. 15 Annex A6), but also recognizes that national field trial samples were small and not specifically designed to examine whether items administered in one mode were systematically easier or harder than those administered in another mode at the national level. Nevertheless, the random assignment of students within field trial schools made it possible to disentangle mode and population effects at the international level. This paper evaluates to what degree the new test administration mode of PISA affected items and the performance of the mathematics domain by comparing the last two main PISA cycles with a focus on the Dutch trend results. This will be done by first evaluating for DIF across modes and then by reanalysing the trend results for the Netherlands. We will re-estimate the Dutch results by only making use of the national data collected in the Netherlands. Consequently, only country-specific item parameters are included in this approach and the national trend estimates are not influenced by differences in mode effects between countries.

These considerations have led to following research questions:

1. To what extent are items used to establish trend effects in PISA 2015 affected by DIF?
2. To what extent do trend estimates for the Netherlands differ when results are only based upon the Dutch data, instead of the complete international data file?

This paper will continue with a description of the changes made in the 2015 PISA methodology and the data included for this study, followed by an overview of DIF and the impact DIF can have on trend results. Results will then be presented and followed by conclusions and discussion.

12.2 Changes in PISA 2015

Since its first administration in 2000, the number of countries and students participating in PISA has grown to a total of over 500,000 students from 72 different countries and economies that took part in 2015. All 35 OECD-member countries participated along with so-called partner countries without membership. All PISA research findings are presented using scales that are standardized to an international average of 500 with a standard deviation of 100. The average of 500 only applies to the OECD countries and was determined in the year in which a specific topic (reading literacy in 2000, mathematics in 2003, and science in 2006) dominated the agenda for the first time. The main mode of test administration in PISA was PBA up to 2012. This was changed in 2015, when most of the participating countries switched to CBA. The main reasons for this change are found in the creation of additional possibilities for measuring specific aspects of the domain constructs, the establishment of a more rigorous procedure for test marking and the availability of detailed metadata, such as response time information (OECD 2017).

Despite these major advantages, there is the potential risk of incomparability of results between cycles, since student performances can be affected by aspects specifically related to the differences between CBA and PBA (Bridgeman et al. 2003; Leeson 2006). These differences have to do with the ease of reading texts, the ease of reviewing or changing answers to questions, the speed of test taking, the clarity of figures and diagrams, and the difference between responding on a computer and on an answer sheet (Kolen and Brennan 2014). Previous research on the impact of these two modes on mathematics testing performance has produced mixed results. Hong et al. (2007) did not find a main effect for the testing mode on the performance of K-12 student mathematics tests, whereas Kingston (2009) reported in a meta-analysis of 81 studies that there was a very small advantage for a paper administration of mathematics tests. Sandene et al. (2005) also reported a higher level of performance with PBA when compared to CBA in a study conducted by the National Assessment of Educational Progress Technology-Based Assessment Project. At the item level in that study, five percent more students responded correctly on paper than on a computer. Despite this mixed evidence, it was hypothesized that mathematics items used in PISA 2012 could be transposed onto a screen without affecting trend data in PISA 2015 by keeping the computer skills required to administer the test to a minimum (OECD 2016a).

A second change in the 2015 PISA methodology was the assessment of item nonresponse. First of all, non-reached items were not only treated as not administered in estimating item parameters, but also not taken into consideration in estimating person parameters in 2015. In previous cycles of PISA, only the former was true, whereas non-reached items were seen as incorrect responses while estimating person parameters (OECD 2016a). During the analyses of this study it turned out that a second change with respect to the treatment of item nonresponse was the classification of item nonresponse into different types of item missingness. This change will be discussed in more detail in the results.

The third change implemented in the PISA 2015 design was the item response theory (IRT) model used to calibrate the response data. While a one-parameter logistic model (1PL) was used from 2000 to 2012, a two-parameter logistic (2PL) IRT model was deployed in 2015. The 2PL model is more flexible, allowing different item discrimination parameters, but those parameters also introduce an empirically-based item contribution to ascertain ability. To minimize trend break effects, as many trend items as possible were estimated according to a 1PL model (OECD 2016a).

Finally, DIF was treated differently in PISA 2015. Up to 2012, only a small number of items were considered as "not administered" for some countries as a result of DIF, mainly due to translation or printing errors (OECD 2016a). In 2015, the PISA consortium took two different types of DIF into account: DIF due to mode effects and DIF due to country effects. Mode effects were taken into account based upon the results of the 2014 field trial and further scaling operations in the main study. Items were classified as either scalar, metric, or non-metric invariant items (Annex A5 OECD 2016c). In order to be able to meaningfully compare the latent means and correlations across groups, both the factor loadings and intercepts should be the same across groups (scalar invariance) (Steenkamp and Baumgartner 1998; Meredith 1993), or in IRT terms: the difficulty and discrimination parameters should have the same value. Scalar invariant items were used for establishing trend effects (51 items in total), the metric invariant items (items where the discrimination parameters were equal across modes, but the items had different difficulty parameters) served to improve measurement precision (30 items), and items where no metric invariance could be established were not used in the 2015 CBA test. Furthermore, in 2015 (a limited number of) specific country-by-cycle item parameters were allowed in the case of country DIF (OECD 2016a). Both the 2012 and 2015 procedures aimed to come to a concurrent item parameter estimation using item responses from all participating countries and only introduced a country specific treatment in case of moderate to severe deviations from the overall item parameters. Within this paper, we will focus on the change in assessment mode and the effect mode differences (between countries) have on item parameters and ultimately the mean country estimates.

12.3 Data

To keep student response burden at an acceptable level while at the same time maintaining construct validity, PISA makes use of a multi-booklet design. Students administer different subsets of all questions included in the test, called test versions or booklets. This paper only included students who have administered one of the regular test versions. For Mathematics 2012, these were test versions 1–13 (OECD 2014). For 2015 PISA, the regular CBA test versions 43–54 (OECD 2017) were included.

The mean proficiency estimates of the 53 countries that participated in both PISA 2012 and in the CBA of PISA 2015 are displayed in Table 12.1, where the countries are ordered by the mean 2015 country ability estimates. The number of students within each country that completed one of the regular test versions are presented in the last two columns of the table. Note that these numbers are somewhat smaller compared the total number of students that participated.

As can be seen in Table 12.1, the majority of the 53 countries—especially those with above-average performance—are confronted with a decrease in mean mathematics performance, as expressed by Cohen’s d (Cohen 1988). The effect size for the difference between the 2012 and 2015 mean estimates in the Netherlands is − 0.12. The numbers of students included in 2015 is much smaller compared to 2012, because mathematics was the core domain for 2012. In the years that a subject is the core domain, all students administer items that measure the core domain. Mostly, only one-third of the students within a country take a test of non-core domain subject in any given cycle.

Table 12.1 PISA participant country ability estimates and student participants

Country	Mean 2012	Mean 2015	Effect size	n 2012	n 2015
Singapore	573.47	564.19	−0.10	5522	2019
Hong Kong	561.24	547.93	−0.15	4509	1777
Macao	538.13	543.81	0.07	5320	1473
Chinese Taipei	559.83	542.32	−0.17	6033	2541
Japan	536.41	532.44	−0.05	6303	2185
Korea	553.77	524.11	−0.31	5031	1833
Switzerland	530.93	521.25	−0.11	11,200	2706
Estonia	520.55	519.53	−0.01	4760	1857
Canada	518.07	515.65	−0.03	21,352	6592
Netherlands	522.97	512.25	−0.12	4322	1743
Denmark	500.03	511.09	0.14	7351	2247
Finland	518.75	511.08	−0.10	8731	1930
Slovenia	501.13	509.92	0.10	5731	2072
Belgium	514.53	506.98	−0.08	8241	3108
Germany	513.52	505.97	−0.09	4834	2095
Poland	517.50	504.47	−0.15	4596	2064
Ireland	501.50	503.72	0.03	5002	2647
Norway	489.37	501.73	0.15	4622	1803
Austria	505.54	496.74	−0.10	4702	2283
New Zealand	499.75	495.22	−0.05	4285	1504
Russian Federation	482.17	494.06	0.15	5207	2015
Sweden	478.26	493.92	0.18	4669	1781
Australia	504.15	493.90	−0.11	14,348	4783
France	494.99	492.92	−0.02	4542	2013
United Kingdom	493.93	492.48	−0.02	12,632	4670
Czech Republic	498.96	492.32	−0.07	5224	2249
Portugal	487.06	491.63	0.05	5651	2398
Italy	485.32	489.73	0.05	30,948	3830
Iceland	492.80	488.03	−0.05	3500	1117
Spain	484.32	485.84	0.02	25,189	2225
Luxembourg	489.85	485.77	−0.05	5246	1721
Latvia	490.57	482.31	−0.11	4306	1591
Lithuania	478.82	478.38	−0.01	4616	2157

(continued)

Table 12.1 (continued)

Country	Mean 2012	Mean 2015	Effect size	n 2012	n 2015
Hungary	477.04	476.83	0.00	4774	1860
Slovak Republic	481.64	475.23	−0.07	4607	2066
Israel	466.48	469.67	0.03	4993	2258
United States	481.37	469.63	−0.14	4947	1873
Croatia	471.13	464.04	−0.08	5003	1919
Greece	452.97	453.63	0.01	5115	1820
Bulgaria	438.74	441.19	0.03	2424	1958
United Arab Emirates	434.01	427.48	−0.07	5246	4629
Chile	422.63	422.67	0.00	3115	2315
Turkey	447.98	420.45	−0.34	4839	1924
Uruguay	409.29	417.99	0.11	2448	1986
Montenegro	409.63	417.93	0.11	4712	1875
Thailand	426.74	415.46	−0.15	6602	2719
Mexico	413.28	408.02	−0.08	15,398	2522
Qatar	376.45	402.40	0.28	10,831	5518
Costa Rica	407.00	400.25	−0.11	2028	2036
Colombia	376.49	389.64	0.19	3643	3872
Peru	368.10	386.56	0.24	2767	2309
Brazil	388.51	377.07	−0.15	8796	7614
Tunisia	387.82	366.82	−0.28	1990	1700

51 items were administered in both 2012 and 2015 PISA,[1] 31 of which have been classified as scalar invariant items that serve as anchor items. Ultimately, changes in overall mean country performances across PISA cycles is based upon student achievement in these items and, as such, they will be the subject of a DIF study in the following section.

12.4 Differential Item Functioning

Concerns related to how the test administration mode might affect the performance of students is the domain of measurement invariance. The measurement instrument—the cognitive test—should function in the same way across varied conditions, as long as these conditions (here the administration mode) are irrelevant to the attribute being measured (Millsap 2011). As soon as comparisons between in this

[1] In PISA 2012, the item labels included an additional "P", indicating they were administered on paper. The 2015 CBA item labels start with an additional "C".

case two different cohorts are made, it is expected that the two tests will not produce systematically different results for groups with the same ability level. If consistent differences in the probability of endorsing an item are found for groups with the same ability level, the item is said to exhibit measurement bias with respect to the mode of administration (Millsap 2011). The evaluation of measurement invariance concerns the assessment how an item functions within different groups.

Therefore, item bias is also referred to by a more neutral term, DIF (Mellenbergh 2011). DIF is directly related to the evaluation of item response probabilities for members of different groups after matching on the latent trait that the test is intended to measure. DIF is a term coined for situations containing a reference group, defined here as the group of students who completed the PISA test using paper and pen, and a focal group, defined as the students who completed the test on a computer. DIF exists when the performance on a particular item or class of items differs between the focal and reference group, controlling for the performance on the test as a whole.

Although the concept of DIF seems straightforward, some problems have been highlighted in among others a recent study by Bechger and Maris (2014) and are mostly related to comparing parameters that are not identified from the observations. Bechger and Maris (2014) proposed using a differential item pair functioning DIF test, which focuses on comparing item pairs instead of seeing DIF as an item property. The difference with traditional procedures is that DIF is defined in terms of the relative difficulties of pairs of items—which are identified from the observations—and not in terms of the difficulties of individual items.

The procedure starts with a separate calibration of the data within each group. There exists an overall test for DIF, which under the null hypothesis that there is no DIF follows a Chi-square distribution with the number of items minus one degrees of freedom (Bechger and Maris 2014). If an item pair in the calibration of one group has a different relative difficulty when compared to the relative difficulty in the calibration of the second group, that item pair is subject to DIF. Differences between item pair difficulties can be tested using a Wald test and the results are usually summarized in a heat map. The plot highlights item pairs that have large inter-group differences in the relative positions of their relative difficulties (Bechger and Maris 2014).

Trends within countries can be estimated by making use of data collected in every country. This is done by making use of concurrent item parameter estimation aimed to place all items on a common international scale. Until PISA 2012 this was established in two stages: First, based upon a separate calibration in each country, items with poor psychometric properties were removed. Only in the second stage, a common scale for all countries was assumed (OECD 2009). In PISA 2015, a concurrent calibration was directly applied. Mainly based upon the field trial results, items administered in both assessment modes were constrained to have the same slope and threshold parameters (scalar invariant), or only the slope parameter was constrained to be the same across modes (metric invariant). Above that, in a relatively small number of cases, item constraints where released to allow the estimation of unique (national) item parameters (PISA 2017). Country trends can, however, also be established by estimating item parameters that only make use of national data, which has similarities to the DIF approach as proposed by Zwitser et al. (2017). This latter approach has the advantage that country specific mode effects are taken into account more explicitly by allowing DIF between countries. For the purpose of estimating country trends based on national data only, this paper will calibrate item responses on the mathematics items collected in 2012 and 2015 from the Netherlands by employing an extended nominal response model (ENORM) using conditional maximum likelihood (CML)[2] and the same classifications of scalar and metric invariant items as used in PISA 2015. Person parameters will be estimated by drawing five plausible values for each student.

The results section will evaluate if the mathematics items measured in 2012 PISA by PBA and 2015 PISA by CBA are subject to DIF. Trend estimates will be compared based upon a separate analysis carried out on the Dutch national data. All analyses were conducted in R (R Core Team 2016) using the package "dexter" (Maris et al. 2018). In order to take the data collection design of PISA into account, the R package "survey" (Lumley 2010) was used to estimate the means for the Netherlands in 2012 and 2015.

12.5 Results

PISA 2012 made a distinction between four different types of missing data for the cognitive tests: (1) Item level nonresponse, which indicated an answer to a question was expected, but no response was given by the student; (2) Multiple or invalid responses, referring to (among others) instances of a student selecting multiple answer categories when only one response was expected; (3) Missing by design, referring to questions that were not included in the test version that the student administered or items that were deleted after the assessment because of misprints or translation errors; and (4) Non-reached items, covering consecutive missing values clustered at the end of a test session, except for those coded as item level nonresponse (OECD 2014).

In 2015, five different missing data types were used and classifications were changed. Item level nonresponse was termed "No response/omit" in 2015 and multiple or invalid responses were labelled simply "invalid". The missing by design category was combined with missingness due to students ending the assessment early. Non-reached items remained untouched but the fifth category, "Not Applicable", noted responses for questions that the respondent was directed to skip and responses that could not be determined due to printing problems or torn booklets (OECD 2017).

Figure 12.1 displays the item response category percentages for each item administered in 2012 and 2015. For comparison's sake, the 2015 invalid and not applicable

[2] The IRT model used in PISA 2012 is the Mixed Coefficients Multinomial Logit Model using marginal maximum likelihood (MML) estimation (Adams et al. 1997). For dichotomous items and polytomous items both the ENORM and the Mixed Coefficients Multinomial Logit Model default to the Rasch or the partial credit model respectively (cf. https://dexterities.netlify.com/2018/08/21/dexter-meets-pisa-1/).

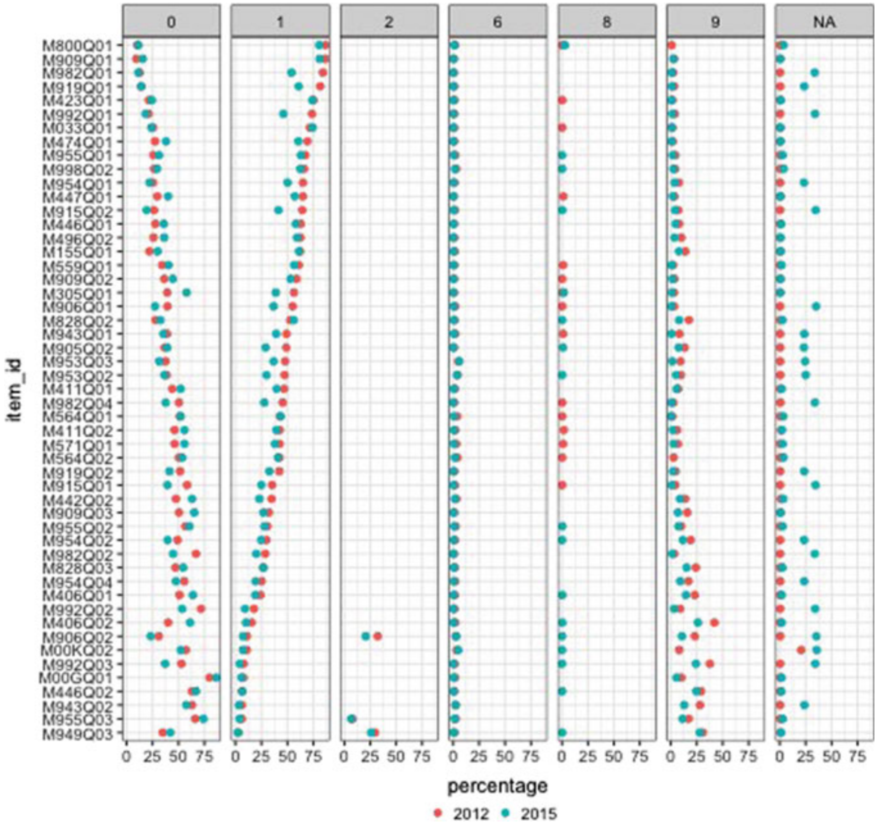


Fig. 12.1 Response category percentages PISA 2012 and 2015

item types were collapsed and compared to the 2012 "Multiple or invalid responses" category.

The first three panels in Fig. 12.1 show respective responses coded as zero (no credit), one (partial credit for polytomously scored items or full credit for dichotomously scored items), and two (full credit). Category six denotes non-reached items, category eight shows not applicable/invalid answers, and category nine corresponds to item nonresponse. Category NA shows that the missing by design (NA) category is also observed within test versions.[3] As mentioned, within the 2012 administration this applies to items that were deleted after assessment because of misprints or translation errors. In 2015 these values within test versions apply to students who did not see the question due to an early ending of the assessment. While that last category had small percentages within the 2012 administration, its values became substantial in 2015 and, for some of the items, up to 30% of the students ended their test earlier than expected. Item nonresponse was slightly higher in 2012 while the response per-

[3] http://www.oecd.org/pisa/data/2015database/Codebook_CMB.xlsx.

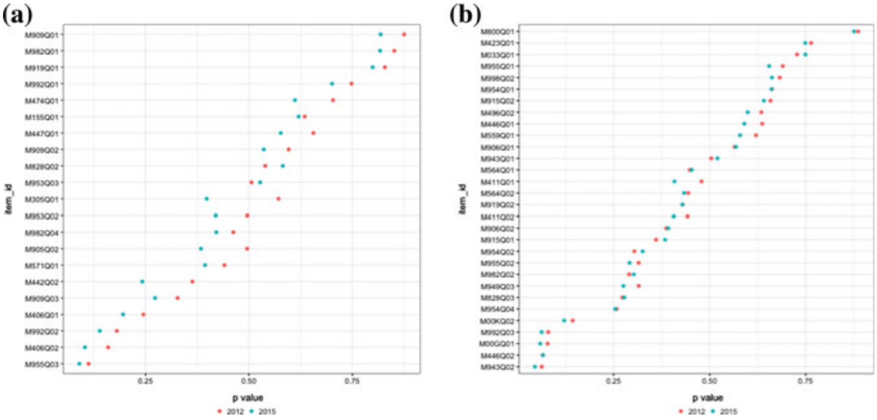


Fig. 12.2 P-values of metric (**a**) and scalar (**b**) invariant items for selected countries in 2012 and 2015

centages for the other missing data categories (non-reached and invalid) were similar across 2012 and 2015. It is important to note that responses coded as non-reached and premature ending of the assessment are considered as "not administered" in the computation of item difficulties and the IRT scaling of items. Figure 12.2 displays the item difficulties, or P-values, of the common items for the selected countries within the included test versions. Items classified by the OECD as metric invariant are found in the left panel and scalar invariant items are presented in the right panel.

As expected, the item difficulty rates of metric invariant items differ substantially between PBA and CBA. As aforementioned item difficulty parameters are also estimated separately for these items. The item difficulty rates of the scalar invariant items are comparable, though the values are somewhat higher in 2012. Nevertheless, the results suggest that the classification into scalar and metric invariant items based upon the field trial results has been advantageous in order to take mode effects into account.

12.5.1 DIF Between Modes

The overall test for DIF between modes indicates that even items classified as scalar invariant items might be subject to DIF (Chi-square = 2209, df = 31, p = <0.01). This indicates that although the procedure used by the consortium to take DIF into account was probably beneficial, there is still reason to believe that results could be subject to DIF. Figure 12.3 provides a visual summary of the item pair DIF Wald test results.

The results found in Fig. 12.3 suggest that especially item M033Q01 shows some indications for DIF. This is actually an item with a higher p-value in 2015. Although

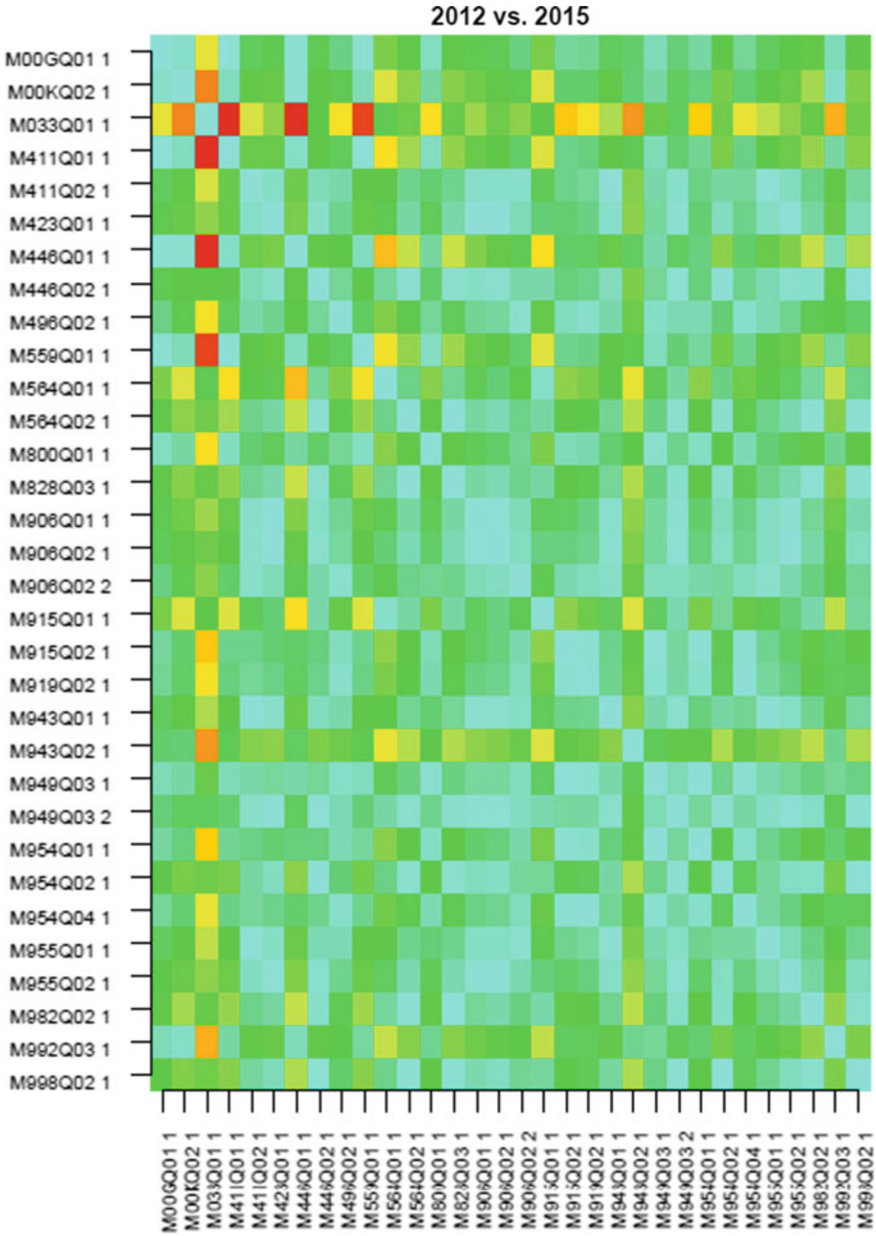


Fig. 12.3 Heat map presenting item pair DIF test results

the relative difficulty of many items in 2015 remain largely unchanged compared to 2012, both the overall test for DIF and the Wald test results for differences between item pair difficulties suggest that results might still be subject to DIF.

12.5.2 Trend Effects in the Netherlands

Figure 12.4 displays the P-values of the trend items in 2012 and 2015 based solely upon the PISA data collected in the Netherlands.

No clear trend can be detected in the unweighted 2012 and 2015 difficulty rates. To assess the mean trend effect using an alternative IRT scaling procedure, only data collected from the Netherlands in 2012 and 2015 were used to estimate item parameters. Based upon these country-specific item parameters the Dutch mean country estimates in 2012 and 2015 have been re-estimated. By applying this approach, the

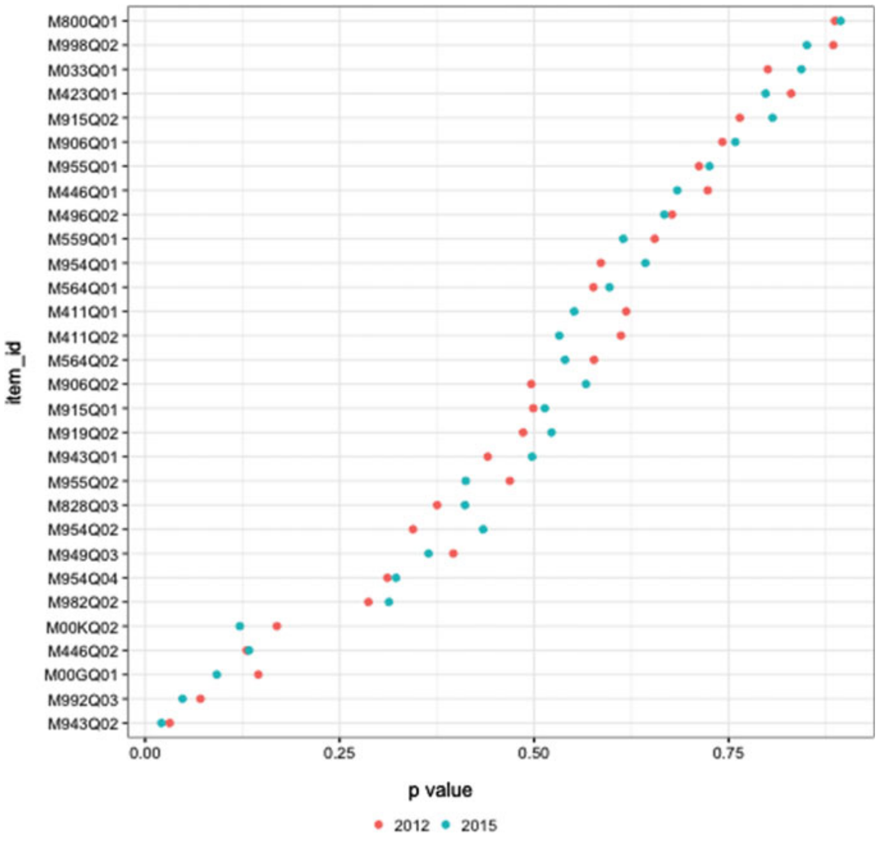


Fig. 12.4 P-values of trend items in the Netherlands for 2012 and 2015

weighted effect size for the differences in mean estimates in the Netherlands is −0.05, about half the size of the reported effect size of −0.12.

12.6 Conclusions and Discussion

Several key changes have been implemented in PISA 2015 and probably the most prominent one was the change of the test administration mode from PBA to CBA. This was a necessary step from many perspectives. The main potential advantage of this move is that a digital test platform facilitates a better representation of the PISA measurement objectives, which emphasize functional knowledge and skills needed to participate in society (OECD 2016a). In particular, the mathematical skills required at work merge with computer use (Hoyles et al. 2002; OECD 2016a). A well-known risk of changing the administration mode is that doing so might jeopardize the comparability of survey results. The contractors of PISA 2015 have taken many measures to accommodate potential trend breaks caused by the switch in the testing mode. A field trial was organized in 2014, prior to the main cycle, to test for mode effects and evaluate the comparability of test results across paper and computer modes (OECD 2016b). Based upon the field trial results, items were classified into scalar, metric and non-metric invariant items with only the first one being used to compare results across modes (OECD 2016c). However, reanalyses of the field trial data collected in Germany found national level mode effects where items administered on a computer were more difficult compared to a paper and pen administration (Robitzsch et al. 2017). Jerrim et al. (2018) concluded (based on a reanalyses of field trial data from Germany, Sweden, and Ireland) that the measures taken during the 2015 cycle have reduced, but not completely resolved, the impact of mode effects.

This study has assessed the extent to which mathematics items, measured in PISA 2012 and 2015 from countries participated using PBA in 2012 and CBA in 2015, are subject to DIF. The performance on scalar invariant items and the evaluation of DIF comparing the main survey results for both years demonstrates that the methodology PISA has adopted was beneficial in accounting for the mode effects. However, the reported decrease for mathematics in the Netherlands between 2012 and 2015, an effect size of −0.12, could not be reproduced with a separate national scaling. Still, a decline in scores was found, but now with a smaller effect size of −0.05. Thus, once DIF between countries is explicitly allowed, the decrease in trend results for the Netherlands is not as large as the reported decrease based upon the calibration procedure applied in PISA 2015. Furthermore, an increase in the number of students ending the assessment early or failing to see a given question was noted.

The reported decrease in mean science scores in 2015 was larger than the decrease in mathematics.[4] This might be due to the introduction of new interactive items, which only took place for the science domain. The decrease in science might be relevant for the scores in mathematics as well, as the majority of reported mathematics plausible values are among other based upon the science scores. The results of this study are only based on the results of students that have been administered one of tests measuring mathematics, which ensures results are not confounded by the potential mode effects of other domains but limits the results to being based upon a subsample of students.

Although many publications on PISA have been made available by the OECD, some information on mode effects and how these effects have been taken into account is missing. For example at the time of writing of this article, the international data from the 2014 field trial were not publicly available. This makes it difficult to fully replicate the analyses carried out by the OECD. Nevertheless, given the results found in this and other recently published studies, the trend outcomes reported for the 2015 PISA should be interpreted with care.

References

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

Bechger, T. M., & Maris, G. (2014). A statistical test for differential item pair functioning. *Psychometrika, 80*(2), 317–340. https://doi.org/10.1007/s11336-014-9408-y.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores* (pp. 395–479). Reading: Addison-Wesley. https://doi.org/10.2307/2283550.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16,* 191–205. https://doi.org/10.1002/j.2333-8504.2001.tb01865.x.

Cohen J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates. https://doi.org/10.2307/2290095.

Feskens, R., Kuhlemeier, H., & Limpens, G. (2016). *Resultaten pisa-2015. Praktische kennis en vaardigheden van 15-jarigen*. Arnhem: Cito. URL: http://www.cito.nl/onderzoekenwetenschap/deelname_int_onderzoek/pisa.

Hong, J., Young, M. J., Brooks, T., Olson, J., & Wang, S. (2007). A meta-analysis of testing mode effects in grade k-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238. https://doi.org/10.1177/0013164406288166.

Hoyles, C., Wolf, A., Molyneux-Hodgson, S., & Kent, P. (2002). *Mathematical skills in the workplace: Final report to the science technology and mathematics council*. Technical report, Institute of Education, University of London. URL http://discovery.ucl.ac.uk/10001565/1/Hoyles2002MathematicalSkills.pdf.

[4] The average of the mean science performance among the 35 OECD countries in 2012 was 501 and decreased to 493 in PISA 2015. For mathematics, this decrease was somewhat less profound: 494 in PISA 2012 to 490 in 2015 (OECD 2016b).

Jerrim, J., Micklewright, J., Heine, J., Salzer, C., & McKeown, C. (2018). Pisa 2015: How big is the ‘mode effect’ and what has been done about it? *Oxford Review of Education, 44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025.

Kingston, N. (2009). Comparability of computer-and paper-administered multiple- choice tests for k-12 populations: A synthesis. *Applied Measurement in Education, 22,* 22–37.

Kolen, M. J., Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York: Springer. https://doi.org/10.1007/978-1-4939-0317-7.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6*(1), 1–24. https://doi.org/10.1207/s15327574ijt0601_1.

Lumley, T. (2010). *Complex surveys*. Wiley-Blackwell. http://dx.doi.org/10.1002/9780470580066.

Maris, G., Bechger, T., Koops, J., Partchev, I. (2018) *dexter: Data management and analysis of tests*. URL https://CRAN.R-project.org/package=dexter. R package version 0.8.1.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/bf02296272.

Mazzeo, J., & von Davier, M. (2008). *Review of the programme for international student assessment (PISA) test design: Recommendations for fostering stability in assessment results*. Technical report, Education Working Papers.

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics. Development, analysis, and application of psychological and educational tests*. The Hague, Netherlands: Eleven international publishing.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/bf02294825.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. London: Routledge. https://doi.org/10.4324/9780203821961.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206.

OECD. (2009). *PISA 2006 technical report*.

OECD. (2014). *Pisa 2012 technical report*. Technical report, OECD, Paris.

OECD. (2016a). *Pisa 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy*. Technical report, OECD, Paris. URL http://dx.doi.org/10.1787/9789264255425-en.

OECD. (2016b). *Pisa 2015 results (volume i): Excellence and equity in education*. Technical report, OECD Publishing, Paris.

OECD. (2016c). *Pisa 2015 results (volume ii): Policies and practices for successful schools*. Technical report, OECD Publishing, Paris.

OECD. (2017). *Pisa 2015 technical report*. Technical report, OECD Publishing, Paris. URL:http://www.oecd.org/pisa/data/2015-technical-report/.

R Core Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Vol. 77). Danisch Institute for Educational Research. 10. 2307/2287805.

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. *Diagnostica, 63*(2), 148–165. https://doi.org/10.1026/0012-1924/a000177.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the naep technology-based assessment project*. Technical report, US Department of Education, National Center for Education Statistics.

Steenkamp, J. M., Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78–90, *25*(1), 78–107. https://doi.org/10.1086/209528.

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at dif in international surveys. *Psychometrika, 82*(1), 210–232. https://doi.org/10.1007/s11336-016-9543-8. ISSN 1860-0980.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 13
Investigating Rater Effects in International Large-Scale Assessments

Hyo Jeong Shin, Matthias von Davier and Kentaro Yamamoto

Abstract The present study investigates rater effects in international large-scale assessments to evaluate the construct validity of constructed-response items scored by human raters. Using the Programme for International Student Assessment data collected in 2015, we illustrate the methods and present the findings about rater effects on constructed-response items in the context of the first fully computer-based international student skill survey. By comparing the unidimensional and the two-dimensional multiple-group item response theory model, it was shown that the latent correlations between human- and machine-coded items were almost perfect. Country means of human- and machine-coded items were generally similar and their differences were small. Further investigation into individual rater effects mostly resulted in small random loadings, which implied that rater effects were negligible in most countries.

13.1 Introduction

Most cognitive assessments are likely to either include multiple-choice (MC) items, constructed-response (CR) items, or both. Unlike the process for scoring MC items, collecting data from CR items often requires human scoring. In the context of international large-scale assessments (ILSAs), ratings from human coders[1] constitute a significant portion of the data and are an essential factor for data quality and, eventually, the reliability and validity of test results. It is common practice to monitor rater reliability by obtaining multiple scores from different raters for a subset of responses

[1] In the context of PISA, human raters are often called "coders." In this chapter, "coders (coding)" and "raters (rating)" are used interchangeably.

H. J. Shin (✉) · K. Yamamoto
Educational Testing Service, Princeton, NJ, USA
e-mail: hshin@ets.org

M. von Davier
National Board of Medical Examiners, Philadelphia, PA, USA

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_13

and evaluating the level of agreement between multiple human raters. For example, human raters in the Programme for International Student Assessment (PISA) are expected to achieve a minimum of 85% agreement with other raters at the item level, based on the subset of common responses. Within each core domain (Mathematics, Reading, Science), an agreement rate of 92% across items is expected. Once the target level of inter-rater reliability is achieved, human-rated scores are assumed to be valid and reliable, and scores from the lead human raters are then used as the final scores in subsequent analytic procedures (such as item calibrations) along with MC item scores (Organisation for Economic Co-operation and Development [OECD] 2017).

Agreement rates between multiple human raters may be sufficient within a given country for a given assessment cycle to examine whether they understand and apply the scoring rubric consistently. In the context of ILSAs, however, it is imperative to also investigate whether raters from different countries and different cycles provide comparable ratings. The main goal of ILSAs is to compare the skills, knowledge, and behaviors of various populations across countries, focusing on group-level scores (Kirsch et al. 2013). This would not be possible without comparability of human-rated scores across countries and cycles; otherwise, the reliability and validity of the test results would be threatened within and across countries and cycles.

Therefore, for CR items to be used effectively in ILSAs, studies have emphasized that raters should be trained to assure that scoring remains valid and comparable across countries and cycles (Mazzeo and von Davier 2008, 2014). Although concerns regarding the validity of human-rated scores and the potential presence of rater bias have been raised in the research literature, a comprehensive investigation has thus far been almost impossible due to a lack of data recording in paper-based assessments (PBA) from the past. For example, in PISA cycles prior to 2015, only a subset of CR items was examined for reliable agreement rates among raters. Coding PBA items in the online coding system was not compulsory. The process saw most participating countries coding CR paper-based items on multiple paper-based coding sheets associated with test booklets, and then entering data into the data management software (OECD 2014). With the introduction of computer-based assessment (CBA) in the 2015 PISA cycle, however, test takers' raw responses to all CBA CR items and corresponding scores, along with the rater assignments on those scores, have been recorded and are easily accessible. Switching to CBA opened up the possibility of investigating the potential presence of rater bias and the reliability and validity of human-rated scores.

Therefore, this study aims to investigate the effect of human rating on construct validity in ILSAs in terms of rater effects that often have been neglected in the past. We use the 2015 PISA data collected for Science, which was the major domain for this cycle and included over 50 human-rated CR items along with over 100 MC items.[2] Two research questions are investigated in this study: (a) the extent to which MC items and CR items similarly measure the underlying latent construct, and (b)

[2] The major domains for PISA included a large number of both new and trend items. The minor domains included a small number of only trend items.

the extent to which individual rater effects exist in the PISA data across countries. In order to answer these research questions, we first apply and compare unidimensional and multidimensional item response models using appropriate variants of these models for analyzing cognitive tests with mixed item types (i.e., MC items and CR items), with different scoring types (i.e., machine- and human-scored items) while accounting for the clustering (i.e., participating countries). We then fit multiple linear regression models specifying rater indicators as input variables to predict the resultant person proficiencies from the multidimensional item response theory (IRT) models. This allows estimation of individual rater effects that are comparable across countries.

13.2 Scoring Human-Coded Items in PISA 2015

13.2.1 Categorization of Items by Item Formats

The definition of item formats can be different and vary by the purpose of the assessment or the target construct. Following the typology of Bennett et al. (1990), an MC item is defined as any item in which the test taker is required to choose an answer from a relatively small set of response options. A CR item is defined as an item that requires the test taker to compose his or her own answer. In PISA, test takers are given a test form consisting of both MC and CR items organized in groups, or testlets, based on a common stimulus. Historically, five format types were identified and initially used in PISA (OECD 2001)[3]: two types of MC (regular and complex multiple-choice items), and three types of CR (closed constructed response, short response, and open constructed response).

Approximately one third of items across the core domains in PISA 2015 were CR items that required human coding, with nearly 50% of them in the Reading domain being CR. To investigate rater effects across countries in ILSA, we used the PISA 2015 Main Survey data and focused on the Science domain, which, as the major domain in the 2015 cycle, had newly developed frameworks and new items (OECD 2016). With the introduction of the CBA mode in PISA from the 2015 cycle, some of the CR items could be scored by a computer while many of the CR items still required coding by human raters. Items with numeric responses (i.e., only numbers, commas, periods, dashes, and back slashes), responses involving choices from a drop-down menu (i.e., radio-button-type items), or selecting rows of data were scored by computer. All others, typically answered by entering text-based answers, were coded by human raters. Hence, the item format and the traditional type of scoring did not necessarily align, so we used "item formats" to distinguish between MC items

[3] Switching to the CBA mode enabled the development and administration of interactive and innovative item formats, such as simulation-based items. In this study, we focus on the distinction between CR and MC, but readers can refer to the PISA 2015 technical report (OECD 2017) for such new item formats.

Table 13.1 Categorization of CBA science items from the PISA 2015 main survey

	Machine-scored	Human-rated
New	69	30
Trend	57	28
Total	126	58

Table 13.2 Number of human coders per country in PISA 2015 science domain

Number of countries	Number of coders	Sample size
2	4	1501–4000
40	8	4001–7000
4	12	7001–9000
6	16	9001–13,000
5	20	13,001–19,000
1	32	19,001–29,000
1	36	More than 29,000

and CR items, and "scoring types" to distinguish between human-rated items and machine-scored items. Because human-rated items were a subset of CR items and machine-scored items included a small number of CR items and all MC items, the focus of comparison was between "human-rated CR items" versus "machine-scored CR and (all) MC items."

There were 184 CBA Science items analyzed in total: 58 human scored (8 polytomous and 50 dichotomous items), and 126 machine scored (5 polytomous and 121 dichotomous items). Ninety-nine items were newly developed items, and 85 were trend items that had been administered in previous cycles. With about a third of new items (30 of 99) being human rated, the importance of studying the validity of human ratings is clear (Table 13.1)

In analyzing the data, we included countries that administered the test via CBA mode only; information about rater assignment linking to their scores was inaccessible via PBA mode. Overall, 59 CBA countries were included in the study. Table 13.2 presents the number of human coders per country in PISA 2015 Science. According to the coding design and procedures (Chapter 13 of the PISA 2015 technical report; OECD 2017), most countries (40 of 59) followed the standard design with 8 coders (grayed in Table 13.2). Two countries with smaller sample sizes had 4 coders, while the 17 remaining countries with larger sample sizes had 12, 16, 20, 32, and 36 coders, respectively, depending on the increasing size of the sampled language group of the assessment.

13.2.2 Coding Design and Procedures

Human coders who evaluate CR item responses and provide scores must be trained to ensure that they adhere to scoring rules, which should be applied consistently for the given assessment, as well as over multiple assessment cycles and across participating countries (Mazzeo and von Davier 2008). Scoring of CR items by human raters is time consuming, expensive, and can be subject to bias, meaning there is potential for inconsistencies across raters, cycles, or countries. Therefore, coder reliability checks in PISA 2015 were for the first time designed to evaluate within- and cross-country levels for all human-rated CR items, facilitated by a coding design that involved *multiple coding*, or coding of the same response independently by different raters. In PISA 2015, it was assumed that the typical number of raw responses to be coded in a single country-language group was around 180,000. Coding design and procedures were made assuming that a single human coder could code 1000 responses per day (it would take 180 days if a single person were to complete the task alone). Multiple coding of all student responses in an international large-scale assessment like PISA is labor intensive and costly. Thus, a subset of test takers' responses was used to evaluate the inter-rater reliability within and across countries.

Within country, 100 student responses per human-coded item were randomly selected for multiple coding. The rest were evenly split among multiple human coders for single coding. The within-country reliability of raters (i.e., exact agreement rates among human raters based on the multiple coding) varied across items and countries. In PISA 2015, 96% of the CBA countries coded every item with proportion agreement higher than 85% in the new Science items. More than 97% of CBA countries had five or fewer items with proportion agreements lower than 85% in the trend Science items. For most CBA countries, the standard inter-rater reliability concerning exact agreement rate was at least 90% for all domains (90% in new Science, 93% in trend Science).

Coder reliability was also evaluated across countries to ensure that the scores on CR items were comparable across countries. In the PISA 2015 cycle, two coders who were bilingual in both English and the language of the assessment additionally scored 10 "anchor responses" for each item. These responses were written in English and were used to compare the application of scoring rules across countries.[4] In general, across-country agreement tended to be lower than within-country agreement; while a mean of 94.2% within-country agreement was observed for trend and new Science items, slightly lower means of 93.6% for trend Science and 93.1% for new Science items were observed across countries. Since the number of anchor responses was only 10 in PISA 2015 and previous cycles, it was increased to 30 for PISA 2018 to make the comparison more robust.

[4] "Anchor responses" were called "control scripts" up to the PISA 2012 cycle, with the number of control scripts varying between 2 and 19 depending on the items (OECD 2014).

13.3 Construct Equivalence of Different Scoring Types in PISA

Messick (1993) argued that “trait (construct) equivalence is essentially an issue of the construct validity of test interpretation and use (p. 61).” With mixed item formats and diverse scoring types in the test, construct equivalence is critical for using and interpreting combined scores from items in different formats and different scoring types as single trait scores. Studies about construct equivalence mostly have focused on the item format, that is, whether MC items and CR items measure the same latent trait and whether the scores from different formats are comparable and valid. For example, Traub (1993) argued “the true-score scales of the MC and CR instruments must be equivalent for the comparison of difficulty to be meaningful” and concluded that there is too little sound evidence in hand regarding the equivalence of item formats, adding that it would vary by domain. More recently, Rodriguez (2003) reviewed the literature and argued that when items are constructed in both formats using the same stem, the correlation between them is higher than using non-stem equivalent items, concluding that the equivalence appeared to be a function of the item design method or the item writer’s intent. In the context of ILSAs, O’Leary (2001) found that item format could influence countries’ rankings in Trends in International Mathematics and Science Study (TIMSS) data and suggested that different experiences of using MC and CR items in different countries can have even more important implications for interpreting test scores. Hastedt and Sibberns (2005) presented a similar conclusion—that there were differences for subgroups of students and for different countries in achievement by item format on TIMSS.

Essentially, construct equivalence of different item formats is closely related to the different scoring types because mostly human raters score CR items. This notion of construct equivalence of different scoring types—whether machine and human-rated items provide consistent information—has not been thoroughly studied in ILSAs yet. Instead, for practical reasons, the focus has been to achieve the operational standards of inter-rater reliability for the CR subset of item responses. Therefore, to examine construct equivalence of the different scoring types, that is, to see if both machine-scored items and human-rated items measure the same construct, we first estimate the correlations between “human-rated CR items” versus “machine-scored CR and (all) MC items.”

13.3.1 Methods

We fit multiple-group IRT models (Bock and Zimowski 1997; von Davier and Yamamoto 2004) the same way as was performed operationally in PISA 2015 (OECD 2017). Multiple-group IRT models enable the estimation of item parameters that are common across different populations, as well as unique group means and standard deviations. Let j denote a person in group k responding in category x_{ij} of item i, and suppose there are g groups and a test composed of n items. Assuming conditional independence of responses, the probability of observing the pattern of response ($\boldsymbol{x}_j$= [$x_{1j}, x_{2j}, \ldots, x_{nj}$]) can be written as

$$P\left(\boldsymbol{x}_j|\theta\right) = \prod_{i=1}^{n} P_i(X_i = x_{ij}|\theta)$$

which applies to all groups and persons, given the person attribute θ. More precisely, the two-parameter logistic model (Birnbaum 1968) for dichotomous items and the general partial credit model (Muraki 1992) for polytomous items were used in modeling the item response function.

Based on these IRT models, items are characterized by item slopes and item locations (difficulties), and the item parameters can be either constrained to be the same across different groups or allowed to be unique for each group. A latent person ability, or attribute θ, follows a continuous distribution with a finite mean and variance in the population of persons corresponding to group k. With the probability density function denoted as $g_k(\theta)$, the marginal probability of response pattern $\boldsymbol{x}_j$ in group k can be expressed as

$$\overline{P_k}\left(\boldsymbol{x}_j\right) = \int_{-\infty}^{\infty} P\left(\boldsymbol{x}_j|\theta\right) g_k(\theta) d\theta.$$

At the item calibration stage in the PISA 2015 Main Survey, each item was allowed to be either common across countries or unique for a specific country-by-language group, or unique within subsets of multiple country-by-language groups (OECD 2017). In this study, we only consider international item parameters because the purpose of the study is to examine rater effects that are comparable across countries. Any systematic rater effect could introduce country-level differential item functioning, and such systematic effects could be already reflected in the unique item parameters allowed for specific countries. Therefore, during the analyses, we fixed all item parameters to the international common item parameters obtained from the PISA 2015 Main Survey for all countries.

In this study, two types of multiple-group IRT models were fit and the results were compared: a unidimensional model that assumed all items measured the Science domain regardless of the scoring type and a two-dimensional model that separated items by scoring types (the first dimension for "human-rated CR items" and the second dimension for "machine-scored CR and [all] MC items.") For this analysis, we used the *mdltm* software (Khorramdel et al., in press; von Davier 2005), which provides marginal maximum likelihood estimation via the expectation-maximization algorithm (Sundberg 1974, 1976; also Dempster et al. 1977).

13.3.2 Findings

Figures 13.1 and 13.2 show distributions of item slopes and item difficulties that were used for fixing the item parameters by item formats. As noted earlier, there were relatively fewer human-coded items (58 vs. 126) in the data. Human-coded items tended to have higher item slopes (i.e., the proportion of items whose slopes are greater than 1) and appeared more difficult compared to machine-coded items (i.e., the percentage of items whose difficulties are greater than zero). This observation is in line with previous studies in large-scale assessments, such as Dossey et al. (1993), who investigated the item formats in the 1992 mathematics assessment of the National Assessment of Educational Progress (NAEP), and Routitsky and Turner (2003), who studied the item formats in PISA 2003. For NAEP, Dossey and colleagues' analyses showed that the extended CR items were much more difficult than MC items and provided considerably more information per item for more proficient students. In a similar vein, using PISA 2003 Field Trial data, Routitsky and Turner asserted that the MC items appeared easier than CR items on average, and CR items showed higher discrimination (i.e., the correlation between item score and the total score) in general.

Table 13.3 shows the comparison of model fit using Akaike information criterion (AIC; Akaike 1974) and Bayesian information criterion (BIC; Schwarz 1978), which penalizes the number of parameters more strongly than AIC. Both model-fit statistics favored the two-dimensional model for the data, which suggested potential differences by scoring types. However, the difference in model-fit improvement based on the Gilula and Haberman (1994) log-penalty measure was negligible. The unidimensional model reached 99.64% model-fit improvement over the baseline model (independence) compared to the more general two-dimensional model. Hence, it

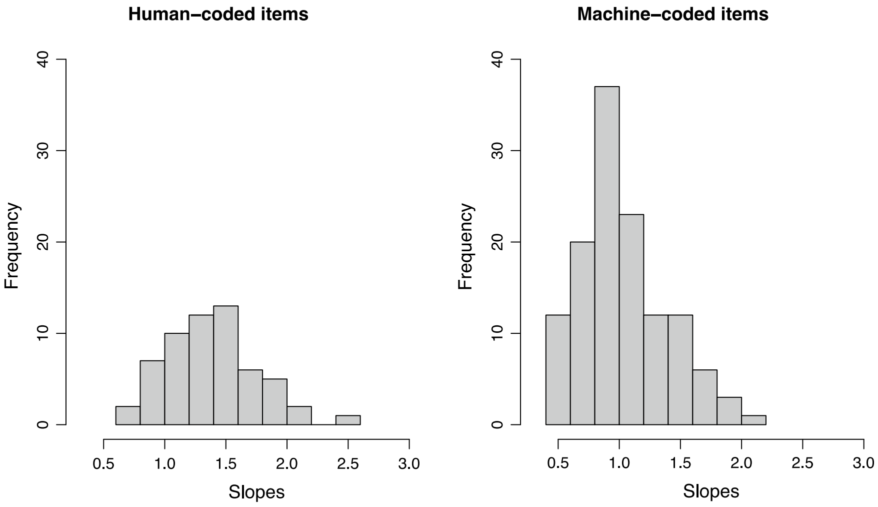


Fig. 13.1 Distribution of item slopes by scoring types (human-coded vs. machine-coded)

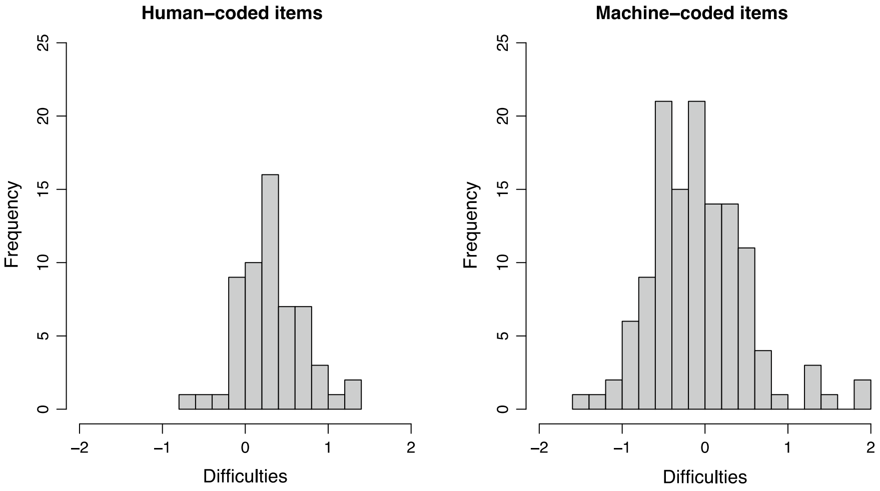


Fig. 13.2 Distribution of item difficulties by scoring types (human-coded vs. machine-coded)

Table 13.3 Comparison of model fit statistics

	AIC	BIC	Log penalty	Percentage improvement
Independence	NA	NA	0.6490	0.00
Unidimensional	9,976,816	9,978,691	0.5668	99.64
Two-dimensional	9,972,179	9,975,302	0.5665	100.00

Note Log Penalty (Gilula and Haberman 1994) provides the negative expected log likelihood per observation; % Improvement compares the log-penalties of the models relative to the difference between the most restrictive and most general model

seems reasonable to assume that MC items and CR items measure a single identifiable latent trait.

Furthermore, the latent correlations were estimated between the two dimensions of scoring types. These correlations were corrected for attenuation by measurement error (Fig. 13.3). The lowest latent correlation among 59 countries was 0.937, and the median latent correlation was 0.974. There were only two countries where the correlation estimates were lower than 0.950. Overall, all countries showed very strong linear associations in their performances on machine- and human-coded items. This implies that human raters may not exert significant effects that should warrant concern about whether human-coded items are considerably different from machine-coded items. Thus, these very high correlations serve as one piece of evidence indicating construct equivalence between different scoring types and the construct validity of the test.

Next, Fig. 13.4 presents the comparison of group-level statistics ($g_k(\theta)$) by different scoring types: the left compares the group (country) means and the right compares the group standard deviations (SD). Regarding the group means, most of

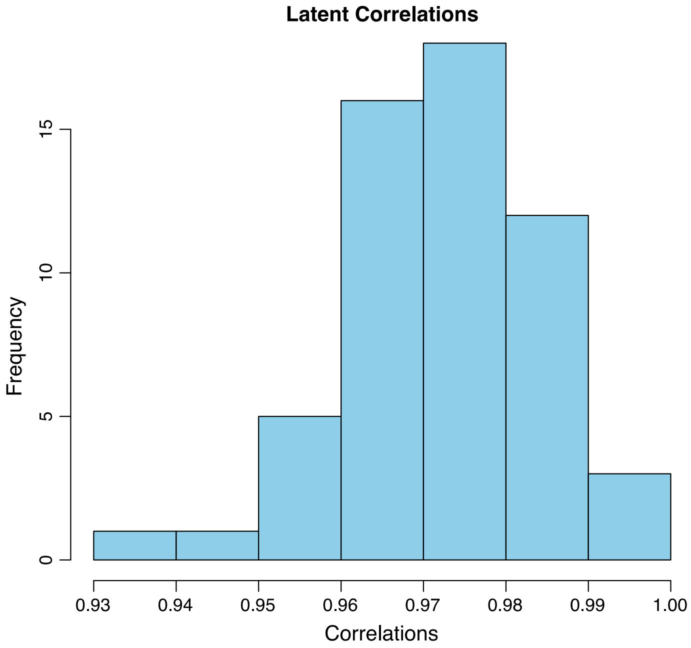


Fig. 13.3 Histogram of latent correlations between machine-coded items and human-coded items

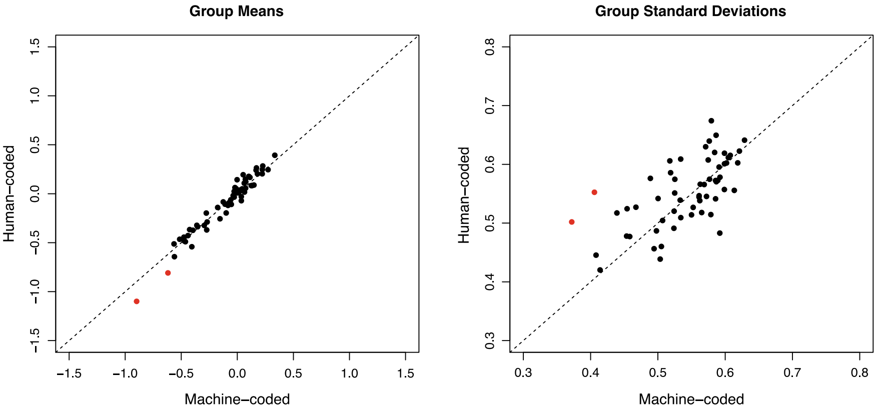


Fig. 13.4 Comparison of group means (left) and standard deviations (right) between machine- and human-coded items

the countries, except for two on the left (solid red dots), showed a consistent pattern in their performance regardless of whether the items were machine or human coded. The shifts in human-coded relative to machine-coded means seem small for all countries, while the largest differences by scoring types were observed in the two lowest-performing countries. Regarding the group SD, there were more notable differences by scoring type, and interestingly, the two lowest performing countries (solid red dots) also showed the largest differences in group SDs: Both showed a higher SD in human-coded items and a smaller SD in machine-coded items. Previous studies noted that the differential behavior of item formats depends on the level of student achievement (DeMars 2000; OECD 2014; Routitsky and Turner 2003). In relation to this, Routitsky and Turner (2003) found a consistent pattern that lower-achieving students performed slightly better on MC than CR items and concluded that lower-achieving countries performed marginally better on MC items because they have a larger proportion of low-achieving students. On the contrary, the OECD (2014, p. 269) calculated the index that indicates achievement on CR items relative to the performance on all other items. The results were that the index was higher in low-performing countries, suggesting that students from low-performing countries were achieving relatively better on the CR items than students from high-achieving countries relative to their achievement on all other items. This pattern might well be attributed to their much lower performance on all other items.

In this study, we suspected that the pattern was observed most likely because these two lowest performing countries could have more missing (omitted) responses on CR relative to MC items. From existing literature it is known that the nonresponse rate is relatively higher for CR items compared to MC items, although that could be the effect of item difficulty or test takers' characteristics (i.e., gender). For example, Routitsky and Turner (2003) reported that CR items tend to comprise difficult sets, with the amount of missing data varying from 9.78 to 57.54% (compared to 1.66–17.62% for MC items), and that item difficulty accounted for about 25% of the missing data (11% for MC items), using PISA 2003 Field Trial data.

Therefore, further investigations were carried out as aimed at skipping response patterns because differences between high and low performers are exaggerated if low performers did not respond to any CR items due to operational scoring rules (e.g., Rose et al. 2017). In present analysis, omitted responses before a valid response were treated as incorrect[5] following scoring rules that have been applied operationally in PISA since its inception. This affects low-performing countries, in particular, as they have omission rates on CR items that are much higher than those observed on MC items relative to other countries. When we calculated the average of the omission rates for each country by scoring types, all 59 country groups showed higher omission rates for human-coded than machine-coded items: The difference in omission rates between human- and machine-coded items ranged between 1.6 and

[5] In PISA, there were two different categories of missing response. If the missing response was followed by valid responses (incorrect or correct) in the subsequent items, it was defined as "omitted." If all the following responses were also missing, it was defined as "not reached." "Omitted" responses were treated as incorrect, but "not reached" responses were treated as missing.

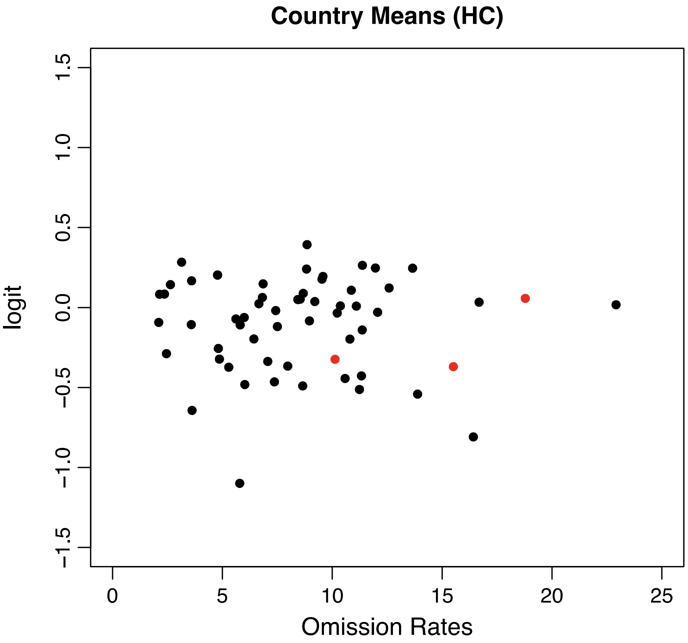


Fig. 13.5 Difference in the country means associated with omission rates

18.4%. Therefore, it was expected that countries with a large number of omits on human-coded items would also tend to have lower scores on those items. Figure 13.5 shows the relationship between those variables when the average omits per country were calculated and associated with the country's performance on the human-coded items. Although the linear association was not seen clearly, three low-performing countries (identified as solid red dots) all showed higher omission rates: Two of them had above 15%, and one had around 10%.

Such a pattern is more apparent when the country means and SDs are sorted by their performance obtained from the unidimensional model that combined two different scoring types. As seen in Fig. 13.6, the difference in the country means by scoring types appears unsubstantial except for the two lowest performing countries. Interestingly, there seems to be a general pattern that lower-performing countries tend to show higher SDs on the human-coded items while higher-performing countries tend to show lower SDs (Figs. 13.6 and 13.7). As shown earlier, this seems to be an artifact due to particularly high nonresponse rates on CR for low-performing countries.

In summary, it seems reasonable to assume that machine- and human-coded items measure the same latent trait, which supports construct equivalence between different scoring types and the construct validity of the test. The two-dimensional model did not add substantial value to that provided by the one-dimensional model, and all countries showed very high correlations of performance measured by human- and machine-

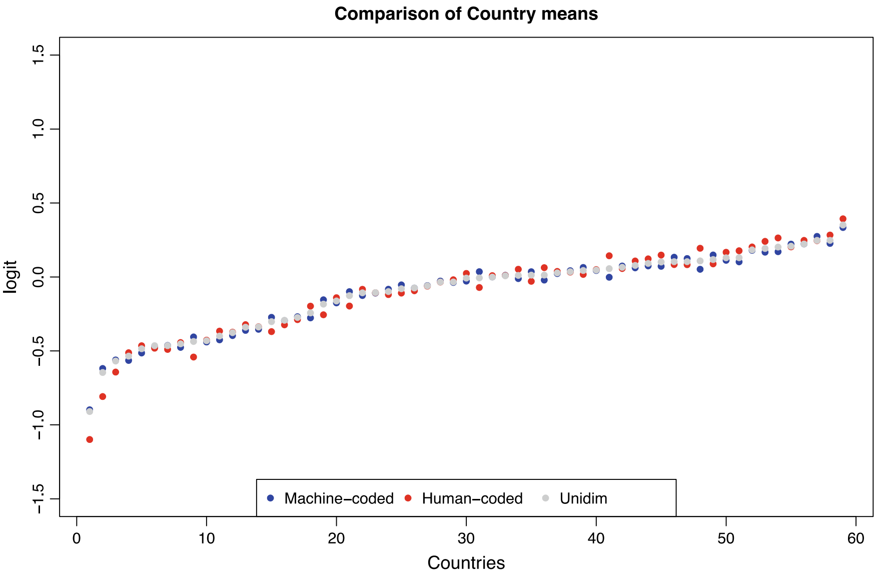


Fig. 13.6 Comparison of country means by scoring types (sorted by group means from the unidimensional IRT model that combined two scoring types)

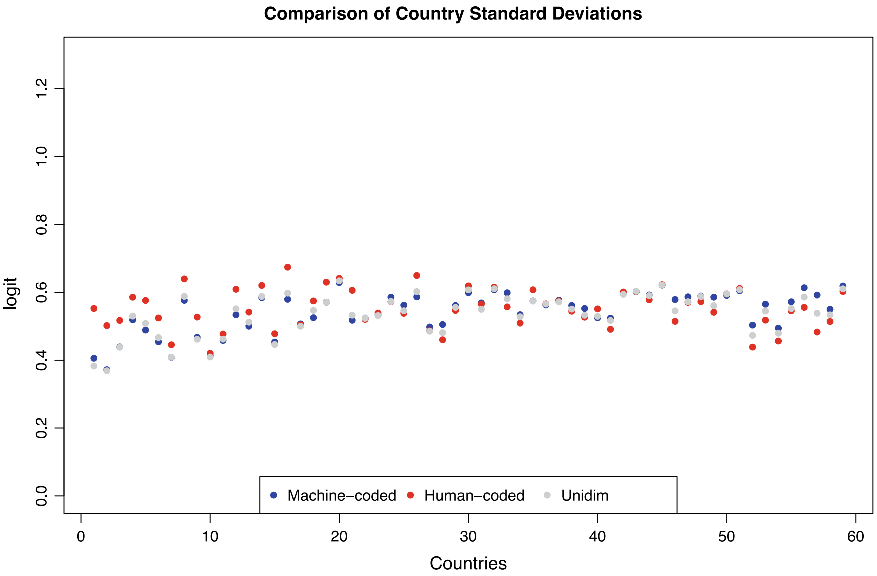


Fig. 13.7 Comparison of country SDs by scoring types (sorted by group means from the unidimensional IRT model that combined two scoring types)

coded items, respectively. Country means of human- and machine-coded items were generally similar and their differences were small for all but the lowest performing countries. The two lowest performing countries showed the largest differences in country means and standard deviations by scoring types, but that pattern seems to be an artifact of scoring rules due to higher omission rates on CR items among low-performing countries. In the next step, we aimed to directly estimate individual rater effects on an internationally comparable scale.

13.4 Rater Effects that Are Comparable Across Countries

13.4.1 Methods

After fitting the two-dimensional multiple-group IRT model separated by scoring types, we fit two types of multiple linear regressions at the international level to estimate the individual rater effects that are comparable across countries. We used the person proficiency estimate as the dependent variable, which was weighted likelihood estimates (WLEs; Warm 1989) based on the second dimension of human-coded items. Below, we use $\widehat{\theta_j^{HC}}$ to denote the WLE of a person j obtained using only the human-coded items and $\widehat{\theta_j^{MC}}$ indicates the WLE for the same person j obtained using only the machine-coded items. To make the interpretation easier, we have standardized those two WLEs to have a mean of 0 and a standard deviation of 1, respectively. In the first model (M1), independent variables included the person proficiency estimates based on the machine-coded items ($\widehat{\theta_j^{MC}}$) and rater-by-country dummy variables. In the second model (M2), only rater-by-country dummy variables were specified to predict the performance on the human-coded items.

$$\widehat{\theta_j^{HC}} = \beta_0 \widehat{\theta_j^{MC}} + \sum_{k=1}^{g} \sum_{r=1}^{R_k} \beta_{rk} N_{rj} d_{rk} \tag{M1}$$

$$\widehat{\theta_j^{HC}} = \sum_{k=1}^{g} \sum_{r=1}^{R_k} \beta_{rk} N_{rj} d_{rk} \tag{M2}$$

Regarding the rater effects, r denotes the rater and R_k the total number of raters in group k. Thus, regression coefficients, β_{rk}, represent the rater effects for rater r in country k, and they are multiplied by N_{rj}, indicating the number of ratings for a person j from rater r associated with the dummy variables (d_{rk}). The number of scores N_{rj} were required because there were multiple raters for each test taker: Different raters were assigned to different item sets and each test taker could have taken a different test form (OECD 2017). In short, each item was coded by a different rater, and each rater was instructed to score a specific set of items. Thus, we calculated the number of item ratings evaluated by individual raters and multiplied those number of ratings to the rater-by-country dummy variables (d_{rk}). These dummy variables take the value 1 for the rater r from the group k, and 0 otherwise. There were 640 dummy variables in total in the data. The number of coefficients is small per country (i.e., no country has more than 36 non-zero dummy variables), while in total we have 640 parameters; per country the number of regression parameters is comparably small. Together with N_{rj} and d_{rk}, the regression coefficients (β_{rk}) represent the average individual rater effects that are weighted by their number of ratings in M2, and weighted average rater effects controlling for the proficiencies on machine-coded items in M1. Finally, the intercept was removed to help the interpretation of the rater effects in a straightforward way, and no country main effects were specified so that rater effects were not centered.

13.4.2 Findings

The overall explained variance at the international level was 0.562 for M1 and 0.135 for M2. A recent study by Shin et al. (in press) analyzed the PISA rater effects of one country and found a negligible effect of raters (4.5% of total variance), arguing that human coding is almost free from subjective bias and that the rater training, scoring rubric, and reliability checks were effective. In this study, the estimate of performance on the machine-coded items ($\widehat{\beta_0}$) in M1 was 0.705 with the standard error of 0.001 (p < 0.001), controlling for the individual rater effects. More importantly, the estimates of individual raters ranged between −0.073 and 0.051 in M1, while the corresponding range was between −0.222 and 0.101 in M2.

Figure 13.8 presents the distributions of 640 individual rater effects ($\widehat{\beta_{rk}}$) that are estimated using an internationally comparable scale across countries. The left panel shows the distribution of estimates obtained from M1 and the right panel from M2. Small random fluctuations in loadings were expected in most countries where rater training worked well. On the other hand, mainly positive loadings were expected in countries where raters tended to be too lenient and negative in countries were expected where raters were too strict. In both cases, the figure clearly shows a majority of small random loadings around zero, which indicated negligible rater effects, hence, rater training seemed to work well in most countries. One notable difference between the two figures was the group of raters with negative loadings (below −0.1) from M2. Nine raters showed severity (below −0.1), and they were all from three low-performing countries, two of which were already identified above in Figs. 13.4 and 13.6 as the lowest performing countries. These raters may have seen many nonresponses, particularly in those low-performing countries. Interestingly, when controlling for the performance on machine-coded items, this pattern diminished. Note that M1 controls for ability measured by MC items so that the rater effects were based on differences relative to the conditional expected performance. Thus, the comparison of these figures suggest that the performance of very low-performing students was attributed to raters' severe scoring practice when performance on machine-coded

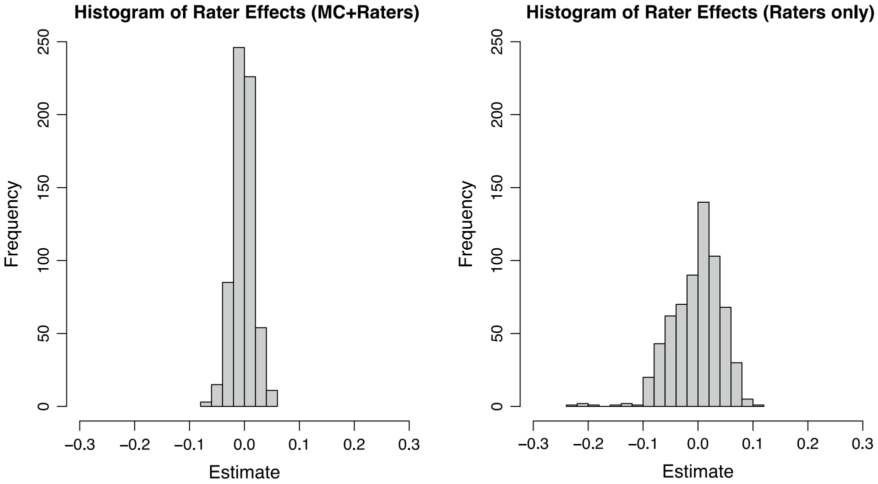


Fig. 13.8 Histogram of rater effects at the international level (left: M1, right: M2)

items was not controlled for. Taken together, small random fluctuations in loadings suggest that rater training worked well in most countries and that human raters did not exert significant effects that should warrant concern about differential rater effects from different countries.

13.5 Conclusion

In the context of ILSAs where a mixed item-format test with a significant portion of CR items is administered, the quality of the scores on the CR items evaluated by human raters is crucial to ensure scoring remains valid and comparable across countries and cycles (Mazzeo and von Davier 2008, 2014). This study investigated rater effects using PISA data collected in 2015 when the assessment switched to CBA as the major mode. Switching from PBA to CBA enabled relatively comprehensive analyses of all CR items, which was previously impossible. By comparing the unidimensional and the two-dimensional multiple-group IRT model, it was shown that the latent correlations between human- and machine-coded items were very high and that the two-dimensional model did not add substantial value to that provided by the one-dimensional model, providing evidence for construct validity by suggesting that the single latent trait is measured by both item types. Country means of human- and machine-coded items were generally similar and their differences were small. Our analysis showed that the two lowest-performing countries also showed the most substantial differences in the country means between two scoring types. Interestingly, larger differences in group SDs were observed as well, and low-performing countries tended to show higher SDs based on the human-coded items compared to the machine-coded items. This seemed to be an artifact of operational scoring rules and higher omitted rates among low-performing countries because high and extremely low performances can be exaggerated if low performers did not respond to any CR items.

Moreover, further investigation of individual rater effects was conducted using multiple linear regressions specifying rater-by-country interactions. The analyses resulted in a distribution of small random loadings, which implied that rater effects were negligible in most countries. There was a group of raters from the three countries whose rater effects were estimated to be too severe, but this effect was diminished when test takers' performance on the machine-coded items were controlled for. In sum, for most countries, rater training seemed to work well, and rater effects were negligible, and hence did not appear to pose any threats to construct validity. At most, two to three countries were identified that might need further investigation on the human-coded CR scores and a review of the scoring rubric and coder training materials. To reiterate, these countries were all low-performing countries where a considerable proportion of students were extremely low achieving, and there seemed no clear evidence of significant rater effects or quality-check efforts specifically on the CR items associated with human coders.

For future studies, rater effects can be investigated concerning the current developments in scoring CR items introduced in ILSAs. For example, the Programme for the International Assessment of Adult Competencies introduced multistage adaptive testing design and a computer-based platform, which automatically and instantaneously scored CR items through the longest common subsequence algorithm (Sukkarieh et al. 2012). More recently in PISA, a machine-supported coding system was implemented operationally for the 2018 cycle and has so far shown to increase the efficiency and accuracy of scoring CR item responses (Yamamoto et al. 2017, 2018). These innovations in scoring CR items using machine learning approaches will reduce the scoring workload of human raters, but it would be worthwhile to investigate whether these tools have a differential effect on scoring correct or incorrect responses, which potentially might affect the estimation of item parameters. A study of the impact of using these tools would be a worthwhile endeavor to assure the measurement invariance and the comparability of item parameters of CR items across countries and cycles.

Finally, systematic rater effects could introduce country-level differential item functioning or indicate the violation of the measurement invariance across countries and cycles. In this study, we have used the data from one assessment cycle and fixed the item parameters using the international parameters obtained at the IRT calibration stage. This was performed intentionally to estimate and separate rater effects that are internationally comparable. A more in-depth look into the comparability across cycles by using multiple cycles of data would be beneficial, and item fit statistics associated with the rater effects might better help the understanding of country- and cycle-level rating processes and behaviors.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items* (Research Report No. RR–90–7). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1990.tb01348.x.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). https://doi.org/10.1007/978-1-4757-2691-6_25.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77. https://psycnet.apa.org/doi/10.1207/s15324818ame1301_3.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*(1), 1–38.

Dossey, J. A., Mullis, I. V. S., & Jones, C O. (1993). *Can students do mathematical problem-solving? Results from constructed-response questions in NAEP's 1992 mathematics assessment*. Washington, DC: National Center for Education Statistics. https://eric.ed.gov/?id=ED362539.

Gilula, Z., & Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *Journal of the American Statistical Association, 89*(426), 645–656.

Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys. *Studies in Educational Evaluation, 31*, 145–161. https://eric.ed.gov/?id=EJ723967.

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzalez & I. Kirsch (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1–11). https://doi.org/10.1007/978-94-007-4629-9.

Khorramdel, L., Shin, H., & von Davier, M. (in press). mdltm (including parallel EM). In M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models—State of the art in modeling, estimation, and applications*.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*. doc.ref. EDU/PISA/GB(2008)28. https://www.researchgate.net/publication/257822388_Review_of_the_Programme_for_International_Student_Assessment_PISA_test_design_Recommendations_for_fostering_stability_in_assessment_results.

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 229–258).

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 61–74). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. https://psycnet.apa.org/record/1993-97248-004.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x.

O'Leary, M. (2001). *Item format as a factor affecting the relative standing of countries in the Trends in International Mathematics and Science Study (TIMSS)*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Organisation for Economic Co-operation and Development. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris, France: OECD Publishing. http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/knowledgeandskillsforlifefirstresultsfrompisa2000-publications2000.htm.

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 technical report*. Paris, France: OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 assessment and analytical framework: Science, reading, math, and financial literacy*. Paris, France: OECD Publishing. http://www.oecd.org/publications/pisa-2015-assessment-and-analytical-framework-9789264281820-en.htm.

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 technical report*. http://www.oecd.org/pisa/data/2015-technical-report/.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163–184. https://doi.org/10.1111/j.1745-3984.2003.tb01102.x.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika, 82,* 795–819. https://doi.org/10.1007/s11336-016-9544-7.

Routitsky, A., & Turner, R. (2003). *Item format types and their influence on cross-national comparisons of student performance*. Presentation given to the Annual Meeting of the American Educational Research Association (AERA) in Chicago, IL. Retrieved from http://works.bepress.com/cgi/viewcontent.cgi?article=1013&context=alla_routitsky.

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

Shin, H., Rabe-Hesketh, S. & Wilson, M. (in press). Trifactor models for multiple-ratings data. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2018.1530091.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics, 1*(2), 49–58.

Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in Statistics—Simulation and Computation. 5*(1), 55–64. https://doi.org/10.1080/03610917608812007.

Sukkarieh, J., von Davier, M. & Yamamoto, K. (2012). *From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks* (Research Report No. RR –12–25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02307.x.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. https://psycnet.apa.org/record/1993-97248-004.

von Davier, M. (2005). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models (Computer software)*. Princeton, NJ: Educational Testing Service.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measuremen*t, *28*(6), 389–406. https://www.ets.org/Media/Research/pdf/RR-03-22-vonDavier.pdf.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450. https://doi.org/10.1007/BF02294627.

Yamamoto, K., He, Q., Shin, H., & von Davier, M. (2017). *Developing a machine-supported coding system for constructed-response items in PISA* (Research Report No. RR-17-47). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12169.

Yamamoto, K., He, Q., Shin, H., & von Davier, M. (2018). Development and implementation of a machine-supported coding system for constructed-response items in PISA. *Psychological Test and Assessment Modeling, 60*(2), 145–164. https://doi.org/10.1002/ets2.12169.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part IV
Computerized Adaptive Testing in Educational Measurement

Chapter 14
Multidimensional Computerized Adaptive Testing for Classifying Examinees

Maaike M. van Groen, Theo J. H. M. Eggen and Bernard P. Veldkamp

Abstract Multidimensional computerized classification testing can be used when classification decisions are required for constructs that have a multidimensional structure. Here, two methods for making those decisions are included for two types of multidimensionality. In the case of between-item multidimensionality, each item is intended to measure just one dimension. In the case of within-item multidimensionality, items are intended to measure multiple or all dimensions. Wald's (1947) sequential probability ratio test and Kingsbury and Weiss (1979) confidence interval method can be applied to multidimensional classification testing. Three methods are included for selecting the items: random item selection, maximization at the current ability estimate, and the weighting method. The last method maximizes information based on a combination of the cutoff points weighted by their distance to the ability estimate. Two examples illustrate the use of the classification and item selection methods.

14.1 Introduction

Computerized classification tests, like computerized adaptive tests, adapt some test elements to the student. Test length and/or item selection is tailored during test administration. The goal of a computerized classification test is to classify the examinee into one of two levels (e.g., master/nonmaster) or into one of multiple levels (e.g., basic/proficient/advanced). Classification methods stop testing when enough confi-

M. M. van Groen (✉) · T. J. H. M. Eggen
Cito, Amsterdamseweg 13, 6814, CM, Arnhem, The Netherlands
e-mail: maaike.vangroen@cito.nl

T. J. H. M. Eggen
e-mail: theo.eggen@cito.nl

T. J. H. M. Eggen · B. P. Veldkamp
Department of Research Methodology, Measurement, and Data Analysis,
University of Twente, P.O. Box 217, 7500, AE Enschede, The Netherlands
e-mail: b.p.veldkamp@utwente.nl

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_14

dence has been established to make the decision. An item selection method selects items for the examinee on the fly.

Computerized classification testing was developed several decades ago for tests intended to measure one construct or dimension. Unidimensional item response theory (UIRT) is used to model the student's performance, to select the items, to decide whether testing can be stopped, and to make the classification decisions. More recently, classification methods were developed for tests measuring multiple constructs or dimensions [(i.e., multidimensionality; Seitz and Frey (2013a, b), Spray et al. (1997), Van Groen et al. (2014b, c, 2016)] using multidimensional item response theory (MIRT). Classification decisions can be made for the entire test or for one or more dimensions or subsets of items.

Two types of multidimensionality Wang and Chen (2004) are discussed in the next section. In the case of between-item multidimensionality, each item is intended to measure just one ability whereas the test as a whole measures multiple abilities. For example, a test contains items for mathematics and language, and each item measures either mathematics or language ability. In case of within-item multidimensionality, items are intended to measure multiple abilities. Here, items measure both mathematics and language. The relative contribution of language and mathematics must vary over items. If not, no distinction can be made between the dimensions and the test should be modeled using UIRT. In what follows, adaptations of two unidimensional classification methods to multidimensional classification testing are discussed. These methods are based on Wald's sequential probability ratio test SPRT; (1947/1973) and Kingsbury and Weiss' confidence interval method (1979). Subsequently, three item selection methods are discussed. The efficiency and accuracy of the classification and item selection methods are investigated in the simulations section. Limitations and recommendations for further research are discussed in the final section.

14.2 Multidimensional Item Response Theory

Multidimensional computerized classification testing requires a statistical framework to model the student's performance, to obtain item parameters, to select the items, to make the classification decisions, and to decide whether testing can be stopped before reaching the maximum test length. MIRT (Reckase 2009) provides such a framework. In a calibrated item bank, model fit is established, item parameter estimates are available, and items with undesired characteristics are removed (Van Groen et al. 2014a). During testing, it is assumed that the item parameters have been estimated with enough precision to consider them known (Veldkamp and Van der Linden 2002).

A vector $\boldsymbol{\theta}$ of p person abilities is used in MIRT to describe the skills and knowledge required for answering an item (Reckase 2009). The dichotomous two-parameter logistic model is used here. The probability of a correct answer, $x_i = 1$, to item i is given by Reckase (2009)

$$\mathrm{P}_i(\boldsymbol{\theta}) = \mathrm{P}_i\left(x_i = 1|\mathbf{a}_i, d_i, \boldsymbol{\theta}\right) = \frac{\exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}{1 + \exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}, \tag{14.1}$$

where $\boldsymbol{a}_i$ is the vector of discrimination parameters and d_i is the easiness of the item.

Ability estimates can be used during testing by the item selection and classification methods in computerized classification testing. Ability is estimated using the likelihood function. The likelihood of a vector of observed responses $\boldsymbol{x}_j = (x_{1j}, ..., x_{kj})$ to items $i = 1, \cdots, k$ for an examinee j with ability $\boldsymbol{\theta}_j$ equals the product of the probabilities of the responses to the administered items (Segall 1996):

$$\mathrm{L}\left(\boldsymbol{\theta}_j|\boldsymbol{x}_j\right) = \prod_{i=1}^{k} \mathrm{P}_i\left(\boldsymbol{\theta}_j\right)^{x_{ij}} \mathrm{Q}_i\left(\boldsymbol{\theta}_j\right)^{1-x_{ij}}, \tag{14.2}$$

where $\mathrm{Q}_i(\boldsymbol{\theta}_j) = 1 - \mathrm{P}_i(\boldsymbol{\theta}_j)$. This likelihood can be used due to the local independence assumption, and it uses the fixed item parameters from the item bank.

The vector of values, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_p)$, that maximizes Eq. 14.2 is used as the ability estimate $\boldsymbol{\theta}_j$ (Segall 1996). Because the equations used to find the maximum likelihood (ML) estimates have no closed-form solution, an iterative search procedure, such as Newton-Raphson, is used. Weighted maximum likelihood (Tam 1992; Van Groen et al. 2016; Warm 1989) estimation reduces the bias in the ML estimates. Alternatively, Bayesian ability estimation approaches are also available (Segall 1996).

The two types of multidimensionality can be distinguished by the structure of their item parameters. If more than one discrimination parameter is nonzero for one or more items, within-item multidimensionality is present (Wang and Chen 2004), and items are intended to measure multiple (or all) abilities. Using a within-item multidimensional model, complex domains can be modeled while taking into account several abilities simultaneously (Hartig and Höhler 2008). Different combinations of abilities can be represented for different items (Hartig and Höhler 2008). If just one discrimination parameter is nonzero per item in the test, the test is considered to have between-item multidimensionality (Wang and Chen 2004). Items are then intended to measure just one ability, and the test contains several unidimensional subscales (Hartig and Höhler 2008).

14.3 Classification Methods

Multidimensional classification testing requires a method that decides whether enough evidence is available to make a classification decision (i.e., testing is stopped and a decision is made). The decision can be a classification into one of two levels (e.g., master/nonmaster) or into one of multiple levels (e.g., basic/proficient/advanced). Two well-known unidimensional methods, the SPRT (Eggen 1999; Reckase 1983; Spray 1993) and the confidence interval method (Kingsbury and

Weiss 1979), were adapted to multidimensional tests. Both methods stop testing when a prespecified amount of confidence has been established to make the classification decision, but the way the decision is made differs. The methods can make a decision based on the entire test and all dimensions, or based on parts of the tests or dimensions. Different implementations of the methods are required for between-item and within-item multidimensionality.

14.3.1 The SPRT for Between-Item Multidimensionality

The SPRT (Wald 1947/1973) was applied to unidimensional classification testing by Eggen (1999), Reckase (1983), and Spray (1993). Seitz and Frey (2013a) applied the SPRT to between-item multidimensionality by making a classification decision per dimension. Van Groen et al. (2014b) extended the SPRT to make classification decisions on the entire between-item multidimensional test.

14.3.1.1 The SPRT for Making a Decision Per Dimension

When making a decision per dimension, a cutoff point, θ_{cl}, is set for each dimension l, $l = 1, ..., p$ (Seitz and Frey 2013a). The cutoff point is used by the SPRT to make the classification decision. Indifference regions are set around the cutoff points and account for the measurement error in decisions for examinees with an ability close to the cutoff point (Eggen 1999). The SPRT compares two hypotheses for each dimension (Seitz and Frey 2013a):

$$\mathrm{H}_{0l} : \theta_{jl} < \theta_{cl} - \delta, \tag{14.3}$$

$$\mathrm{H}_{al} : \theta_{jl} > \theta_{cl} + \delta, \tag{14.4}$$

where δ is the distance between the cutoff point and the end of the indifference region.

The likelihood ratio between the hypotheses for dimension l after k items are administered is calculated for the SPRT (Seitz and Frey 2013a) after each item administration by

$$\mathrm{LR}\left(\theta_{cl} + \delta; \theta_{cl} - \delta\right) = \frac{\mathrm{L}\left(\theta_{cl} + \delta; \mathbf{x}_{jl}\right)}{\mathrm{L}\left(\theta_{cl} - \delta; \mathbf{x}_{jl}\right)}, \quad l = 1, \cdots, p, \tag{14.5}$$

in which $\mathrm{L}\left(\theta_{cl} + \delta; \mathbf{x}_{jl}\right)$ and $\mathrm{L}\left(\theta_{cl} - \delta; \mathbf{x}_{jl}\right)$ are calculated using Eq. 14.2 with those items included that load on dimension l.

Decision rules are then applied to the likelihood ratio to decide whether to continue testing or to make a classification decision (Seitz and Frey 2013a; Van Groen et al. 2014b):

$$\begin{array}{ll} \text{administer another item if} & \beta/(1-\alpha) < \mathrm{LR}(\theta_{cl}+\delta;\theta_{cl}-\delta) < (1-\beta)/\alpha; \\ \text{ability below } \theta_{cl} \text{ if} & \mathrm{LR}(\theta_{cl}+\delta;\theta_{cl}-\delta) \le \beta/(1-\alpha); \\ \text{ability above } \theta_{cl} \text{ if} & \mathrm{LR}(\theta_{cl}+\delta;\theta_{cl}-\delta) \ge (1-\beta)/\alpha, \end{array} \tag{14.6}$$

where α and β specify the acceptable classification error rates (Spray et al. 1997). Previous research has shown that the size of α, β and sometimes also δ has a limited effect on classification accuracy (Van Groen et al. 2014a, c). When classifications need to be made into one of multiple levels, the decision rules are applied sequentially for each of the cutoffs until a decision can be made (Eggen and Straetmans 2000; Van Groen et al. 2014b).

14.3.1.2 The SPRT for Making a Decision on All Dimensions

In many testing situations, a pass/fail decision is required on the entire test in addition to classifications on dimensions of the test. The likelihood then includes all dimensions and items (Van Groen et al. 2014b):

$$\mathrm{LR}\left(\boldsymbol{\theta}_c+\boldsymbol{\delta};\boldsymbol{\theta}_c-\boldsymbol{\delta}\right) = \frac{\mathrm{L}\left(\boldsymbol{\theta}_c+\boldsymbol{\delta};\mathbf{x}_j\right)}{\mathrm{L}\left(\boldsymbol{\theta}_c-\boldsymbol{\delta};\mathbf{x}_j\right)}, \tag{14.7}$$

where $\boldsymbol{\theta}_c$ and $\boldsymbol{\delta}$ include all dimensions. The decision rules for the entire test then become (Van Groen et al. 2014b)

$$\begin{array}{ll} \text{administer another item if} & \beta/(1-\alpha) < \mathrm{LR}(\boldsymbol{\theta}_c+\boldsymbol{\delta};\boldsymbol{\theta}_c-\boldsymbol{\delta}) < (1-\beta)/\alpha; \\ \text{ability below } \boldsymbol{\theta}_c \text{ if} & \mathrm{LR}(\boldsymbol{\theta}_c+\boldsymbol{\delta};\boldsymbol{\theta}_c-\boldsymbol{\delta}) \le \beta/(1-\alpha); \\ \text{ability above } \boldsymbol{\theta}_c \text{ if} & \mathrm{LR}(\boldsymbol{\theta}_c+\boldsymbol{\delta};\boldsymbol{\theta}_c-\boldsymbol{\delta}) \ge (1-\beta)/\alpha. \end{array} \tag{14.8}$$

The SPRT can also make decisions based on a subset of the dimensions or a subset of the items (Van Groen et al. 2014b). This implies that only the selected dimensions and items are included in the likelihoods. This makes it possible to include an item in several decisions.

14.3.2 The Confidence Interval Method for Between-Item Multidimensionality

An alternative unidimensional classification method, developed by Kingsbury and Weiss (1979), uses the confidence interval surrounding the ability estimate. The method stops testing as soon as the cutoff point is outside the confidence interval.

Seitz and Frey (2013b) applied this method to tests with between-item multidimensionality to make a classification decision per dimension. Van Groen et al. (2014c) extended the method to make decisions based on a some or all dimensions within the test.

14.3.2.1 The Confidence Interval Method for Making a Decision Per Dimension

Seitz and Frey (2013b) used the fact that the likelihood function of a between-item multidimensional test reduces to a unidimensional function for each dimension. This unidimensional likelihood function is used to make the classification decisions. The method uses the following decision rules (Van Groen et al. 2014c):

$$\begin{array}{ll} \text{administer another item if} & \hat{\theta}_{jl} - \gamma \cdot \mathrm{se}(\hat{\theta}_{jl}) < \theta_{cl} < \hat{\theta}_{jl} + \gamma \cdot \mathrm{se}(\hat{\theta}_{jl}); \\ \text{ability below } \theta_{cl} \text{ if} & \hat{\theta}_{jl} + \gamma \cdot \mathrm{se}(\hat{\theta}_{jl}) < \theta_{cl}; \\ \text{ability above } \theta_{cl} \text{ if} & \hat{\theta}_{jl} - \gamma \cdot \mathrm{se}(\hat{\theta}_{jl}) > \theta_{cl}, \end{array} \tag{14.9}$$

where $\hat{\theta}_{jl}$ is the estimate examinee's current ability for dimension l, γ is a constant related to the required accuracy (Eggen and Straetmans 2000), and $\mathrm{se}(\hat{\theta}_{jl})$ is the estimate's standard error (Hambleton et al. 1991):

$$\mathrm{se}(\hat{\theta}_{jl}) = \frac{1}{\sqrt{\mathrm{I}(\hat{\theta}_{jl})}}, \tag{14.10}$$

where $\mathrm{I}(\hat{\theta}_{jl})$ is the Fisher information available to estimate θ_{jl} (Mulder and Van der Linden 2009). Fisher information is given by Tam (1992)

$$\mathrm{I}(\hat{\theta}_{jl}) = \sum_{i=1}^{k} a_{il}^2 \mathrm{P}_i(\hat{\theta}_{jl}) \mathrm{Q}_i(\hat{\theta}_{jl}). \tag{14.11}$$

14.3.2.2 The Confidence Interval for Making a Decision on All Dimensions

Seitz and Frey's (2013b) application of the confidence interval method can be used to make decisions per dimension, but not to make decisions on the entire test. An approach developed by Van der Linden (1999) can be used as a starting point for applying the confidence interval method to between-item multidimensional classification testing (Van Groen et al. 2014c).

Van der Linden (1999) considered the parameter of interest to be a linear combination of the abilities, $\boldsymbol{\lambda}'\boldsymbol{\theta}$, where $\boldsymbol{\lambda}$ is a combination of p nonnegative weights. In contrast to Van Groen et al. (2014c), we set $\boldsymbol{\lambda}$ equal to the proportion of items intended to measure the dimension at maximum test length. The composite ability is given by Van der Linden (1999)

$$\zeta = \sum_{l=1}^{p} \theta_l \lambda_l. \tag{14.12}$$

The confidence interval method requires the standard error of the ability estimates (Yao 2012):

$$\mathrm{se}(\zeta) = \sqrt{\mathrm{V}(\zeta)}, \tag{14.13}$$

with variance

$$\mathrm{V}(\zeta) = \boldsymbol{\lambda}'\mathrm{V}(\boldsymbol{\theta})\boldsymbol{\lambda}, \tag{14.14}$$

and $\mathrm{V}(\boldsymbol{\theta}) = \mathrm{I}(\boldsymbol{\theta})^{-1}$. The nondiagonal elements of matrix I are zero, and the diagonal elements can be calculated using Eq. 14.11.

The decision rules are (Van Groen et al. 2014c)

$$\begin{array}{ll} \text{administer another item if} & \hat{\zeta}_j - \gamma \cdot \mathrm{se}(\hat{\zeta}_j) < \zeta_c < \hat{\zeta}_j + \gamma \cdot \mathrm{se}(\hat{\zeta}_j); \\ \text{ability below } \zeta_c \text{ if} & \hat{\zeta}_j + \gamma \cdot \mathrm{se}(\hat{\zeta}_j) < \zeta_c; \\ \text{ability above } \zeta_c \text{ if} & \hat{\zeta}_j - \gamma \cdot \mathrm{se}(\hat{\zeta}_j) > \zeta_c, \end{array} \tag{14.15}$$

where ζ_c denotes the cutoff point and $\hat{\zeta}_j$ is the estimated composite for the student (Van Groen et al. 2014c). The cutoff point is determined with Eq. 14.12 using the cutoff points for the dimensions. If a decision on several, but not all, dimensions simultaneously is required, only those dimensions should be included in the equations.

14.3.3 The SPRT for Within-Item Multidimensionality

The SPRT can also be used for tests with within-item multidimensionality. This does require the computation of the so-called reference composite (Van Groen et al. 2016). The reference composite reduces the multidimensional space to an unidimensional line (Reckase 2009; Wang 1985, 1986). This line is then used to make the classification decision (Van Groen et al. 2016).

The reference composite describes the characteristics of the matrix of discrimination parameters for the items in the item bank (Reckase 2009) or test. The direction of the line is given by the eigenvector of the $\mathbf{aa}'$ matrix that corresponds to the largest eigenvalue of this matrix (Reckase 2009). The elements of the eigenvector determine the direction cosines, $\alpha_{\xi l}$, for the angle between the reference composite and the dimension axes.

$\boldsymbol{\theta}$-points can be projected onto the reference composite (Reckase 2009). A higher value on the reference composite, ξ_j, denotes a more proficient student than does a lower value (Van Groen et al. 2014c). Proficiency on the reference composite can be calculated using an additional line through the $\boldsymbol{\theta}_j$-point and the origin. The length of this line is given by Reckase (2009)

$$\mathrm{L}_j = \sqrt{\sum_{l=1}^{p} \hat{\theta}_{jl}^2}, \tag{14.16}$$

and the direction cosines, α_{jl}, for this line and dimension axis, l, are calculated using (Reckase 2009)

$$\cos \alpha_{jl} = \frac{\hat{\theta}_{jl}}{L_j}, \quad l = 1, \cdots, p. \tag{14.17}$$

The angle, $\alpha_{j\xi} = \alpha_{jl} - \alpha_{\xi l}$, between the reference composite and the student's line is used to calculate the estimated proficiency, $\hat{\xi}_j$, on the reference composite:

$$\hat{\xi}_j = L_j \cos \alpha_{j\xi}. \tag{14.18}$$

The reference composite can now be used to make classification decisions with the SPRT (Van Groen et al. 2016). The cutoff point for the SPRT, ξ_c, and δ^{ξ} are specified on the reference composite. The boundaries of the indifference region need to be transformed to their $\boldsymbol{\theta}$-points using

$$\boldsymbol{\theta}_{\xi_{c+\delta}} = \cos \boldsymbol{\alpha}_{\xi} \times (\xi_c + \delta^{\xi}); \tag{14.19}$$

$$\boldsymbol{\theta}_{\xi_{c-\delta}} = \cos \boldsymbol{\alpha}_{\xi} \times (\xi_c - \delta^{\xi}), \tag{14.20}$$

where $\boldsymbol{\alpha}_{\xi}$ includes the angles between the reference composite and all dimension axes. The likelihood ratio for the SPRT becomes (Van Groen et al. 2016)

$$\mathrm{LR}\left(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}\right) = \frac{\mathrm{L}\left(\boldsymbol{\theta}_{\xi_{c+\delta}}; \mathbf{x}_j\right)}{\mathrm{L}\left(\boldsymbol{\theta}_{\xi_{c-\delta}}; \mathbf{x}_j\right)}, \tag{14.21}$$

which can be used to make multidimensional classification decisions with the following decision rules (Van Groen et al. 2016):

$$\begin{array}{ll} \text{administer another item if} & \beta/(1-\alpha) < \mathrm{LR}(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}) < (1-\beta)/\alpha; \\ \text{ability below } \xi_c \text{ if} & \mathrm{LR}(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}) \leq \beta/(1-\alpha); \\ \text{ability above } \xi_c \text{ if} & \mathrm{LR}(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}) \geq (1-\beta)/\alpha. \end{array} \tag{14.22}$$

Decisions can be made using the reference composite for different subsets of items and for more than two decision levels (Van Groen et al. 2016).

14.3.4 The Confidence Interval Method for Within-Item Multidimensionality

The confidence interval method (Kingsbury and Weiss 1979) can also be used for tests with within-item multidimensionality. Again, the reference composite is used to make a classification decision (Van Groen et al. 2014c).

To determine whether testing can be stopped after an item is administered, the examinee's ability is estimated after each item. This estimate is projected onto the reference composite. The proficiency on the reference composite is then transformed to the corresponding point in the multidimensional space (Van Groen et al. 2014c):

$$\boldsymbol{\theta}_{\hat{\xi}_j} = \cos \boldsymbol{\alpha}_{\xi} \times \hat{\xi}_j. \tag{14.23}$$

The reference composite is considered to be a combination of abilities. The weights between the abilities, $\boldsymbol{\lambda}_{\xi}$, are based on the angles between the reference composite and the dimension axes; $\lambda_{\xi l} = 1/\alpha_{\xi l}$ (Van Groen et al. 2014c).

The standard error for the confidence interval method is given by Yao (2012)

$$\text{se}(\xi) = \sqrt{\text{V}(\xi)}, \tag{14.24}$$

with

$$\text{V}(\xi) = \boldsymbol{\lambda}'_{\xi}\text{V}(\boldsymbol{\theta}_{\xi})\boldsymbol{\lambda}_{\xi}. \tag{14.25}$$

The variance at $\boldsymbol{\theta}_{\hat{\xi}_j}$ is approximated by the inverse of the information matrix at $\boldsymbol{\theta}_{\hat{\xi}_j}$. The decision rules are (Van Groen et al. 2014c)

$$\begin{array}{ll} \text{administer another item if} & \hat{\xi}_j - \gamma \cdot \text{se}(\hat{\xi}_j) < \xi_c < \hat{\xi}_j + \gamma \cdot \text{se}(\hat{\xi}_j); \\ \text{ability below } \xi_c \text{ if} & \hat{\xi}_j + \gamma \cdot \text{se}(\hat{\xi}_j) < \xi_c; \\ \text{ability above } \xi_c \text{ if} & \hat{\xi}_j - \gamma \cdot \text{se}(\hat{\xi}_j) > \xi_c, \end{array} \tag{14.26}$$

where ξ_c denotes the cutoff point on the reference composite. The cutoff point is determined based on the cutoff points for each dimension. Again, decisions can be made using the reference composite for different subsets of items and for more than two decision levels (Van Groen et al. 2016).

14.4 Item Selection Methods

Computerized classification testing requires a method to select the items during test administration. Most methods select an optimal item given some statistical criterion. Limited knowledge is available for selecting items in multidimensional classification testing (Seitz and Frey 2013a; Van Groen et al. 2014b, c, 2016).

Van Groen et al. (2014c) based their item selection on knowledge about item selection methods from unidimensional classification testing (Eggen 1999; Spray and Reckase 1994; Van Groen et al. 2014a, c) and multidimensional computerized adaptive testing for ability estimation (Reckase 2009; Segall 1996; Yao 2012).

Following Van Groen et al. (2014c), three different types of item selection methods are included here for between-item and within-item multidimensionality. In unidimensional classification testing, items are often selected that maximize information at the current ability estimate or at the cutoff point (Eggen 1999; Spray 1993; Thompson 2009). Maximization at the cutoff point becomes more complicated when multiple cutoff points are included. One type of the methods available for selection with multiple cutoff points, that is weighting methods, is also used here (Van Groen et al. 2014a). The third item selection method selects items at random. This provides a benchmark that can be used to compare the efficiency and accuracy of the more complex methods. The method that selects items based on the ability estimate and the weighting method have different implementations for between- and within-item multidimensionality.

14.4.1 Item Selection Methods for Between-Item Multidimensionality

Unidimensional item selection methods can be used for tests with between-item multidimensionality (Van Groen et al. 2014b), but items must be selected per dimension. Here this is accomplished using the Kingsbury and Zara (1989) approach. This approach selects the items from the dimension for which the difference between the desired and achieved percentage is the largest (Van Groen et al. 2014c) and can be applied to the three described item selection methods. The unidimensional item selection method that maximizes information at the ability estimate is described first. The unidimensional weighting method is described thereafter.

14.4.1.1 Item Selection Using the Ability Estimate

In unidimensional classification testing, Fisher information is often maximized at the current ability estimate (Eggen 1999; Spray 1993). The aim of maximizing information at the current ability estimate is to reduce the estimate's standard error. In between-item multidimensional testing, this boils down to (Van Groen et al. 2014b)

$$\max \ \mathrm{I}_i(\hat{\theta}_l), \quad \text{for } i \in \mathrm{V}_{al}, \tag{14.27}$$

where V_{al} is the set of items available for selection for dimension l.

14.4.1.2 Item Selection Using the Weighting Method

Information is often maximized at the cutoff point in unidimensional classification testing (Eggen 1999; Spray 1993). In between-item multidimensional tests, items can be selected that maximize the information at the cutoff point using

$$\max \ \mathrm{I}_i(\theta_{cl}), \quad \text{for } i \in \mathrm{V}_{al} \tag{14.28}$$

as a consequence of the selection per dimension in such tests.

This method can be used if just one cutoff point is specified per dimension. Several methods have been developed for item selection with multiple cutoff points for unidimensional classification testing (Eggen and Straetmans 2000; Van Groen et al. 2014a; Wouda and Eggen 2009), The weighting method combines the objective functions per cutoff point into one weighted objective function. The weight for the cutoff points depends on the distance of the cutoff point to the current ability estimate (Van Groen et al. 2014a). As a result, the test is adapted to the individual student's ability. The item is selected for dimension l that fulfills (Van Groen et al. 2014b)

$$\max \sum_{c=1}^{C} \frac{1}{|\hat{\theta}_{jl} - \theta_{cl}|} \mathrm{I}_i(\theta_{cl}), \ \text{for } i \in \mathrm{V}_{al}. \tag{14.29}$$

14.4.2 Item Selection Methods for Within-Item Multidimensionality

In contrast to between-item multidimensional tests, selecting items for within-item multidimensional tests has to take the multidimensional structure of the test into account. This implies that multidimensional item selection methods have to be used. However, these methods were developed for multidimensional computerized adaptive testing to obtain an efficient and precise ability estimate. Nevertheless, Segall's (1996) method, which selects the items at the current ability estimate, can be used for classification testing. This method was adapted by Van Groen et al. (2016) for selection at a weighted combination of the cutoff points.

14.4.2.1 Item Selection Using the Ability Estimate

Segall (1996) developed an item selection method focused on estimating ability as precisely as possible. Precision is reflected by the size of the confidence ellipsoid surrounding the estimate. The item is selected that results in the largest decrement of the volume of the confidence ellipsoid. The size of the ellipsoid can be approximated by the inverse of the information matrix, so the item is selected that maximizes (Segall 1996)

$$\max\ \det\left(\sum_{i=1}^{k}\mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{ij}) + \mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{k+1,j})\right), \quad \text{for } k+1 \in V_{k+1}, \tag{14.30}$$

where $\sum_{i=1}^{k}\mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{ij})$ denotes the information that has been collected thus far and $\mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{k+1,j})$ is the information that a potential item provides. Because the item is selected that results in the largest determinant of the information matrix, the volume of the confidence ellipsoid is minimized (Reckase 2009).

14.4.2.2 Item Selection Using the Weighting Method

The weighting method for within-item multidimensional classification testing is based on maximization at a weighted combination of the cutoff points (Van Groen et al. 2016). The weights for the dimensions depend on the distance between the cutoff points and the student's proficiency. The weighting method then selects the item using (Van Groen et al. 2016)

$$\max\sum_{c=1}^{C}\frac{1}{|\xi_{\hat{\theta}_j} - \xi_c|}\mathrm{I}_i(\theta_{\xi_c}), \text{ for } i \in \mathrm{V}_{al}. \tag{14.31}$$

14.5 Examples

The accuracy and efficiency of the classification and item selection methods are illustrated using two example simulation studies. The first example deals with a test in which between-item multidimensionality is present. The second example shows a test with a within-item multidimensional structure. The examples demonstrate the differences in accuracy and efficiency in specific situations with specific settings for the classification methods.

14.5.1 Example 1: Between-Item Multidimensionality

The End of Primary School Test (Cito 2012) in the Netherlands can be calibrated with a between-item multidimensional model. The test provides a recommendation on the most suitable level of secondary education to pupils. As a test result, a scale score is given that relates to an advise on the most appropriate level of secondary education (Van Boxtel et al. 2011). The test is typically modeled using six unidimensional scales; four scales for Language (100 items), one scale for Mathematics (60 items), and one scale for Study Skills (40 items).

14.5.1.1 Simulation Design for Example 1

The example is based on the item parameters, ability distribution, and correlation structure from a NOHARM 4 (Fraser and McDonald 2012) calibration with 147,099 students. A dataset of 10,000 students was generated with abilities drawn from the observed correlation structure. The same dataset was used for all simulations. The generated abilities were used to determine the true classification for each student. The minimum test length was set at 20 items, with a maximum of 200 items. Three cutoffs were specified based on the 2012 ability and score distributions. Simulations were run with the SPRT and the confidence interval method. For the SPRT, the values for α and β were set equal and varied between 0.1 and 0.2, and δ was set to 0.4. Other values for δ resulted in similar test lengths and accuracy. For the confidence interval method, γ varied between values corresponding to 90–97.5% confidence intervals for each of the individual cutoff points. Items were selected using random selection, maximization at the ability estimate, and the weighting method. The content structure of the item pool was respected in the item selection using the Kingsbury and Zara approach (1989).

The choices made here were different from those made for actual test administration for the purpose of simplicity and comparability with example 2. The number of cutoffs was limited to three in example 1. The actual test provides eight different advises and also reports at the subject level. The choices made here imply that the presented simulation results cannot be translated to actual test administration.

14.5.1.2 Simulation Results, Example 1

The results of the simulations for example 1 are presented in Figs. 14.1 and 14.2. The former indicates the average test length (ATL) over all students per condition. The latter denotes the proportion of correct decisions (PCD) as an indication of the accuracy of the classification decisions. The PCD compares the true classification and the decision made by the classification methods.

The plot for average test length for the SPRT shows that higher values for α and β result in a shorter test. The plot for the confidence interval method shows that when the confidence interval is increased, the test length also increases. The SPRT results in longer tests than the confidence interval method for these settings of the SPRT and confidence interval method. Random item selection results in the longest tests. Maximization of information at the ability estimate results in the shortest tests.

The plot for the proportion of correct decisions for the SPRT indicates that 80–87% of the classifications are accurate. The plot for the confidence interval method indicates that the settings for γ have a limited influence on classification accuracy. The plot also shows that random selection results in slightly less accurate decisions. The SPRT results in more accurate decisions than the confidence interval method.

Based on the these simulations, one might conclude that the SPRT results in longer tests with more accurate decisions. However, this might be caused by the settings of the SPRT and the confidence interval method in the example.

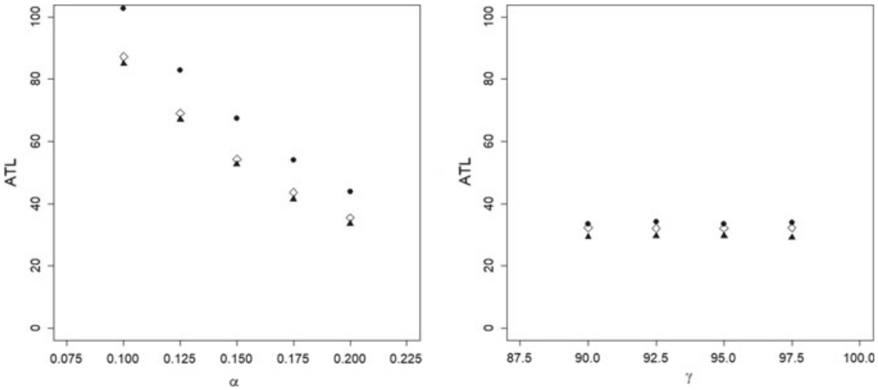


Fig. 14.1 Average test length, example 1. Left plot: SPRT; right plot: confidence interval method. $\delta = 0.4$. Dots: random selection; triangles: maximization at the ability estimate; diamonds: weighting method

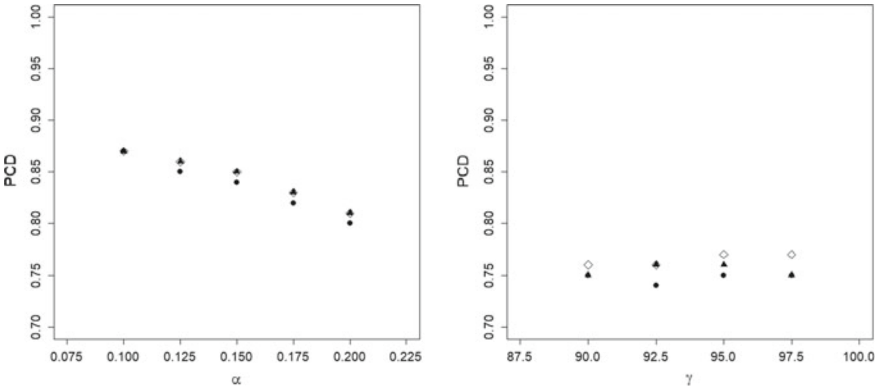


Fig. 14.2 Proportion of correct decisions, example 1. Left plot: SPRT; right plot: confidence interval method. $\delta = 0.4$. Dots: random selection; triangles: maximization at the ability estimate; diamonds: weighting method

14.5.2 Example 2: Within-Item Multidimensionality

An within-item multidimensional ACT mathematics item pool, as in Ackerman (1994) and Veldkamp and Van der Linden (2002), was used in the second example. The item parameters had been obtained previously using an orthogonal two-dimensional NOHARM II (Fraser and McDonald 1988) calibration. Items are intended to measure coordinate geometry, elementary algebra, intermediate algebra, pre-algebra, plane geometry, and trigonometry, and they require three skill categories; analysis, application, and basic skills (Veldkamp and Van der Linden 2002).

14.5.2.1 Simulation Design, Example 2

The standard normal multivariate ability distribution was used to generate 10,000 students. The cutoff points were specified on the reference composite based on the 33th and 66th percentiles of the underlying dimensions. Simulations were generated with $\alpha = \beta$ between 0.025 and 0.10, and δ was set at 0.40. γ varied between values corresponding to 90–97.5% confidence intervals. A maximum of 50 items was selected using random selection, maximization at the ability estimate, and the weighting method. No content restrictions were made in the simulations.

14.5.2.2 Simulation Results, Example 2

The results of the simulations for example 2 can be seen in Figs. 14.3 and 14.4. The plot for average test length for the SPRT shows that higher values for α and β result in a shorter test. As expected, the average test length increased when a larger confidence interval was required. The plots also suggest that test length is often shorter for the confidence interval method. However, this depends on the settings for the SPRT and the confidence interval method. Random selection results in the longest tests, and the weighting method results in the shortest test. The effect of the item selection method appears to be consistent regardless of the classification method.

The plot for the proportion of correct decisions for the SPRT indicates that the values for α and β did not influence the accuracy of the decisions. In contrast, the size of the confidence interval did influence the accuracy. As expected, random selection resulted in the least accurate decisions. The confidence interval method resulted in less accurate decisions than did the SPRT. This might be caused by the shorter tests for the confidence interval method.

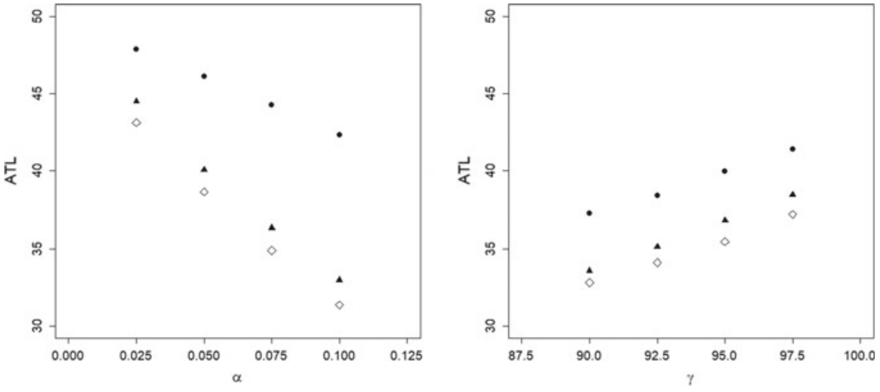


Fig. 14.3 Average test length, example 2. Left plot: SPRT; right plot: confidence interval method. $\delta = 0.4$. Dots: random selection; triangles: maximization at the ability estimate; diamonds: weighting method

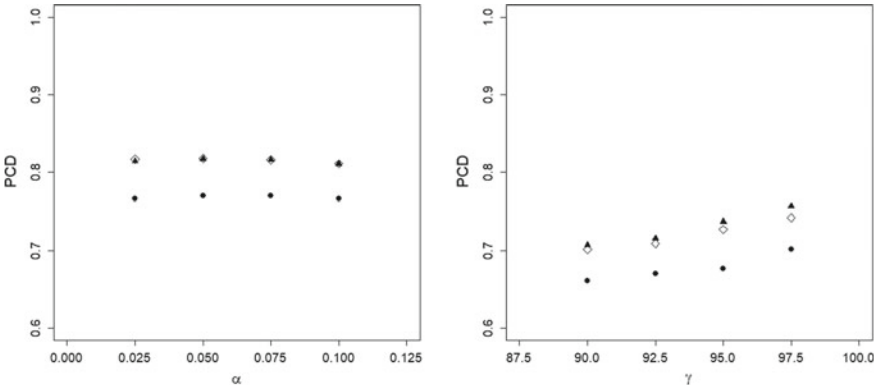


Fig. 14.4 Proportion of correct decisions, example 2. Left plot: SPRT; right plot: confidence interval method. $\delta = 0.4$. Dots: random selection; triangles: maximization at the ability estimate; diamonds: weighting method

Based on the current settings for the SPRT and confidence interval method, one might conclude that the SPRT resulted in longer tests in this example dataset. However, those longer tests result in more accurate classification decisions. This finding clearly indicates the well-known trade-off between accuracy and test length.

14.6 Conclusions and Discussion

Multidimensional computerized classification testing can be used when classification decisions are required for constructs that have a multidimensional structure. Here, two methods for making those decisions were included here for two types of multidimensionality. In the case of between-item multidimensionality, each item is intended to measure just one dimension. By contrast, in case of within-item multidimensionality, items are intended to measure multiple or all dimensions.

Wald's (1947) sequential probability ratio test has previously been applied to multidimensional classification testing by Seitz and Frey (2013a) and Van Groen et al. (2014b) for between-item multidimensionality and by Van Groen et al. (2016) for within-item multidimensionality. Kingsbury and Weiss's (1979) confidence interval method was applied to between-item multidimensionality by Seitz and Frey (2013b) and Van Groen et al. (2014c), and by Van Groen et al. (2014c) to within-item multidimensionality. The present study's implementation of the confidence interval method for between-item dimensionality differs from the implementation in Van Groen et al. (2014c). Here, different weights were used that reflect the number of items per dimension relative to the test length.

Three methods were included for selecting the items: random item selection, maximization at the current ability estimate, and the weighting method. The last method maximizes information based on a combination of the cutoff points weighted by their distance to the ability estimate. Two examples illustrated the use of the classification and item selection methods.

Classifications were made into one of three classification levels on the entire test. Van Groen et al. (2014c) showed classifications on both the entire test as well as on some of the items or dimensions. They also required that decisions had to be made on all partial classification decisions before a classification decision could be made on the entire test. Here, only one decision was made per test.

Examples were included to demonstrate the use of the classification and item selection methods. Simulations included only two different item banks, a limited number of settings for the classification methods, and a limited number of students. More thorough studies are needed in order to draw conclusions regarding the most effective settings and classification methods. Van Groen et al. (2014c) provided some further suggestions for comparison of the classification methods. Moreover, only three item selection methods were included here. Many more item selection methods (Reckase 2009; Segall 1996; Yao 2012) exist for multidimensional adaptive testing. These should be investigated as well. Finally, α, β, δ, and γ were set equal for all cutoffs points. The use of different values for different cutoff points should be investigated as well.

Two classification methods were included here. A third unidimensional classification method, the generalized likelihood ratio test developed by Bartroff, Finkelman, and Bartroff et al. (2008), has never been applied to multidimensional classification testing. This method is based on the sequential probability ratio test. This suggests that it should be feasible to expand the method to multidimensional classification testing.

Research on multidimensional classification testing is still limited to the work of (Seitz and Frey 2013a, b), Spray et al. (1997), and the current authors (Van Groen et al. 2014b, c, 2016). Further studies will be required before multidimensional classification testing can be applied in practice. To date, too little is know about the classification methods, the item selection methods in this context, and the settings for the classification methods to administer multidimensional classification tests with confidence.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*. https://doi.org/10.1207/s15324818ame0704_1.

Bartroff, J., Finkelman, M. D., & Lai, T. L. (2008). Modern sequential analysis and its application to computerized adaptive testing. *Psychometrika*. https://doi.org/10.1007/s11336-007-9053-9.

Cito. (2012). *Eindtoets Basisonderwijs 2012 (End of primary school test 2012)*. Arnhem, The Netherlands: Cito.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*. https://doi.org/10.1177/01466219922031365.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131640021970862.

Fraser, C., & McDonald, R. P. (1988). *NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory* (Computer Software).

Fraser, C., & McDonald, R. P. (2012). *NOHARM 4: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* (Computer Software).

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology*. https://doi.org/10.1027/0044-3409.216.2.89.

Kingsbury, G. G., & Weiss, D. J. (1979). *An adaptive testing strategy for mastery decisions (Research Report 79–5)*. Minneapolis, M.N.: University of Minnesota Press.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education*. https://doi.org/10.1207/s15324818ame0204_6.

Mulder, J., & Van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*. https://doi.org/10.1007/S11336-008-9097-5.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237–254). New York, NY: Academic Press.

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

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*. https://doi.org/10.1007/BF02294343.

Seitz, N.-N., & Frey, A. (2013a). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling*, *55*(1), 105–123.

Seitz, N.-N., & Frey, A. (2013b). *Confidence interval-based classification for multidimensional adaptive testing* (Manuscript submitted for publication).

Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (Report No. ACT-RR-93-7). Iowa City, IA: American College Testing.

Spray, J. A., Abdel-Fattah, A. A., Huang, C.-Y., & Lau, C. A. (1997). *Unidimensional approximations for a computerized adaptive test when the item pool and latent space are multidimensional* (Report No. 97-5). Iowa City, IA: American College Testing.

Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the national meeting of the National Council on Measurement in Education, New Orleans, LA.

Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait* (Unpublished doctoral dissertation, Columbia University, New York, NY).

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164408324460.

Van Boxtel, H., Engelen, R., & De Wijs, A. (2011). *Wetenschappelijke verantwoording van de Eindtoets 2010 (Scientific report for the end of primary school test 2010)*. Arnhem, The Netherlands: Cito.

Van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986024004398.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014a). Item selection methods based on multiple objective approaches for classification of respondents into multiple levels. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621613509723.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014b). Multidimensional computerized adaptive testing for classifying examinees on tests with between-dimensionality. In M. M. van Groen (Ed.), *Adaptive testing for making unidimensional and multidimensional classification decisions* (pp. 45–71) (Unpublished doctoral dissertation, University of Twente, Enschede, The Netherlands).

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014c). Multidimensional computerized adaptive testing for classifying examinees with the SPRT and the confidence interval method. In M. M. van Groen (Ed.), *Adaptive testing for making unidimensional and multidimensional classification decisions* (pp. 101–130) (Unpublished doctoral dissertation, University of Twente, Enschede, The Netherlands).

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2016). Multidimensional computerized adaptive testing for classifying examinees with within-dimensionality. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621616648931.

Veldkamp, B. P., & Van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*. https://doi.org/10.1007/BF02295132.

Wald, A. (1973). *Sequential analysis*. New York, NY: Dover (Original work published 1947).

Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effect of latent space misspecification on the application of IRT (Research Report MW: 6–24-85)*. Iowa City, IA: University of Iowa Press.

Wang, M. (1986). *Fitting a unidimensional model to multidimensional item response data*. Paper presented at the Office of Naval Research Contractors Meeting, Gatlinburg, TN.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621604265938.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*. https://doi.org/10.1007/BF02294627.

Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika*. https://doi.org/10.1007/S11336-012-9265-5.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 15
Robust Computerized Adaptive Testing

Bernard P. Veldkamp and Angela J. Verschoor

Abstract In order to optimize measurement precision in computerized adaptive testing (CAT), items are often selected based on the amount of information they provide about a candidate. The amount of information is calculated using item- and person parameters that have been estimated. Usually, uncertainty in these estimates is not taken into account in the item selection process. Maximizing Fisher information, for example, tends to favor items with positive estimation errors in the discrimination parameter and negative estimation errors in the guessing parameter. This is also referred to as capitalization on chance in adaptive testing. Not taking the uncertainty into account might be a serious threat to both the validity and viability of computerized adaptive testing. Previous research on linear test forms showed quite an effect on the precision of the resulting ability estimates. In this chapter, robust test assembly is presented as an alternative method that accounts for uncertainty in the item parameters in CAT assembly. In a simulation study, the effects of robust test assembly are shown. The impact turned out to be smaller than expected. Some theoretical considerations are shared. Finally, the implications are discussed.

15.1 Introduction

In computerized adaptive testing (CAT), the items are administered such that the difficulty level is tailored to the test taker's ability level. Adaptive testing turns out to entail a number of advantages. Candidates only have to answer items that are paired to their ability level, test length can be reduced in comparison to linear test forms, and test administration can be more flexible in terms of time and location as a result of individualized testing. CATs could be offered continuously, on flexible locations, on portable devices, and even via the Web. The advantages of CAT are very appealing for

B. P. Veldkamp (✉)
University of Twente, Enschede, The Netherlands
e-mail: b.p.veldkamp@utwente.nl

A. J. Verschoor
Cito, Arnhem, The Netherlands

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_15

candidates who live in a 21st century world, with tablets, mobile phones and who are continuously online. Computerized adaptive testing is a more and more popular test administration mode in educational, psychological, and health measurement. Many algorithms for tailoring the difficulty level of the test to the individual's ability level have been proposed in the literature (e.g. Eggen 2004, page 6). These algorithms generally consist of the following steps:

1. Before testing begins, the ability estimate of the candidate is initialized (e.g., at the mode of the ability distribution, or based on historical data).
2. Items are selected from an item bank to be maximally informative at the current ability estimate. Sometimes, a number of specifications related to test content or other attributes have to be met, which restricts the number of items available for selection. In this step, an exposure control method is commonly applied to prevent overexposure of the most popular items.
3. Once an item is selected, it is administered to the candidate.
4. The responses are scored.
5. An update of the ability estimate is made after each administration of an item.
6. Finally, the test ends whenever a stopping criterion has been met, for example when a fixed number of items have been administered or when a minimum level of measurement precision has been obtained.

One of the prerequisites of CAT is that a calibrated item pool is available and that the item parameters have been estimated with enough precision to be treated as fixed values. These parameters are used during test administration to calculate the amount of information each item provides and to estimate the ability levels. Unfortunately, item parameters are calibrated with a finite sample of candidates. The resulting item parameter estimates might be unbiased, but they still contain measurement error. This measurement error, which causes uncertainty in the true values of the item parameters, is a source of concern. Previous research on item calibration error in adaptive testing (van der Linden and Glas 2000) already mentioned that items with high discrimination parameters tend to be selected more often from the bank, when items are selected based on the amount of information they provide at the estimated ability level. Especially, positive estimation errors in the discrimination parameters have quite some impact on the amount of information provided. Overestimation of item discrimination will increase the probability that the item will be selected. This phenomenon is also referred to as the problem of capitalization on chance.

Both Tsutakawa and Johnson (1990) and Hambleton and Jones (1994) already studied the effects of item parameter uncertainty on automated assembly of linear test forms. Hambleton and Jones found out that not taking the uncertainty into account resulted in serious overestimation (up to 40%) of the amount of information in the test. To illustrate the effect, Veldkamp (2013) illustrated this with a simulated item pool that consisted of 100 items. The parameters for all of these items were drawn from the same multivariate item parameter distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)$. The mean values of $\boldsymbol{\mu}$ were equal to the true item parameters of a parent item. The discrimination parameter of the parent was equal to a = 1.4, the difficulty parameter equal to b = 0.0, and the guessing parameter equal to c = 0.2. The variance covariance matrix $\boldsymbol{\Sigma}$ was equal

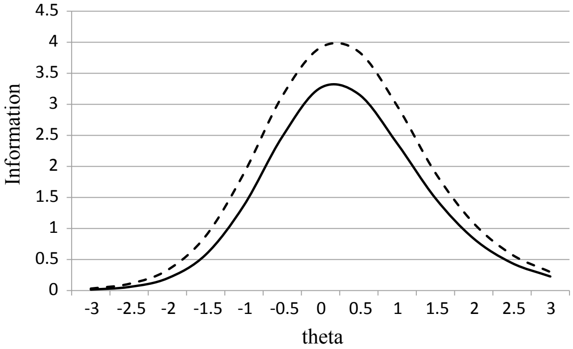


Fig. 15.1 Test information function: ATA (dashed line) or true (solid line)

to the diagonal matrix with the standard errors of estimation ($SE\,a = 0.05$, $SE\,b = 0.10$, $SE\,c = 0.02$) on the diagonal. Because of this, the item parameters only varied due to uncertainty in the parameter estimates. Resulting parameters fell in the intervals $a \in [1.29, 1.52]$, $b \in [-0.31, 0.29]$, and $c \in [0.14, 0.28]$. To illustrate the effects of item parameter uncertainty, a linear test of ten items was selected from this item bank. Fisher information at $\theta = 0.0$ was maximized during test assembly. A comparison of test information functions was made between this test and a test consisting of 10 items with parameter equal to the parameters of the parent item. The results are shown in Fig. 15.1. As can be seen, the test information is overestimated by 20% when uncertainty due to simulated item calibration errors was not taken into account.

Hambleton and Jones (1994) demonstrated that the impact of item parameter uncertainty on automated construction of linear tests depended on both the calibration sample size and the ratio of item bank size to test length. When their findings are generalized to computerized adaptive testing, the impact of calibration sample size is comparable. Calibration error will be larger for smaller samples. For the ratio of item bank size to test length the effects are even larger. In CAT, only one item is selected at a time. The ratio of item pool size to test length is therefore even less favorable. Van der Linden and Glas (2000), studied the impact of capitalization on chance for various settings of CAT in an extensive simulation study, and they confirmed the observations of Hambleton and Jones (1994). In other words, capitalization on chance is a problem in CAT when items are selected based on the amount of information they provide. As a result, the measurement precision of the CATs might be vastly overestimated. Item selection algorithms, therefore have to be modified to account for capitalization on chance.

15.2 Robust Test Assembly

Automated test assembly problems can be formulated as mixed integer linear programming problems. An extensive introduction on how to formulate the mixed integer linear programming problems can be found in van der Linden (2005). These mixed integer programming problems have a general structure where one feature of the test is optimized and specifications for other features are met. For example, the amount of information can be maximized while specifications with respect to the content, the type of items, and the test length have to be met. When a mixed integer linear programming approach is used, the parameters in the model are assumed to be fixed. Due to, for example, calibration error, there is some uncertainty in the parameters and robust optimization methods have to be used. The general idea underlying robust optimization is to take uncertainty into account when the problem is solved in order to make the final solution immune against this uncertainty (Ben Tal et al. 2009).

One of the early methods to deal with item parameter uncertainty in optimization problems was proposed by Soyster (1973). He proposed a very conservative approach, where each uncertain parameter in the model was replaced by its infimum. In this way, a robust lower bound to the solution of the optimization problem could be found. The resulting lower bound turned out to be very conservative though, since it assumed a maximum error in all the parameters, which is extremely unlikely to happen in practice. A modified version of this method was applied to automated assembly of linear tests by de Jong et al. (2009). They took uncertainty due to calibration error into account. The calibration errors were assumed to be normally distributed. But instead of using the infima of these distributions, they subtracted one posterior standard deviation from the estimated Fisher information as a robust alternative. This approach was even studied more into detail by Veldkamp et al. (2013), who studied the effects of uncertainties in various item parameters on Fisher information in the assembly of linear test forms.

A more realistic method to deal with uncertainty in optimization problems was proposed by Bertsimas and Sim (2003). They noted that it almost never happens in practice that uncertainty plays a role for all of the parameters in the model. Instead, uncertainty in a few of the parameters really affects the final solution. They proposed an optimization method where uncertainty only plays a role for Γ of the parameters. For this situation, they proved that finding an optimal solution when at most Γ parameters are allowed to change, is equal to solving $(\Gamma + 1)$ mixed integer optimization problems. In other words, this robust optimization method will be more time consuming, but we can still apply standard software for solving mixed integer programming methods. In automated test assembly, calibration errors are assumed to be normally distributed, and extreme overestimation or underestimation of the item parameters is only expected for a few items the item pool. This resembles the observations of Bertsimas and Sim that uncertainty only affects the final solution for some of the parameters. Therefore, the mixed integer optimization methods for automated test assembly proposed in van der Linden (2005) can still be applied, although the test assembly models are more complicated and more time consuming to solve. For an application of Bertsimas and Sim (2003) to linear test assembly, see Veldkamp (2013).

15.3 Robust CAT Assembly

Good results were obtained for some practical test assembly problems with the modified Soyster method (see de Jong et al. 2009) and the Bertsimas and Sim method (see Veldkamp 2013). Both methods replace estimated item parameters by a more conservative value either for all or for some of the items, by subtracting one or three standard deviations. These robust optimization methods originate from the field of combinatorial optimization.

A different approach, that originated in the field of psychometrics, can be found in Lewis (1985) where expected response functions (ERFs) are proposed to correct for uncertainty in the item parameters (Mislevy et al. 1994) in the process of constructing fixed-length linear tests. To apply the approach to CAT assembly, ERFs have to be apply at the item pool level. This might result in a starting point for a robust CAT assembly procedures.

15.3.1 Constructing a Robust Item Pool

The calibration error follows a normal distribution. When the distribution of the errors is used, it can be derived which percentage of items will have a certain deviation from the mean. Straightforward application of the cumulative normal distribution illustrates that for 2.5% of the items, a larger deviation than 1.96 times the standard deviation is expected. When the assumption is being made that uncertainty hits were it hurts most, all the items in the pool can be ordered based on the maximum amount of information they provide for any ability value, and expected deviation is subtracted from the estimated information. This can be formulated as:

$$I_i^R(\theta) = I_i(\theta) - z_i * SD(I_i(\theta)), \quad i = 1, \ldots, I, \tag{15.1}$$

where i is the index of the item in the ordered bank, I is the number of items in the bank, $I_i^R(\theta)$ is the robust information provided at ability level θ, z_i corresponds to the $100 \cdot i/(I+1)$-th percentile of the cumulative normal distribution function, and $SD(I_i(\theta))$ is the standard deviation of the information function based on estimated item parameters. A comparable procedure can be applied in a Bayesian framework, however, to calculate z_i the posterior distribution has to be used.

15.3.2 Numerical Example to Illustrate the Concept of Robust Item Pools

An operational item pool of 306 items can be used to illustrate the effects of expected response function, or in our application, robust response function. The items can be calibrated with a three-parameter logistic model (3PLM):

$$P_i(\theta) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}, \tag{15.2}$$

where a is the discrimination, b is the difficulty, and c is the guessing parameter. BILOG MG 3 was applied to estimate all item parameters, based on a sample of 41,500 candidates. Besides the estimated item parameters, that ranged from $a \in [0.26, 1.40]$, $b \in [-3.15, 2.51]$, and $c \in [0.00, 0.50]$, the calibration error was also reported in terms of standard deviations (sd a = 0.02, sd b = 0.044, sd c = 0.016). These standard deviations are relatively small, but that was expected because of the large sample of candidates. The larger the sample, the smaller the calibration errors.

Based on the estimated item parameters, the maximum amount of information over all theta levels (Hambleton and Swaminathan 1985, p. 107) was calculated for all items in the pool, and they were ordered from large to small. The information provided by the 50 most informative items is shown in Fig. 15.2.

Robust information could be calculated by subtracting the expected deviation for all of the items using Eq. (15.1).

A small experiment can show the impact of robust item pools. First of all, the deviation between the information provided by each item and its robust counterpart were calculated. Besides, for each item, three simulated counterparts were created by drawing item parameters from the multivariate normal distribution with a mean equal to the estimated item parameters and standard deviations equal to the calibration error. In this way, three simulated item pools were created. Deviations between the information provided by each item and its simulated counterparts were calculated as well. These deviations are shown in Fig. 15.3.

Fig. 15.2 Maximum amount of information provided by the 50 most informative items

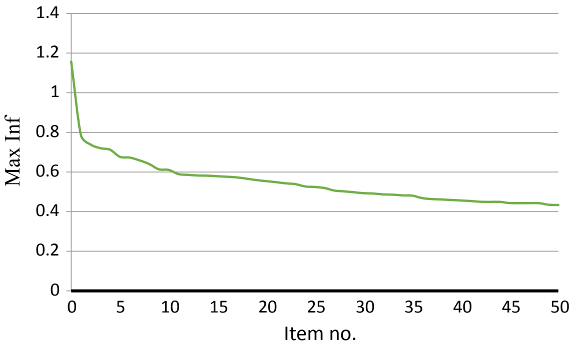
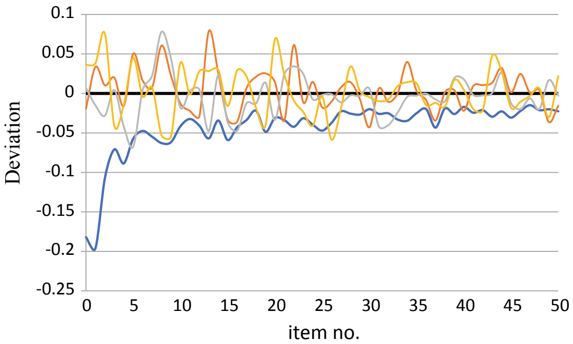


Fig. 15.3 Deviations from the maximum information for the robust information (thick line) and various simulated item banks (thin lines) for the 50 most informative items



From this experiment, several conclusions can be drawn. First of all, the robust counterparts provide less information that the original items. It should be noted however that the differences becomes smaller and smaller when the original items are less informative. Since the calibration errors differ for the various items, the deviance does not decrease monotonously. For the deviances between the original item and its simulated counterparts, it can be observed that the deviances are sometimes positive and sometimes negative. The most informative items had the largest calibration errors, therefore the largest deviations were observed for these items. Finally, it could be seen that simulated items could be less informative then their robust counterparts. The reason is that for most items in the middle of the ordered item pool, the robust counterparts are almost equal to the original items, even when the item parameters were estimated with considerable calibration error. This is in line with the observation that uncertainty was assumed to hit most for the most informative items.

15.3.3 Towards an Algorithm for Robust CAT

From this small experiment it can also be learned that for the 25 most informative items, the simulated items are more informative than their robust counterparts. In other words, the robust item information is still conservative. To deal with this conservatism, the Bertsimas and Sim method can be applied for item selection in robust CAT. This method assumes that uncertainty only affects the solution for at most Γ items in the test. The following pseudo-algorithm (Veldkamp 2012) describes the application of the Bertsimas and Sim method for selecting the gth item in CAT for a fixed length test of G items. It is a modified version or the original Bertsimas and Sim algorithm. In the first step, a robust item pool is used to calculate the conservative values d_i in the optimization model. Besides, the optimization model in Eqs. (15.3)–(15.6) has been formulated in such way that only the one item is selected that provides most information at the current ability estimate:

1. Calculate $d_i = I_i(\theta^{g-1}) - I_i^R(\theta^{g-1})$ for all items.

2. Rank the items such that $d_1 \geq d_2 \geq \ldots \geq d_n$
3. For $l = 1, \ldots, (G - (g - 1)) + 1$ find the item that solves:

$$G^l = \max\left\{\sum_{i=1}^{I} I_i(\hat{\theta}^{g-1})x_i - \left[\sum_{i=1}^{l}(d_i - d_l)x_i + \min(G - g, \Gamma)d_l\right]\right\} \tag{15.3}$$

subject to:

$$\sum_{i \in R^{g-1}} x_i = g - 1 \tag{15.4}$$

$$\sum_{i=1}^{I} x_i = g \tag{15.5}$$

$$x_i \in \{0, 1\} \quad i = 1, \ldots, I. \tag{15.6}$$

4. Let $l^* = \arg \max_{l=1,\ldots,n} G^l$.
5. Item g is the unadministered item in the solution of G^{l^*}.

In step 3 of the pseudo algorithm, (G-(g-1)) + 1 MIPs are solved, where (G-(g-1)) is the amount of items still to be selected. For the MIPs, it holds that x_i denotes whether item i is selected ($x_i = 1$) or not ($x_i = 0$) (see also Eq. [15.6]), and R^{g-1} is the set of items that have been administered in the previous $(g - 1)$ iterations. Equations (15.4)–(15.5) ensure that only one new item is selected. Finally, in (15.3) the amount of robust information in the test is maximized. This objective function consists of a part where the information is maximized and a part between square brackets that corrects for overestimation of the information. By solving (G-(g-1)) + 1 MIPs and choosing the maximum, a robust alternative for the test information that is not too conservative can be calculated. For details and proofs see Veldkamp (2013) and Bertsimas and Sim (2003).

15.4 Simulation Studies

To validate this algorithm for robust CAT, several simulation studies were conducted. The first study was conducted to illustrate the impact of item parameter uncertainty on CAT and to investigate whether robust item pools could reduce the effects. In the second study, the algorithm for robust CAT was studied. The Γ parameter, which indicates the number of items for which uncertainty is assumed to have impact on the resulting test, was varied to find out how this parameter influences the precision of the ability estimates. In the third study, the effects of five different methods for dealing with uncertainty in CAT were compared. First of all, we implemented Robust

CAT, where uncertainty in some of the items is assumed to impact ability estimation. The second method was more conservative. It is only based on the robust item pool, introduced in this chapter, where Fisher information of the items was corrected for expected uncertainty in the item parameters. In the second method, items were selected from the robust item pool. The third alternative was based on the work of Olea et al. (2012). They proposed to implement exposure control methods for dealing with uncertainty in the item parameters. Since exposure control methods limit the use of the most informative items, the use of the items with largest positive estimation errors will be limited as well. As a consequence, the impact of uncertainty in the item parameters on ability estimation will be neutralized. The fourth method combines Robust CAT and the exposure control method. Also because in practical testing situations, exposure control methods always have to be implemented to prevent that the most informative items in the pool become known (e.g. Sympson and Hetter 1985; van der Linden and Veldkamp 2004, 2007). Finally, the fifth alternative was to implement the Soyster (1973) method, where maximum values for the uncertainty for all the items was assumed. This method serves as a yardstick. It is very conservative, but takes all possible uncertainties into account.

15.4.1 Study 1

For the first simulation study, an item pool of 300 2PL-items was simulated, where the discrimination parameters a_i, $i = 1, \ldots, 300$, were randomly drawn according to $\log(a_i) \sim N\big(0, 0.3^2\big)$, and the difficulty parameters b_i, $i = 1, \ldots, 300$, were randomly drawn according to $b_i \sim N(0, 1)$. In this way, the true item parameters were simulated. Item parameter uncertainty was simulated by adding some random noise to these parameters according to $a_{ir} \sim N\big(a_i, (0.1)^2\big)$ and $b_{ir} \sim N\big(b_i, (0.3)^2\big)$. Test length was set equal to 20 items and items were selected based on maximum Fisher information. A number of 50,000 respondents were simulated for each $\theta \in \{-3, -2.75, \ldots, 3\}$. First, CATs were simulated based on the bank with uncertainty in the item parameters. Then, test information and RMSE were calculated based on the item parameters with uncertainty, true item parameters and based on robust item parameters.

15.4.2 Study 2

For the second study, the proposed method for robust CAT was implemented in R. Various settings of Γ were compared with the case where uncertainty was assumed to impact all items in the test. The item pool that was also used to illustrate the concept of robust item pools was applied. For this item pool, uncertainty in the parameter estimates was only small (average uncertainties in the parameters equal to $\Delta\text{a} = 0.02$, $\Delta\text{b} = 0.044$, $\Delta\text{c} = 0.016$). To calculate the robust item pool, expected information was calculated for all the items by taking only the uncertainty in the discrimination parameters into account (see also Veldkamp et al. 2013). In order to investigate the impact of the Γ parameter on CAT, uncertainty was assumed to impact the results for 25, 50, 75% and for all the items. This simulation study was much smaller than the first one. We simulated 1000 respondents for each of the ability values in the grid (−3, −2.5, …, 3). Test length was set equal to 20 items.

15.4.3 Study 3

In the third study, the five methods for dealing with uncertainty were compared with the Regular CAT, where uncertainty was not taken into account. In this study, also 1000 respondents were simulated for each of the ability values in the grid (−3, −2.5, …, 3). For the robust CAT method, Γ was set equal to 50% of the items. To study the impact of test length, the methods were compared for various test lengths. It varied from n = 5, n = 10, n = 20 to n = 40 items. In earlier studies on item selection in CAT (e.g. Matteucci and Veldkamp 2012) it turned out that differences between item selection methods only resulted in differences in ability estimates for short CATs with ten or fifteen items. The question remains whether the same findings hold for methods dealing with the impact of uncertainty on CAT.

15.4.4 Study Setup

Simulations for Study 1 were performed using dedicated software in C++, based on maximizing Fisher information and Warm's (1989) WLE estimator. Simulations for Study 2 and Study 3 were performed using the R software-package. The catR-package was used for implementing the CAT (Magis and Barrada 2017; Magis and Raîche 2012). In this package, several options are available. We applied the default settings with Bayes modal estimation, starting value equal to $\theta_0 = 0$ for all candidates, and a fixed test length as stopping criterion. The exposure control method in the package is based on Sympson-Hetter. To implement robust CAT in this package, we had to fix the number of items for which uncertainty had an impact in advance. In the robust CAT method, uncertainty in *at most Γ items* is assumed to impact the solution, but in the implementation, uncertainty in *exactly Γ items* was assumed to impact the solution. As a consequence, the robust CAT method became slightly more conservative.

15.5 Results

In Study 1, we compared CAT based on a robust item pool with CAT based on true item parameters and item parameters with uncertainty in them. Average test information functions are shown in Fig. 15.4.

The items in the robust item pool have been corrected for possible overestimation of the parameters. The resulting average information for CATs based on the robust item pool is lower than the information provided by CATs based on an item pool with uncertainty. In the middle of the ability distribution, the difference is only 2%, but towards the tails it is close to 10%. CATs were also simulated based on item parameters that were not disturbed by uncertainty. For these items it holds that they really provide most of their information when the theta estimated equals the difficulty parameter. Towards the tails of the distribution, there was quite a difference in average test information function. In the middle of the distribution, CATs are almost as or even more informative than CAT based on the disturbed item parameters. Root mean squared errors (RMSEs) for the various ability values are shown in Fig. 15.5.

Standard CAT, where uncertainty is not taken into account, resulted in an RMSE that is 6–17% higher than a CAT using the same item pool, but now with the item parameters assuming their true values. Thus, the efficiency of Standard CAT was

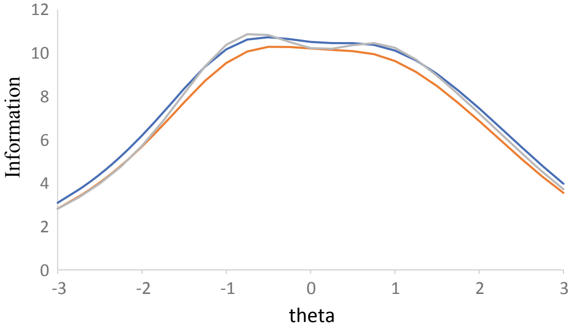


Fig. 15.4 Test information function for CAT with uncertainty in the item parameters (blue), robust item pool (orange) and based on real item parameters (green)

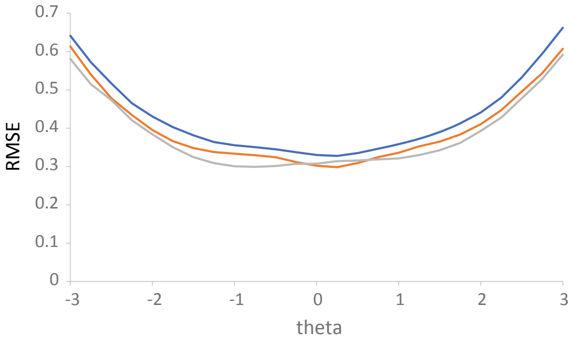


Fig. 15.5 RMSE for CAT with uncertainty in the item parameters (green), without uncertainty (blue) and based on robust item pool (blue)

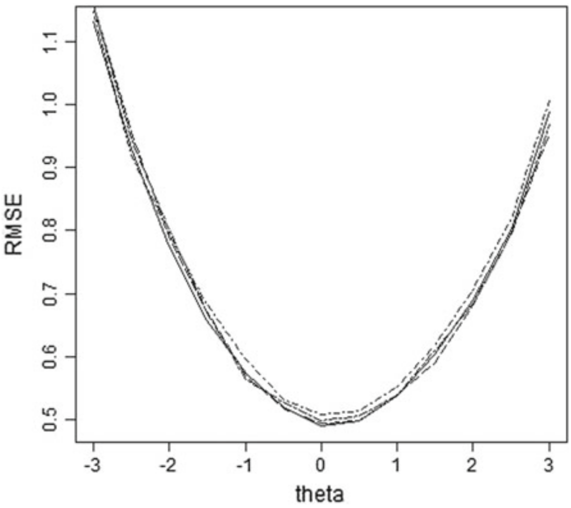


Fig. 15.6 RMSE for Robust CAT with Γ = 25% (solid line), Γ = 50% (large dashes), Γ = 75% (small dashes) and Γ = 100% (dash/dotted line) of the items

Table 15.1 Resulting average RMSEs for various methods for dealing with uncertainty for various test lengths (n)

Test length	n = 5	n = 10	n = 20	n = 40
Standard CAT	1.40	1.08	0.80	0.57
Robust CAT	1.40	1.09	0.82	0.59
Robust item pool	1.40	1.10	0.83	0.58
Exposure control	1.45	1.10	0.84	0.59
Robust item pool and exposure control	1.44	1.13	0.86	0.60
Soyster's method	1.49	1.15	0.88	0.63

overestimated by the same 6–17%. CAT based on robust item pools performed much better. RMSE was 5–9% higher than in simulations with perfectly known parameters, thus it overestimated the efficiency by 5–9%.

The second study focused on the method of robust CAT. In Fig. 15.6, the RMSE of the ability estimates is shown for various ability levels and various settings of Γ. The results for Γ = 25% (solid line) and Γ = 50% (large dashes) cannot be distinguished. For Γ = 75% (small dashes) the RMSE is slightly higher for abilities close to $\theta = 0$. For Γ = 100% (dash/dotted line), the RMSE is slightly higher for all ability levels. Overall, the differences in RMSE are very small.

The third study compared various methods for dealing with uncertainty in CAT. Impact of uncertainty was studied for various test lengths. Average RMSE was calculated over all ability values.

In Table 15.1, the results of various methods for dealing with uncertainty are shown for various test lengths.

Overall, it can be noticed that longer tests provide more information, and the differences between various methods in RMSE become smaller. Standard CAT resulted in the smallest RMSE. Robust CAT performed only slightly worse. Robust item pools performed almost comparable to robust CAT. Both methods based on exposure control performed slightly worse. As expected, the combination of robust item pools and exposure control performed even worse than the exposure control method. Finally, Soyster's method, which is very conservative by nature, performed the worst. Some small deviances of this general pattern were noted, but this might be due to the relatively small sample size in this study.

15.6 Conclusion

In this chapter, the outline of a procedure for robust CAT was presented as an answer to the problem over capitalization on uncertainty in the item parameters in CAT. In this method, a robust item pool based on expected Fisher information and the robust item selection method of Bertsimas and Sim (2003) are combined. First, it was demonstrated how robust item pools can be used in CAT. In a large simulation study, it was illustrated that robust item pools can be implemented successfully, and that the resulting CATs are much closer to the real values than standard CAT that does not take uncertainty in the item parameters into account. Figure 15.6 illustrates how various implementations of robust CAT provide different results. $\Gamma = 100\%$ of the items is equivalent to selecting all the items from the robust item pool, where the other values of Γ only select a percentage of the items from this pool. The impact of Γ on the RMSE turned out to be small, but for $\Gamma \leq 50\%$ of the items, the best results were obtained. An explanation for the small impact of Robust CAT might be found in the construction of the robust item pool and the nature of CAT. In the robust item pool, expected information is calculated based on the assumption of a normally distributed estimation error. Large adaptations of the provided information are only made for small number of items. As was illustrated in Fig. 15.3, differences between Fisher information and robust item information are only small for most of the items. On top of that, only a few items will be selected per candidate where the robust item information is really much smaller than Fisher information due to adaptive item selection. Larger differences might be found in case of larger estimation errors in the item pool.

The method of Robust CAT was also compared with other methods for dealing with uncertainty in the item parameters in CAT. Robust CAT generally provided the smallest RMSEs. Only applying robust item pools, performed almost as well. Besides, the exposure control method did not perform that much worse. More conservative methods like the combination of a robust item pool with exposure control and Soyster's method had larger RMSEs. It should be remarked however, that the differences are relatively small.

All of these results were based on averages over large numbers of replications. It might be interesting to see what happens at the individual level. The Robust CAT method was developed to prevent overestimation of the precision at the individual level as well. The exposure control method, on the other hand, does not take overestimation at the individual level into account. For example, when the maximum exposure rate of the items is set equal to $r_{max} = 0.2$, this implies that items with overestimated discrimination parameters will still be used for 20% of the candidates. Especially for small CATs with test length smaller than 20 items, the impact might be considerable. Further research will be needed to reveal for which percentage of the candidates is affected.

Finally, it needs to be mentioned that Belov and Armstrong (2005) proposed using an MCMC method for test assembly that imposes upper and lower bounds on the amount of information in the test. Since there is no maximization step in their approach, item selection is not affected by the capitalization on chance problem. On the other hand, this approach does not take uncertainty in the item parameters into account at all. This could lead to infeasibility problems (Huitzing et al. 2005), as illustrated in Veldkamp (2013). Besides, MCMC test assembly was developed for the assembly of linear test forms, and therefore application to CAT is not straightforward.

References

Belov, D. I., & Armstrong, D. H. (2005). Monte Carlo test assembly for item pool analysis and extension. *Applied Psychological Measurement, 29,* 239–261.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust optimization*. Princeton, NJ: Princeton University Press.

Bertsimas, D., & Sim, M. (2003). Robust discrete optimization and network flows. *Mathematical Programming, 98,* 49–71.

De Jong, M. G., Steenkamp, J.-B. G. M., & Veldkamp, B. P. (2009). A model for the construction of country-specific yet internationally comparable short-form marketing scales. *Marketing Science, 28,* 674–689.

Eggen, T. T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. (Unpublished doctoral thesis, Enschede).

Hambleton, R. H., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7,* 171–186.

Hambleton, R. H., & Swaminathan, H. (1985). *Item response theory, principles and applications*. Boston, MA: Kluwer Nijhoff Publishing.

Huitzing, H. A., Veldkamp, B. P., & Verschoor, A. J. (2005). Infeasibility in automated test assembly models: A comparison study of different methods. *Journal of Educational Measurement, 42,* 223–243.

Lewis, C. (1985). *Estimating individual abilities with imperfectly known item response functions.* Paper presented at the Annual Meeting of the Psychometric Society, Nashville, TN.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, 76*(1), 1–19.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software, 48*(8), 1–31.

Matteucci, M., & Veldkamp, B. P. (2012). On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. *Statistical Methods and Applications.* (Online First).

Mislevy, R. J., Wingersky, M. S., & Sheehan, K.M. (1994). *Dealing with uncertainty about item parameters: Expected response functions* (Research Report 94-28-ONR). Princeton, NJ: Educational Testing Service.

Olea, J., Barrada, J. R., Abad, F. J., Ponsoda, V., & Cuevas, L. (2012). Computerized adaptive testing: The capitalization on chance problem. *The Spanish Journal of Psychology, 15,* 424–441.

Soyster, A. L. (1973). Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research, 21,* 1154–1157.

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).

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371–390.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer Verlag.

van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13,* 35–53.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining Item exposure rates in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29,* 273–291.

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32,* 398–417.

Veldkamp, B. P. (2012). Ensuring the future of computerized adaptive testing. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 35–46).

Veldkamp, B. P. (2013). Application of robust optimization to automated test assembly. *Annals of Operations Research, 206*(1), 595–610.

Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement, 37*(2), 123–139.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 16
On-the-Fly Calibration in Computerized Adaptive Testing

Angela Verschoor, Stéphanie Berger, Urs Moser and Frans Kleintjes

Abstract Research on Computerized Adaptive Testing (CAT) has been rooted in a long tradition. Yet, current operational requirements for CATs make the production a relatively expensive and time consuming process. Item pools need a large number of items, each calibrated with a large degree of accuracy. Using on-the-fly calibration might be the answer to reduce the operational demands for the production of a CAT. As calibration is to take place in real time, a fast and simple calibration method is needed. Three methods will be considered: Elo chess ratings, Joint Maximum Likelihood (JML), and Marginal Maximum Likelihood (MML). MML is the most time consuming method, but the only known method to give unbiased parameter estimates when calibrating CAT data. JML gives biased estimates although it is faster than MML, while the updating of Elo ratings is known to be even less time consuming. Although JML would meet operational requirements for running a CAT regarding computational performance, its bias and its inability to estimate parameters for perfect or zero scores makes it unsuitable as a strategy for on-the-fly calibration. In this chapter, we propose a combination of Elo rating and JML as a strategy that meets the operational requirements for running a CAT. The Elo rating is used in the very beginning of the administration to ensure that the JML estimation procedure is converging for all answer patterns. Modelling the bias in JML with the help of a relatively small, but representative set of calibrated items is proposed to eliminate the bias.

A. Verschoor (✉) · F. Kleintjes
Cito, P.O. Box 1034, 6801 MG Arnhem, Netherlands
e-mail: angela.verschoor@cito.nl

F. Kleintjes
e-mail: frans.kleintjes@cito.nl

S. Berger · U. Moser
Institute for Educational Evaluation, Associated Institute of the University of Zurich,
Wilfriedstrasse 15, 8032 Zürich, Switzerland
e-mail: stephanie.berger@ibe.uzh.ch

U. Moser
e-mail: urs.moser@ibe.uzh.ch

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_16

16.1 Introduction

In recent years a comprehensive body of research on computerized adaptive testing (CAT) has been performed. Kingsbury and Zara (1989) refined item selection with a method for content control. Sympson and Hetter (1985) dealt with overexposure, while Revuelta and Ponsoda (1998) treated the subject of underexposure. Eggen and Verschoor (2006) proposed a method for difficulty control.

Still, the process to develop an operational CAT is an expensive and time consuming task. Many items need to be developed and pretested with relatively high accuracy. Data collection from a large sample of test takers is a complex task, while it is crucial to be able to calibrate these test items efficiently and economically (Stout et al. 2003). Item calibration is especially challenging when developing a CAT targeted to multiple school grades (Tomasik et al. 2018), or when the target population is small and participation low. Several rounds of pretesting may be necessary in those cases to arrive at accurate estimates of item characteristics. Unfortunately, a threat of pretesting is that motivational effects can cause disturbances in the estimation process (e.g., Mittelhaëuser et al. 2015).

There are enough reasons to look for alternative procedures that can be employed to diminish the burden of pretesting for use in digital testing, especially in CAT. In this chapter, strategies for on-line calibration that can replace pretesting entirely, or at least to a certain extent, are evaluated. Ideas of extending a calibrated item pool are not new. They range from extending an existing item pool with a limited number of new items – replenishment – to periodic calibration of all items and persons in real time: *on-the-fly calibration*. Until recently the latter possibilities were limited by computational potential and infrastructure. But in view of the increased demand for CAT, on-the-fly calibration is drawing attention.

16.1.1 Replenishment Strategies and On-the-Fly Calibration

Replenishment strategies were developed in order to extend existing item pools or replace outdated items. Thus, a precondition for replenishment strategies is that a majority of items have been previously calibrated. Only a relatively small portion of items can be pretested. Test takers will usually take both operational and pretest items, carefully merged, so that they cannot distinguish between the two types. The responses on the pretest items often are not used to estimate the ability of the test takers. This procedure is referred to as seeding. Usually, but not necessarily, responses for the seeding items are collected and only when a certain minimum number of observations has been reached, the item pool will be calibrated off-line. In on-line methods, the seeding items are calibrated regularly in order to optimize the assignment of the seeding items to the test takers, using provisional parameter estimates of those seeding items.

Stocking (1988) was the first to investigate on-line calibration methods, computing a maximum likelihood estimate of a test taker's ability using the item responses and parameter estimates of the operational items. Wainer and Mislevy (1990) described an on-line MML estimation procedure with one EM iteration (OEM) for calibrating seeding items. All these methods can be applied to replenish the item pool, that is, to add a relatively small set of new items to the pool in order to expand the pool or to replace outdated items.

Makransky (2009) took a different approach and started from the assumption that no operational items are available yet. He investigated strategies to transition from a phase in which all items are selected randomly to phases in which items are selected according to CAT item selection algorithms. Fink et al. (2018) took this approach a step further by postponing the calibration until the end of an assessment period and delaying reporting ability estimates, so that the method is suitable for high stakes testing for specialist target populations like university examinations.

Usually, however, an operational CAT project is not started in isolation. Frequently, several items are being reused from other assessments or reporting must take place in relation to an existing scale. In those circumstances, data from these calibrated items and scales may be used as a starting point to collect data for the CAT under consideration. Using this approach it is interesting to investigate how many and which type of calibrated items are required to serve as reference items. Furthermore, we assume a situation in which instant parameter updating and reporting is a crucial element, and thus the time needed for calibration should be kept at a pre-specified minimum.

16.1.2 On-the-Fly Calibration Methods

In this chapter we concentrate on the 1PL or Rasch model (Rasch 1960) for which the probability of giving a correct answer by a test taker with ability parameter θ on an item with difficulty parameter β is given by

$$P(X = 1|\theta) = \frac{e^{(\theta-\beta)}}{(1 + e^{(\theta-\beta)})}. \tag{16.1}$$

For item calibration, we consider three methods:

- a rating scheme by Elo (1978) which has been adopted by the chess federations FIDE and USCF in 1960;
- the JML procedure described by Birnbaum (1968);
- the MML method proposed by Bock and Aitkin (1981).

16.1.2.1 Elo Rating

According to the Elo rating system, player A with rating r_A has an expected probability of

$$E_{AB} = \frac{10^{(r_A - r_B)/400}}{1 + 10^{(r_A - r_B)/400}} \tag{16.2}$$

of winning a chess game against player B with rating r_B. After the game has been played, the ratings of the players are updated according to

$$r_A^{'} = r_A + K(S - E_{AB}) \tag{16.3}$$

and

$$r_B^{'} = r_B - K(S - E_{AB}), \tag{16.4}$$

where S is the observed outcome of the game and K a scaling factor. From Equations (16.2)–(16.4) it can be seen that Elo updates can be regarded as a calibration under the Rasch model, albeit not one based on maximization of the likelihood function. Several variations exist, especially since K has been only loosely defined as a function decreasing in the number of observations. Brinkhuis and Maris (2009) have shown, however, that conditions exist under which the parameters assume a stationary distribution.

It is clear that the Elo rating scheme is computationally very simple and fast, and therefore ideally suited for instantaneous on-line calibration methods. However, little is known about the statistical properties of the method, such as the rate of convergence, or even if the method is capable at all of recovering the parameters at an acceptable accuracy. The Elo rating method is widely used in situations in which the parameters may change rapidly during the collection of responses, such as in sports and games. The method has been applied in the Oefenweb (2009), an educational setting in which students exercise frequently within a gaming environment.

16.1.2.2 JML

JML maximizes the likelihood L of a data matrix, where N test takers have each responded to n_j items with item scores x_{ij} and sum score s_j, where P_j is the probability of a correct response and $Q_j = 1 - P_j$ refers to the probability of an incorrect response. The likelihood function can be formulated as

$$L = \prod_{j=1}^{N} \frac{n_j!}{s_j!(n_j - s_j)!} P_j^{s_j} Q_j^{n_j - s_j}. \tag{16.5}$$

By maximizing the likelihood with respect to a single parameter, while fixing all other parameters, and rotating this scheme over all parameters in turn, the idea is that the global maximum for the likelihood function will be reached. In order to maximize the likelihood, a Newton-Raphson procedure is followed that takes the form

$$\beta_i^{t+1} = \beta_i^t - \frac{\sum_{j=1}^{N} P_{ij} - x_{ij}}{\sum_{j=1}^{N} P_{ij} Q_{ij}} \tag{16.6}$$

for updating item parameter β_i, and

$$\theta_j^{t+1} = \theta_j^t + \frac{\sum_{i=1}^{n_j} P_{ij} - x_{ij}}{\sum_{i=1}^{n_j} P_{ij} Q_{ij}} \tag{16.7}$$

for updating person parameter θ_j.

Like the Elo rating, JML is a simple and fast-to-compute calibration method. However, there are a few downsides. In the JML method, the item parameters are structural parameters. The number of them remains constant when more observations are acquired. The person parameters, on the other hand, are incidental parameters, whose numbers increase with sample size. Neyman and Scott (1948) showed in their paradox that, when structural and incidental parameters are estimated simultaneously, the estimates of the structural parameters need not be consistent when sample size increases. This implies that when the number of observations per item grows to infinity, the item parameter estimates are not guaranteed to converge to their true values. In practical situations, JML might still be useful as effects of non-convergence may become apparent only with extreme numbers of observations per item. The biggest issue with the use of JML, however, is a violation of the assumption regarding the ignorability of missing data. Eggen (2000) has shown that under a regime of item selection based on ability estimates, a systematic error will be built up. He has shown that only for MML, violation of the ignorability assumption has no effect.

16.1.2.3 MML

MML is based on the assumption that the test takers form a random sample from a population whose ability is distributed according to density function $g(\theta|\tau)$ with parameter vector τ. The essence of MML is integration over the ability distribution, while a sample of test takers is being used for estimation of the distribution parameters. The likelihood function

$$L = \prod_{j=1}^{N} \int P(x_j|\theta, \xi, \tau) g(\theta_j|\tau) d\theta_j \tag{16.8}$$

is maximized using the expectation-maximization algorithm (EM). ξ is the vector of item parameters here. EM is an iterative algorithm for maximizing a likelihood function for models with unobserved random variables. Each iteration consists of an

E-step in which the expectation of the unobserved data for the entire population is calculated, and an M-step in which the parameters are estimated that maximize the likelihood for this expectation. The main advantage of the MML calibration method is that violation of the ignorability assumption does not incur biased estimators. A disadvantage, however, is that MML is usually more time consuming than the Elo rating or JML, and thus it might be too slow to be employed in situations where an instantaneous update is needed.

In order to replenish the item pool, Ban et al. (2001) proposed the MEM algorithm, where in the first iteration of the EM algorithm, the parameters of the ability distribution for the population are estimated through the operational items only. However, if the number of operational items is very small, this procedure results in highly inaccurate estimates. For the problem at hand, where not many operational items are available, the MEM method is not applicable, and the regular MML might be considered.

16.1.3 The Use of Reference Items in Modelling Bias

Although bias cannot be avoided when calibrating CAT data with JML, previously collected responses for a part of the items can be used to model this bias. We assume that the bias is linear, i.e. that by applying a linear transformation $\hat{\beta}_i' = a\hat{\beta}_i + b$, estimators $\hat{\beta}_i'$ eliminate the bias. When JML is used to calibrate all items in the pool, the reference items, whose previously estimated parameters we trust, can be used to estimate transformation coefficients a and b. Let μ_r and σ_r be the mean and standard deviation of the previously estimated parameters of the reference items, while μ_c and σ_c are the mean and standard deviations of the parameter estimates for the same items, but now taken from the current JML-calibration, then the transformation coefficients are $a = \frac{\sigma_c}{\sigma_r}$ and $b = \frac{\mu_c - \mu_r}{\sigma_r}$. All reference items retain their trusted parameter values, while all new items in the calibration are updated to their transformed estimates.

Even though MML calibrations do not incur bias when applied to CAT data, it is technically possible to follow a similar procedure, but this should not result in substantially improved parameter estimates.

16.1.4 The Need for Underexposure Control

A solution must be found for a complication that will generally only occur in the very first phases when few observations are available. For some response patterns, the associated estimates of the JML and MML methods assume extreme values. In the case of perfect and zero scores, the estimations will even be plus or minus infinity. On the other hand, items are usually selected according to the maximization of Fisher information. This means that once an item has assumed an extreme parameter value, it will only be selected for test takers whose ability estimation is likewise extreme. And thus, those items tend to be selected very rarely. If the parameter estimation is based on many observations, this is fully justified. But if this situation arises when only a few observations have been collected, these items will effectively be excluded from the item pool without due cause.

In previous studies, this situation was prevented by defining various phases in an operational CAT project, starting with either linear test forms (Fink et al. 2018) or random item selection from the pool (Makransky 2009). When the number of observations allowed for a more adaptive approach, a manual decision regarding transition to the next phase was taken. Since we assume a situation in which decisions are to be taken on the fly and have to be implemented instantaneously, manual decisions are to be avoided. Therefore, a rigid system to prevent underexposure must be present so that items with extreme estimations but with very few observations will remain to be selected frequently. Similar to the findings of Veldkamp et al. (2010), we propose an underexposure control scheme based on eligibility function

$$f(n_i) = \begin{cases} X - n_i(X-1)/M, & n_i < M \\ 1, & \text{else} \end{cases} \tag{16.9}$$

where n_i is the number of observations for item i, M is the maximum number of observations for which the underexposure control should be active, and X is the advantage that an item without any observations gets over an item with more than M observations. This means effectively that there is no transition from different phases for the entire item pool, but each transition will take place on the level of individual items.

Overexposure may further improve the estimation procedure, but since items with only few observations form a larger impediment to the cooperation between estimation procedure and item selection than a loss of efficiency caused by overexposure, overexposure control is out of scope for this study.

16.1.5 A Combination of Calibration Methods

Although Joint Maximum Likelihood (JML) calibration would meet operational requirements for running a CAT in terms of computational performance, its bias and its inability to estimate parameters for perfect or zero scores makes it unsuitable as a strategy for on-the-fly calibration. To overcome this issue, we propose a combination of Elo rating and JML as a strategy that meets the operational requirements for running a CAT. The Elo rating is used in the very beginning of the administration to ensure that the JML estimation procedure is converging for all answer patterns. Modelling the bias in JML with the help of a relatively small, but representative set of calibrated items, is proposed to eliminate the bias in JML. Although Marginal Maximum Likelihood (MML) estimation is not known to give biased estimates, it may turn out that this method also benefits from modelling the bias and starting with an Elo rating scheme.

Next to the use of well-calibrated reference items there is the need for underexposure control to ensure that we will collect an acceptable minimum of observations for all items to make calibration feasible.

16.2 Research Questions

Research on on-the-fly item calibration in an operational setting for use in a CAT has different aspects. The operational setting requires the process to be accurate enough and quick enough. The main research question is: Does the proposed approach of a combination of Elo and JML yield sufficient accuracy for both item and persons parameters, under operational conditions? If the research question can be positively answered, there is no need for a large calibrated item pool for a CAT implementation. New items will be embedded in a set of reference items, while instantaneous on-the-fly calibration enables us to base test takers' ability estimates on both new and reference items. At the same time, the use of reference items help to improve item parameter estimations.

Elaborating on the main research question, we formulate the following research sub questions:

- Can we model bias in JML by means of reference items? How do we select the reference items in modelling the bias in JML in relation to their scale? How many reference items are required?
- How well does the proposed calibration method recover item and person parameters? Can we give guidelines when to switch from Elo to JML?
- How well does the proposed method perform in terms of computational time?

Two simulation studies were conducted. The first study concentrated on the investigation of the use of reference items in the elimination of bias. The second study concentrated on a comparison of the methods with respect to their ability to recover the item and person parameters, in relation to the computation time needed.

16.3 Simulation Studies

All simulations were based on the same item pool consisting of 300 items, 100 of which had "true" parameters drawn randomly from a standard normal distribution, supplemented by 200 items drawn from a uniform distribution from the interval (−3.5, …, 3.5). Furthermore, all simulated students were drawn from a standard normal distribution and took a test with fixed length of 20 items. Simulated students were assigned to the CAT one after another. The first item for each simulated student was drawn randomly from the entire item pool. For selecting the remaining 19 items, the weighted maximum likelihood estimator (WMLE) by Warm (1989) was used for the provisional ability, while items were selected that had the highest eligibility function value as given in (16.9), multiplied by the Fisher information at the provisional ability. At the start of each run, all item parameter estimates were set to zero, apart from the reference items that retained their trusted value. After each simulated student had finished the test, the item pool was calibrated. Two criteria were used to evaluate the measurement precision of the calibration methods: average bias and root mean square error (RMSE) of the item and person parameter estimators.

The Elo rating was used in the very beginning of the simulation runs. Using the Elo rating, in combination with the underexposure control, effectively prevented the situation that extreme estimations occurred and that items were thus no longer selected. All items that had a minimum of 8 observations and had a response vector consisting of both correct and incorrect answers were calibrated with JML or with MML. At this moment, extreme parameter estimates purely due to chance begin to become somewhat unlikely. For the items not having this threshold yet, the Elo rating was retained.

16.3.1 Use of Reference Items in Elimination of Bias

To validate the reference items approach with JML, we compared the results from JML with and without reference items. After sorting the item pool in ascending difficulty, the reference items were chosen to be evenly spread in an appropriate difficulty range. It may be assumed that such an evenly spread reference set will yield more accurate estimations for the transformation coefficients than, for example, a normal distribution. Items with more extreme difficulties, however, may be selected less frequently and less optimally than items with moderate difficulty. This difference in selection may well influence the extent to which the reference items model the bias. Therefore, two different sets of reference items were simulated in order to investigate which set is able to eliminate the bias best. The first set comprised 20 items with parameters from the interval (−2.5, …, 2.5), while in the second set, 20 items from the full range were selected. The item parameters in both reference sets were evenly spread. In summary, we investigated three different conditions in the first simulation study: Two different ranges of reference items (limited vs. full range), while the third condition was the application of JML without reference items. Each condition was repeated 1000 times and each of these runs consisted of 3000 simulated students. This coincided with the moment that the average number of observations is 200 per item, which is a reasonable choice for calibrations under the 1PL model. Furthermore, bias for each item, averaged over the simulation runs, was investigated when 600 simulated students had finished their tests. This moment coincides with the moment that the average number of observations is 40. This point can be seen as being in a transition from the situation in which no information at all is available to the situation in which calibrations under the 1PL model are starting to become stable.

16.3.1.1 Results

Figure 16.1 shows two scatter plots of $\hat{\beta}$ against β at the end of a single run, i.e. when each item had on average 200 observations. The left scatter plot depicts the estimates for the JML, while the right one shows the estimates in case of JML with the limited set of reference items. It shows a typical situation that can be observed in all simulation runs. It can be seen that a linear transformation based on a set of reference items is effective in reducing the bias, even if it can be surmised that the reduction is not perfect for items with extreme parameter values. Nevertheless, the JML without reference items shows bias to such a large extent that it has not been considered any further in the remainder of the simulations.

In Fig. 16.2, the bias in $\hat{\beta}$ is given. The bias for the limited set of reference items is presented in the left graphs, in the right graphs the bias for the full set. The top row shows the bias for each item, averaged over the simulation runs, when 600 simulated students have finished their test, the bottom row at the end of the runs. When an average of 40 observations per item have been collected, the exclusion of extreme items in the reference set lead to a pronounced overcompensation, while using items from the full difficulty range resulted in overcompensation only at the extremes. The situation at an average of 200 observations per item is different: using reference items from the full difficulty range lead to a clear outward bias for the medium items and an inward bias for the extreme items. In the early phases of the simulations, a linear transformation based on the full set gave the best results, but overcompensation caused a situation in which the error in parameters gradually increased. As it may be surmised that items with medium difficulty will be selected more frequently than items with extreme difficulty, it can be expected that the ability estimates will be more accurate when extreme items will be excluded from being reference items. A second

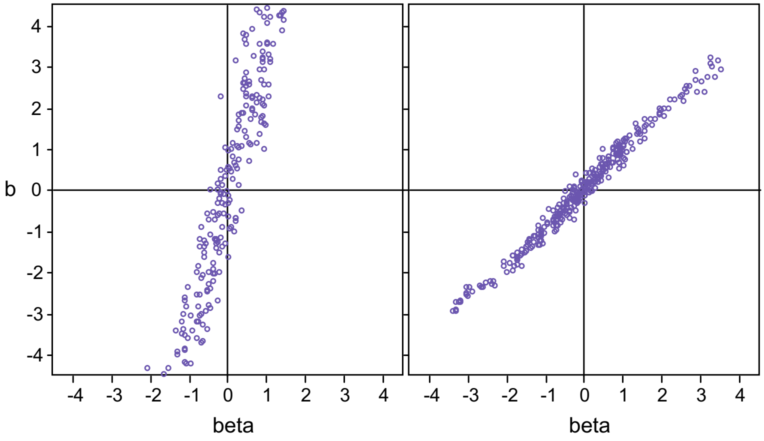


Fig. 16.1 Scatter plot of $\hat{\beta}$ against β for JML without (left) and JML with reference items (right – limited set) at an average of 200 observations per item

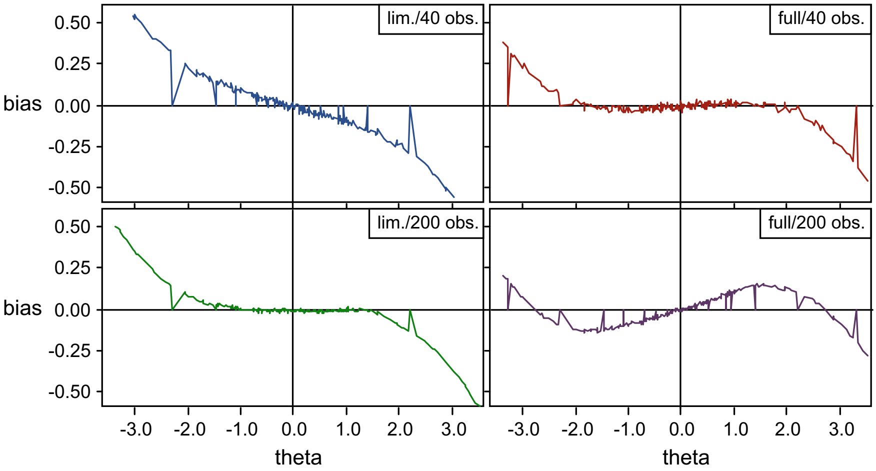


Fig. 16.2 Bias of $\hat{\beta}$ for JML with two different sets of reference items at an average of 40 and 200 observations per item

disadvantage for using extreme items in the reference set is the overcompensation as it can be seen when a considerable number of observations has been collected. Therefore, only JML with a set of reference items selected from a limited difficulty range was considered in the second study. This condition is from here on referred to plainly as JML.

16.3.2 Comparison of the Methods

The settings of the second set of simulations were similar to those of the first set. Like the first set of simulations, we evaluated the bias for person and item parameters after 3000 simulated student, thus making a direct comparison possible. As an extension, we reported the RMSE for the person parameter after each simulated student, as well as the average of the RMSEs for all item parameters. In order to investigate convergence of the methods beyond an average of 200 observations per item, each run consisted of 15,000 simulated students instead of 3000, thus extending the simulation to an average of 1000 observations per item. The conditions that were investigated were:

- Elo rating
- JML
- MML
- MML with reference items.

16.3.2.1 Results

Figure 16.3 shows the bias of $\hat{\beta}$ for the four different simulation conditions, taken after 3000 simulated students. The top left graph shows the results for the Elo method, the top right graph for JML, and bottom left for MML, while in the bottom right graph the results for MML with reference items are presented.

As can be inferred from Eqs. (16.3) and (16.4), the Elo rating update scheme is an effective method in preventing extreme estimations in cases of perfect and zero scores, but in the long run this appears to become a disadvantage. The Elo rating shows a large inward bias in Fig. 16.3, while MML shows an outward bias. The two conditions using reference items showed the best results. JML yielded an increasing inward bias for extreme item parameters, and the use of reference items in the case of MML compensated the bias almost completely.

Recovery of person parameters is equally important as the recovery of item parameters, if not more important. The purpose of a test usually is an evaluation of test takers, item calibration is merely a means to achieve this. Figure 16.4 shows the bias in $\hat{\theta}$. The bias in $\hat{\theta}$ reflected the situation for $\hat{\beta}$: The Elo rating showed substantial inward bias, MML an outward bias, while the use of reference items reduced bias also in the person parameter estimates. In the case of JML, the extreme ability ranges showed some inward bias while a small overcompensation took place for MML.

Figure 16.5 illustrates the development of RMSE($\hat{\beta}$) during the simulation runs, now averaged over all items. As can be seen, the average RMSE($\hat{\beta}$) rapidly diminished in the first phases of the process. The Elo rating, combined with the underexposure control, effectively prevented estimations with extreme values when the first simulated students took their tests. The disadvantage of the Elo rating becomes

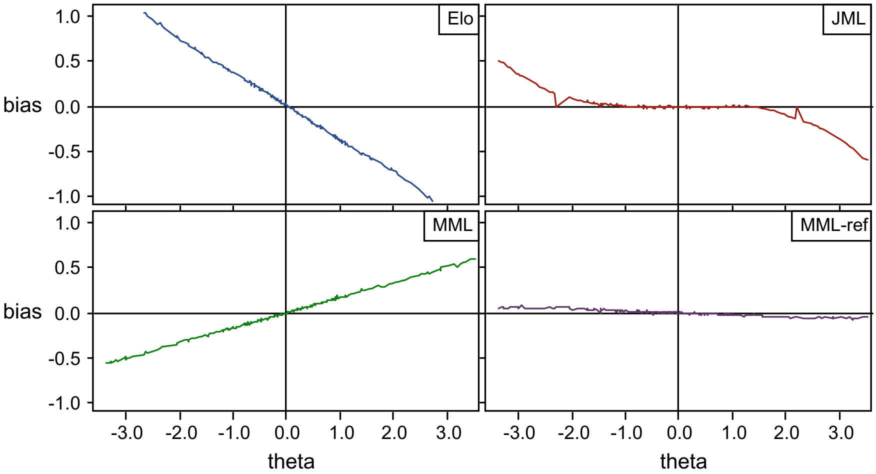


Fig. 16.3 Bias of $\hat{\beta}$ for Elo, JML, MML and MML with reference items at an average of 200 observations per item

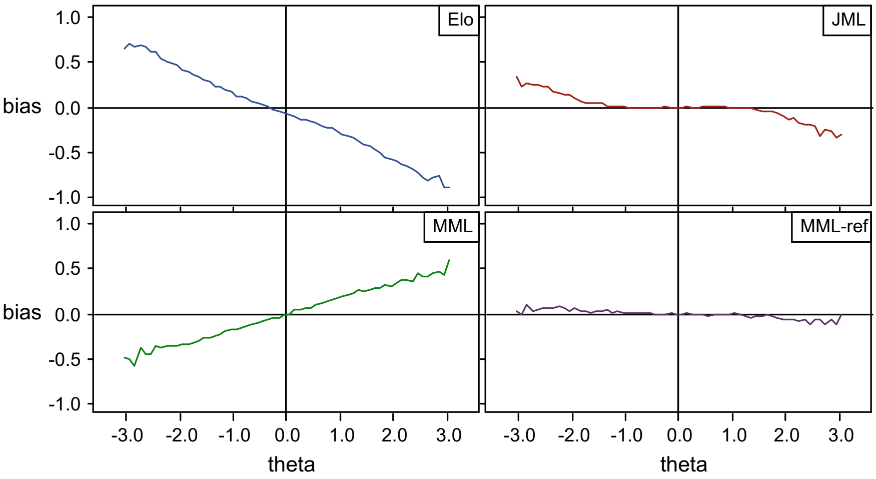


Fig. 16.4 Bias of $\hat{\theta}$ for Elo, JML, MML and MML with reference items at an average of 200 observations per item

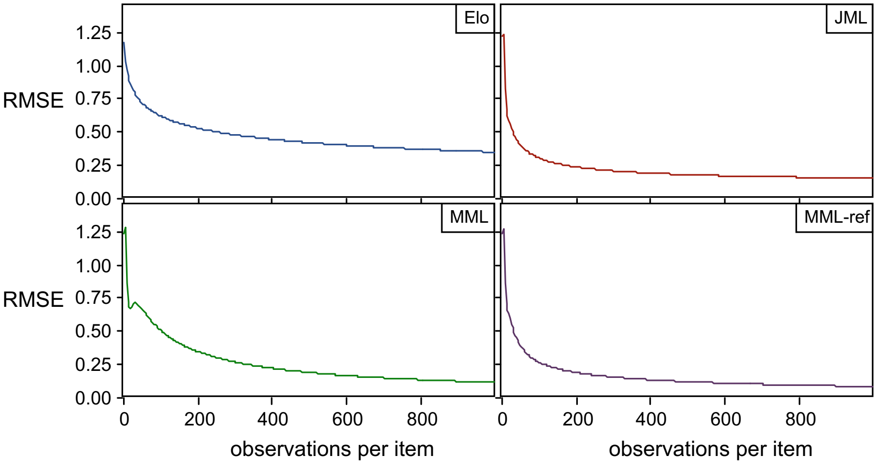


Fig. 16.5 Average RMSE($\hat{\beta}$) at increasing numbers of observations

apparent later on in the runs: The Elo rating is slowest to converge of the methods under consideration. At an average of 1000 observations per item, RMSE($\hat{\beta}$) was on average 0.351. This situation was reached at an average of approximately 70 observations per item for JML and 60 for MML with reference items. JML converged faster in the initial phase, but with increasing numbers of observations, MML started to outperform JML. MML with reference items had a clear advantage over JML from 50 observations per item. The average RMSE($\hat{\beta}$) was 0.383 at that moment.

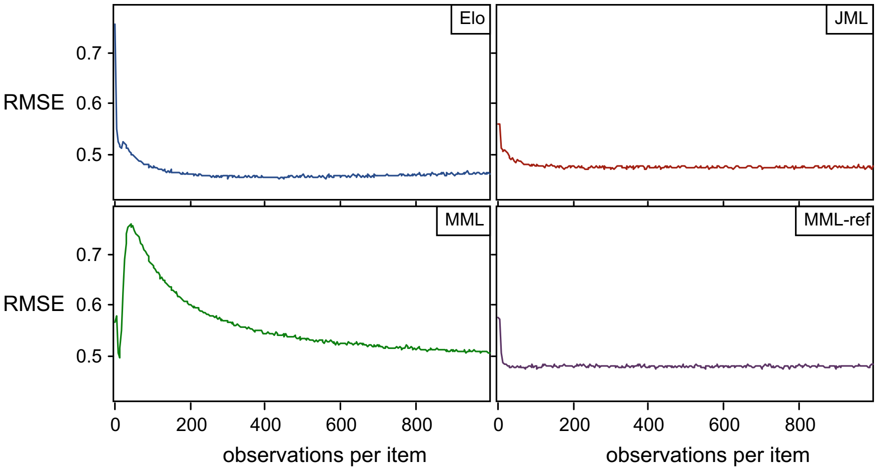


Fig. 16.6 Average RMSE($\hat{\theta}$) at increasing numbers of observations

Without using reference items, MML outperformed JML from approximately 500 observations per item, when the average RMSE($\hat{\beta}$) was 0.191.

Figure 16.6 presents a different situation. It shows the development of average RMSE($\hat{\theta}$). In all conditions, apart from the case of MML without reference items, average RMSE($\hat{\theta}$) decreased rapidly. At approximately 50 observations per item, average RMSE($\hat{\theta}$) decreased below 0.50 for the Elo ratings. This value was reached at approximately 30 observations per item in the JML and MML with reference items. In the MML condition without reference items, the lowest value for average RMSE($\hat{\theta}$) was 0.51 when an average of 1000 observations per item were collected. As a baseline, a simulation was conducted where all item parameters assumed their true value in the item selection algorithm and thus no calibration took place. In this situation, average RMSE($\hat{\theta}$) reached a value of 0.48. This value was reached only in the Elo rating condition, albeit because $\hat{\theta}$ showed a large inward bias.

16.3.2.2 Computational Performance

One important issue for on-the-fly calibration in operational settings still remains open: computational performance. Ideally, a calibration method gives highly accurate estimations almost instantly, but in case that such an ideal is not reached, a compromise between accuracy and speed must be made. Unfortunately, in operational settings speed depends not only on the method to be used, but also on other factors like system architecture and interfacing between various software modules. Therefore a complete evaluation of the compromise between accuracy and speed falls outside the scope of this chapter. Nevertheless, measurements of computer times during the simulations do give an indication of the relative performance of the calibration methods under consideration.

The Elo rating method is clearly the fastest method, since it makes use of only the responses of one test taker in each cycle and does not consider the entire data matrix. Running on a laptop equipped with an i5-5Y57 CPU at 1.6 GHz, one Elo update cycle typically used less than 0.1 ms, regardless of the number of previous responses on the respective items. On the other hand, JML and MML were substantially more computation intensive and showed different time complexity. After 300 simulated students, that is, when on average 20 observations were collected, the average time for one JML calibration cycle took approximately 800 ms. The MML calibration took on average approximately 2.3 s to complete. At 600 simulees, the computation times had increased to 1.6 and 3.8 s for JML and MML, respectively. The computation time increased roughly linearly to the end of the simulation runs, when each JML cycle needed an average of 42 s and a single MML cycle used approximately 140 s to complete. During all phases of the runs, the MML showed negligible differences in computation time between the condition with and without reference items.

16.4 Discussion

From a measurement perspective, MML clearly outperforms JML and Elo ratings, while the use of reference items has an advantage in stabilizing the scale in the early phases of data collection for an item pool while in operation. In the case of JML, the use of reference items should always be considered. But also in the case of MML, reference items are an effective means to eliminate bias. Computational performance, however, heavily depends on external factors such as system architecture and cannot be evaluated in a simulation study as we have conducted. What is clear, though, is that the Elo ratings are extremely fast to compute so that they can be employed in synchronous processes immediately after receiving a response. This is a reason why many large-scale gaming and sports environments use some variant of this method. Furthermore, it is a very useful method in the beginning of data collection processes, when the number of observations is still very low. On the other hand, JML and MML are considerably slower, so that it can be expected that employment of those methods will be deemed feasible only in asynchronous processes. As a consequence, the update of parameters will not be instantaneous but will take a few seconds or more to be effective. In a setting with many thousands of test takers taking the test concurrently, this would be an unwanted situation and the Elo rating system would probably be the method of choice.

When more accurate estimations than those provided by the Elo rating are needed, MML can be used. If reference items are available, JML is an option as well, whereby it could be expected that MML with reference items produce more accurate estimations at the cost of a somewhat heavier computational burden as compared to JML. In both cases it can be expected that item parameter estimations will have stabilized when an average of approximately 200 observations per item have been collected.

At this stage, each consecutive calibration will not change parameter values substantially. Thus, it might be wise not to calibrate the items after each test taker, but to use a less intensive routine. Such a regime could well be an off-line calibration, whereby, for example, item fit statistics can be inspected and actions undertaken when items appear to be defective.

Two important questions were left unanswered in this study: “How many reference items are needed?” and “How many new items can be added?”. It is obvious that the answer to both questions depend on factors such as model fit and accuracy of the reference item estimates. As all responses were simulated under ideal circumstances, that is, all items showed a perfect model fit and all reference items were modelled as having no error, we assume that using two reference items would have been enough to accurately estimate the transformation coefficients, resulting in findings similar to the ones reported here.

The first simulation study, validating the reference items approach, showed that a linear transformation does not completely remove the bias incurred by violation of the ignorability assumptions. A more advanced model of the bias involving a non-linear transformation would probably improve the parameter recovery for the JML calibration, potentially up to the level currently seen with MML using reference items.

Based on findings from the simulation studies, we suggest to apply the following strategy in an on-the-fly calibration in CAT: The entire process of data collection and on-the-fly calibration in an operational CAT setting should start with an Elo rating update scheme in order to avoid extreme parameter estimates, combined with a rather rigorous underexposure control scheme. After this initial phase, a switch to JML using reference items should be made. When increased accuracy is worth the additional computational burden of using MML, a new switch could be made and the on-the-fly calibration can either be slowed down or switched off entirely when parameters appear to be stabilized.

References

Ban, J., Hanson, B., Wang, T., Yi, Q., & Harris, D. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, *38*, 191–212. https://doi.org/10.1111/j.1745-3984.2001.tb01123.x.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental scores*. Reading, MA: Addison-Wesley.

Bock, R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02293801.

Brinkhuis, M. & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems*. Technical Report MRD 2009-1, Arnhem: Cito.

Eggen, T. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika*, *65*, 337–362. https://doi.org/10.1007/BF02296150.

Eggen, T., & Verschoor, A. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, *30*, 379–393. https://doi.org/10.1177/0146621606288890.

Elo, A. (1978). *The rating of chess Players, past and present*. London: B.T. Batsford Ltd.

Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, *60*, 327–346.

Kingsbury, G., & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375. https://doi.org/10.1207/s15324818ame0204_6.

Makransky, G. (2009). An automatic online calibration design in adaptive testing. In D. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*: Reston, VA: Graduate Management Admission Council.

Mittelhaëuser, M., Béguin, A., & Sijtsma, K. (2015). The effect of differential motivation on irt linking. *Journal of Educational Measurement*, *52*, 339–358. https://doi.org/10.1111/jedm.12080.

Neyman, J., & Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32. https://doi.org/10.2307/1914288.

Oefenweb B.V. (2009). Math Garden [Computer Software].

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danmarks Pædagogiske Institut.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 311–327. https://doi.org/10.1111/j.1745-3984.1998.tb00541.x.

Stocking, M. (1988). *Scale drift in on-line calibration*. Research Report (pp. 88–28). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00284.x.

Stout, W., Ackerman, T., Bolt, D., & Froelich, A. (2003). *On the use of collateral item response information to improve pretest item calibration*. Technical Report (pp. 98–13). Newtown, PA: Law School Admission Council.

Sympson, J., & Hetter, R. (1985). *Controlling item exposure rates in computerized adaptive testing*. San Diego: Paper presented at the annual conference of the Military Testing Association.

Tomasik, M., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, *9*, 1–17. https://doi.org/10.3389/fpsyg.2018.02245.

Veldkamp, B., Verschoor, A., & Eggen, T. (2010). A multiple objective test assembly approach for exposure control problems in computerized adaptive testing. *Psicologica*, *31*, 335–355.

Wainer, H., & Mislevy, R. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computer Adaptive Testing: A Primer: Hillsdale*. NJ: Lawrence Erlbaum.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. https://doi.org/10.1007/BF02294627.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 17
Reinforcement Learning Applied to Adaptive Classification Testing

Darkhan Nurakhmetov

Abstract This study investigates how computerized adaptive classification testing task can be considered as a sequential decision process and made accessible to Reinforcement Learning. The proposed method performs a sequential item selection that learns which items are most informative, choosing the next item depending on the already administered items and the internal belief of the classifier. A simulation study shows its efficiency for tests which require to make a confident classification decision with as few items as possible. Solutions for a variety of practical problems using the proposed method are considered in this study.

17.1 Introduction

The key component of a computerized classification testing (CCT)/adaptive testing (CAT) program is the "adaptive" item selection algorithm, which should find the best suited items for each test taker based on his or her estimated ability.

Several item selection methods have been proposed based on item response theory, including the maximization of item information at the current θ estimate (Reckase 1983), the cut-score (Spray and Reckase 1994, 1996), maximization of item information across a region for θ (Eggen 1999), a log-odds ratio (Lin and Spray 2000), and maximization of item information across θ (Weissman 2004).

In addition to correct classification, several other concerns have to be taken into account when selecting the next item in an operational CAT program. Various non-statistical constraints need to be considered, such as content balancing and exposure control. Several heuristics are proposed for this purpose, including the weighted deviation modelling (WDM) method (Stocking and Swanson 1993), the normalized weighted absolute deviation heuristic (NWADH; Luecht 1998), the maximum priority index (MPI) method (Cheng and Chang 2009); randomization strategies, e.g. randomesque strategy (Kingsbury and Zara 1989); conditional selection strategies,

D. Nurakhmetov (✉)
University of Twente, Enschede, The Netherlands
e-mail: d.nurakhmetov@utwente.nl

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_17

e.g. targeted exposure control strategy (Thompson 2002), shadow test approach (van der Linden and Veldkamp 2005); combined strategies (Eggen 2001), e.g. combined application of Sympson-Hetter strategy (Sympson and Hetter 1985) and Progressive strategy (Revuelta and Ponsoda 1998); maximum information with content control and exposure control (Eggen and Straetmans 2000).

Differences in item selection methods may lead to different choices of items for the same test takers and, consequently, to different classification decisions. The choice of the item selection method is therefore one of the most important parts of computerized adaptive testing.

Test content balancing and item exposure control are separate steps that are generally integrated in the item selection heuristics by limiting the items available for selection or by adding small probability experiments to limit over- or underexposure of items. This study applies a new adaptive item selection method. It explores the possibility to solve problems, such as exposure control and content balancing, that may occur in computerized adaptive classification testing (CACT), and simultaneously attempts to optimize item selection.

17.2 Method

Fisher information is commonly used for item selection, but this information measure is based on an estimate of the ability parameter, and the ability estimate is not very stable, particularly in the beginning of a CAT administration. Therefore, when the estimate is not close to the true value, using the Fisher information criterion might result in inefficient item selection. The observation that item selection procedures may favor items with optimal properties at wrong ability values is generally known as the attenuation paradox (Lord and Novick 1968, Sect. 16.5).

The foundation of new methodology for incorporating expert test development practices in the construction of adaptive classification tests is the application of *reinforcement learning*. It has been applied successfully to various selection problems, including robotics, elevator scheduling, telecommunications, a few games, such as backgammon, checkers and go (see Sutton and Barto 1998). Reinforcement learning refers to goal-oriented algorithms, which learn how to attain a complex objective or maximize along over a particular dimension over many steps.

There are two major factors that make reinforcement learning powerful: the use of samples to optimize performance and the use of function approximation to deal with large environments. Reinforcement learning can solve the difficult problem of correlating immediate actions with the delayed returns they produce. Reinforcement learning can be used in the situation, where there is no information about the environment and the only way to collect it is to interact with it. This could be considered to be a genuine learning problem, which is fit to adaptive classification testing, for example to get information about a test taker's state (classification state) it is needed to give him/her an item (action) and get a response.

In short, an improved approach of item selection is presented, which not only optimally spans the item pool (input space), but also optimizes with respect to data consumption, that is, minimizes the test length needed to classify respondents. Going beyond traditional item selection methods in the computerized adaptive testing, in this paper, we lay out and demonstrate an approach of selecting items in sequence, making the decision which item to select next dependent on previously selected features and the current internal state of the supervised method that it interacts with. In particular, our sequential item selection (SIS) algorithm will embed Reinforcement Learning (RL) into classification tasks, with the objective to reduce data consumption and associated costs of item selection during classification. The main question of this chapter is: "Where do I have to look next, in order to keep data consumption and costs low while maintaining high classification results?" or "Which item should I deliver to the test taker next, in order to keep test constraints satisfied while maintaining high classification results?"

The Framework is mapped out in the following section. After introducing the general idea, the formal definition of sequential classifiers and rephrasing the problem as a Partially Observed Markov Decision Process (POMDP) is represented (see Astrom 1965). In addition, an algorithm to take exposure and content control into consideration is introduced.

17.3 Framework

17.3.1 General Idea

Machine learning is often applied to classification problems: mapping an input x to one of a finite set of class labels C. Regarding *classification testing*, the formulation of the initial problem remains the same: based on a series of item responses of the test takers (vector x), class $\boldsymbol{C}$ is estimated.

In classification testing, item selection is needed as key option for good classification results: filtering out less informative items, while paying attention to exposure and content control. Reformulating this to classic machine learning: item selection is a combinatorial optimization problem that tries to identify those items, which will minimize the generalization error, with different constraints. In particular, item selection can be seen as a process that tries to reduce the amount of redundant data.

Turning classification into a sequential decision process results in item selection and classification becoming an adaptive and intertwined process: deciding which item to select next depends on the previously-selected items and the behavior of the classifier on them. This will be achieved by using a fully trained classifier as an environment for a Reinforcement Learning agent, that learns which item to select next, receiving the reward on successful classification of the partially uncovered input response pattern. The general schema of reinforcement learning is described below (Fig. 17.1).

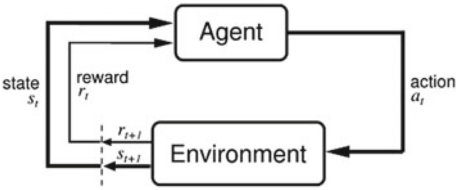


Fig. 17.1 Typical framing of reinforcement learning scenario

Figure 17.1 represents how reinforcement learning works in general: an agent interacts with an environment by taking actions, which is translated into a reward and a representation of the state, which are fed back to the agent. Typical framing of RL can be mapped to CACT as follows:

Agent. An agent takes actions. In CACT, the algorithm which select items or terminate a test is an agent.

Action. Action is the set of all possible moves the agent can make. An action is almost self-explanatory, but it should be noted that agents choose among a list of possible actions. In CACT, the list includes all the items and the action which refers to termination of the test.

Environment. The environment is the world through which the agent moves. The environment takes an agent's current state and action as input, and returns as output the agent's reward and its next state. In CACT, the environment could be the rules of our classification process, such as, the metric used for making decisions, the maximum number of items in the test, the exposure rate, etc.

State. A state is a concrete and immediate situation in which an agent finds itself; i.e. a specific moment, an instantaneous configuration that puts an agent into relation to other significant things. In CACT, it can be the number of items that have already been administered, the probabilities of selecting next items, etc.

Reward. A reward is the feedback by which we measure the success or failure of the agent's actions. Rewards can be immediate or delayed. In CACT, rewards can be granted after an agent terminates the test. The reward is delayed, when it is granted after the test has ended. An immediate reward in CACT can be provided as a penalty for the agent, when an agent chooses an item that has already been administered, or an item that would bring the exposure rate over the limit.

17.3.2 Sequential Classification

First of all, it is needed to formulate CACT as a Partially Observed Markov Decision Process (POMDP), making the problem sequential and thus accessible to Reinforcement Learning algorithms. For detailed information about POMDP, the reader is referred to Astrom (1965). For now we focus on the implementation of POMDP. Therefore, some specific notation and some definitions have to be introduced.

Specific notation used in this chapter:

( )—ordered sequences,
{ }—unordered sequences,
$a \bullet k$—appending an element k to a.

Besides, several concepts have to be defined:

Power sequence. Related to power sets, a power sequence, denoted as powerseq(M), is defined as a set of all permutations of all elements in the power set of M, including empty sequence ( ). As an example, for $M = \{0,1\}$, the resulting powerseq(M) = {( ), (0), (1), (0,1),(1,0)}.

Episode. Each test administration is called an episode. During an episode, the item history $h_t \in \text{powerseq}(F)$ is the sequence of all previously selected items in an episode up to and including the current item at time t.

Cost. Costs associated with accessing an item f are represented as negative scalars $r_f^- \in \mathbb{R}, r_f^- < 0$.

Reward. A non-negative global reward is defined as $r^+ \in \mathbb{R}, r^+ \geq 0$ for correctly classifying an input.

Classifier. Classifiers in general are denoted with the symbol K. A sequential classifier is defined by K*, and it is to be a functional mapping from the power sequence of item responses to a set of classes, i.e., $K^* : \text{powerseq}\big(\{f(x)\}_{f \in F}\big) \rightarrow \boldsymbol{C}$.

A requirement of CACT as POMDP is to process the sequence one input at a time in an online mode, rather than classifying the whole sequence at once. Besides, as output, a class label is added to each input. Therefore, K* requires some sort of memory. Recurrent Neural Networks (RNN) are known to have implicit memory that can store information about inputs seen in the past (see Hopfield 1982; Hochreiter and Schmidhuber 1997). In this chapter RNN and RNN with Gated Recurrent units (GRU) are used (Cho et al. 2014).

Dealing with a POMDP implies that we need to extract an observation from the classifier that summarizes the past into a stationary belief. Most classifiers base their class decision on some internal belief state. A Feed Forward Network (FFN), for example, often uses a softmax output representation, returning a probability p_i in [0,1] for each of the classes with $\sum_{i=1}^{\boldsymbol{C}} p_i = 1$. If this is not the case (e.g., for purely discriminative functions like a Support Vector Machine), a straightforward belief representation of the current class is a k-dimensional vector with 1-of-k coding.

To finally map the original problem of classification under the objective to minimize the number of items to be administered to POMDP, the elements of the 6-tuple $(S, A, O, \mathrm{P}, \Omega, \mathcal{R})$ which describes POMDP can be described as follows. The state $s \in S$ at timestep t comprises the current input response pattern x, the classifier K*, and the previous item history $h_{t\text{-}1}$, so that $s_t = (x, K^*, h_{t-1})$. This triple suffices to fully describe the decision process at any point of time. Actions $a_t \in A$ are chosen from the set of items $F\backslash h_{t\text{-}1}$, i.e., previously administered items are not available. The observation $o_t \in O$ is represented by the classifier's internal belief of the class after seeing the values of all items in h_{t-1}, written as $o_t = b(x, K^*, h_{t-1}) = b(s_t)$. The probabilities p_i for each class serve as an observation to the agent: $o_t = b(x, K^*, h_{t-1}) = \big(p_1, p_2, \ldots, p_{|\boldsymbol{C}|}\big)$.

Assuming a fixed x and a deterministic, pretrained, classifier K*, the state and observation transition probabilities P and Ω collapse and can be described by a deterministic transition function T, resulting next state $s_{t+1} = T_x(s_t, a_t) = (x, K^*, h_{t-1} \cdot a_t)$ and next observation $o_{t+1} = b(s_{t+1})$. Finally, the reward function $\mathcal{R}^a_{ss'}$ returns the reward r_t at timestep t for transitioning from state s_t to s_{t+1} with action a_t. Given c as the correct class label, it is defined as:

$$r^t = \begin{cases} r^+ + r^-_{a_t} if K^*\big((h_\tau(x))_{0<\tau\leq t}\big) = c \\ r^-_{a_t} \qquad\qquad\qquad\qquad else \end{cases}$$

17.3.3 Item Selection

In CACT, one needs to ensure that an item (action) is only chosen at most once per test taker (per episode), i.e., the set of available actions at each given decision step is dependent on the history h_t of all previously selected actions (items) in an episode (for one test taker). Note that this does not violate the Markov assumption of the underlying MDP, because no information about available actions flows back into the state and therefore the decision does not depend on the item history.

Value-based reinforcement learning can solve this problem. The action value V is the expected return for selecting action in a state and following policy. By manually changing all action-values V(o, at) to $-\infty$ after choosing action (item) at, this leads to all actions not previously chosen in the current episode having larger value and be preferred over a_t. A compatible exploration strategy for this action selection without replacement is Boltzmann exploration (Cesa-Bianchi et al. 2017). Here a probability of choosing an action is proportional to its value under the given observation:

$$p(a_t|o_t) = \frac{e^{V(a_t|o_t)/\tau}}{\sum_a e^{V(a_t|o_t)/\tau}},$$

where τ is a temperature parameter that is slowly reduced during learning process for greedier selection towards the end. Thus, when selecting action a_{t+1}, all actions in h_t have a probability of $e^{-\infty} = 0$ of being chosen again. At the end of an episode, the original values are restored.

Having defined the original task of classification as a POMDP and solved the problem of action selection without replacement (item selection), it is possible to use existing algorithms for solving this class of problems. Since the transition function is unknown to the agent, it needs to learn from experience, and a second complication is the continuous observation space.

17.3.4 Algorithm

In this chapter, the actor-critic algorithm is used (see Mnih et al. 2016). The Actor-critic algorithm maintains a policy $\pi((t_t, a_t)|s_t;\theta)$ and an estimate of the value function $V(s_t;\theta_v)$, where θ—parameters, which are learned by iteratively minimizing a sequence of loss functions, and policy represents how items (actions) are picked from the item pool; t—termination gate—action which represent the termination step of the test. Policy and a value can be briefly described as follows:

Policy. A policy is the strategy that the agent employs to determine the next action based on the current state. It maps states to actions, the actions that promise highest reward.

Value. The expected long-term return, as opposed to the short term reward. Value is defined as the expected long-term return of the current state under the policy.

The actor-critic algorithm operates in a forward view and uses a mix of n-step returns to update both the policy and the value-function. The general schema of the actor-critic algorithm is shown in Fig. 17.2. The policy and the value function are updated after every t_{max} actions or when a terminal state is reached. The update performed by the algorithm can be seen as $\nabla_{\theta'}\log\pi((t_t, a_t)|s_t;\theta')A\big(s_t, a_t;\theta,\theta'\big)$ where $A\big(s_t, a_t;\theta,\theta'\big)$ is an estimate of the advantage function given by $\sum_{i=0}^{k-1}\gamma^i r_{t+i} + \gamma^k V(s_{t+k};\theta_v) - V(s_t;\theta_v)$, where k can vary from state to state and is upper-bounded by t_{max}.

The state sequence $\boldsymbol{s}$ is hidden and dynamic, controlled by an RNN sequence model. The Actor-critic algorithm performs a "classification" action a_T at the T-th step, which implies that the termination gate variables equal $t_{1:T}= (t_1= 0, \ldots, t_{T-1}= 0,\ t_T = 1)$. Basically, at the T-th step, the actor-critic model (ACM) gets enough information to terminate tests and make a classification decision. The ACM learns a stochastic policy $\pi((t_t, a_t)|s_t;\theta)$ with parameters θ to get a distribution of termination actions, to continue administering items, or to stop, and of a 'classification' action,

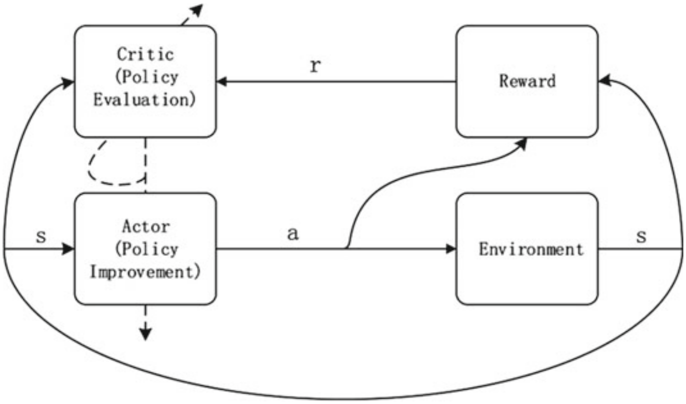


Fig. 17.2 Typical framing of actor-critic algorithm

if the model decides to stop at the current step. The termination step T can vary from test taker to test taker. However, it is possible to set the maximum of T, to ensure that a test ends anyway at step T. Then, if T is reached, classification decision is made by existing information from steps 1 to T.

The parameters θ are trained by maximizing the total expect reward. The expected reward for a test taker is defined as:

$$J(\theta) = \mathbb{E}_{\pi(t_{1:T}, a_T; \theta)}\left[\sum_{t=1}^{T} r_t\right]$$

The reward can only be received at the final termination step when a classification action a_T is performed. We define $r_T = 1$ if $t_T = 1$ and the classification is correct, and $r_T = -1$ otherwise. The rewards on intermediate steps are zeros, except for cases when items are repeatedly selected, in that case rewards are also -1. This is done to penalize the agent for picking same items for the same test taker; and cases when selecting an item results in exceeding the limit in exposure rate, in that cases reward -0.5. J can be maximized by directly applying gradient-based optimization methods. The gradient of J is given by:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi(t_{1:T}, a_T; \theta)}[\nabla_\theta \log\pi((t_{1:T}, a_T; \theta)r_T]$$

Motivated by the REINFORCE algorithm (Williams 1992), the following computation of $\nabla_\theta J(\theta)$ is estimated:

$$\begin{aligned}&\mathbb{E}_{\pi(t_{1:T}, a_T; \theta)}[\nabla_\theta \log\pi((t_{1:T}, a_T; \theta)r_T]\\ &= \sum_{(t_{1:T}, a_T) \in \mathbb{A}^\dagger} \pi(t_{1:T}, a_T; \theta)[\nabla_\theta \log\pi((t_{1:T}, a_T; \theta)r_T - b_T],\end{aligned}$$

where $\mathbb{A}^\dagger$ represents all the possible episodes, and T, $t_{1:T}$, a_T and r_T are the termination step, termination action, classification action, and reward, respectively, for the $(t_{1:T}, a_T)$ episode, while. $b_{T,}$ which is called the reward baseline in the RL literature is introduced to lower the variance (Sutton 1984). It is common to select $b_T = \mathbb{E}_\pi[r_T]$(Sutton et al. 1999), and this term can be updated via an online moving average approach: $b_T = \lambda b_T + (1 - \lambda)r_T$. However, this might lead to slow convergence in training the ACM. Intuitively, the average baselines $\{b_T; T = 1 \ldots T_{max}\}$ are global variables, independent of test takers. It is hard for these baselines to capture the dynamic termination behavior of an ACM. Since an ACM may stop at different time steps (different test length) for different test takers, the adoption of a global variable without considering the dynamic variance for each test taker is inappropriate. To resolve this weakness in traditional methods and to account for the dynamic characteristics of an ACM, an instance-dependent baseline method to calculate $\nabla_\theta J(\theta)$ is proposed. The gradient can be rewritten as:

$$\nabla_{\theta} J(\theta) = \sum_{(t_{1:T}, a_T) \in \mathbb{A}^{\dagger}} \pi(t_{1:T}, a_T; \theta)[\nabla_{\theta} \log \pi((t_{1:T}, a_T; \theta) r_T - b],$$

where the baseline $b = \sum_{(t_{1:T}, a_T) \in \mathbb{A}^{\dagger}} \pi(t_{1:T}, a_T; \theta) r_T$ is the average reward on the $|\mathbb{A}^{\dagger}|$ episodes for the n-th training test taker. It allows different baselines for different training test takers. This can be beneficial since the complexity of training test takers varies significantly.

17.4 Experiments

In this section, evaluation of performance of the ACM is presented. A simulation study utilized a Monte Carlo simulation methodology, with 1000 test takers multiplied by 100 administrations simulated, to evaluate differences in efficiency and accuracy. The population of test takers was randomly drawn and selected from a *N(0,1)* distribution. With Monte-Carlo simulation, item responses are generated by comparing a randomly generated number $0.0 < r < 1.0$ to the probability of an each possible response for each test taker to each item. The probability is calculated using the Generalized Partial Credit Model (GPCM) (see Muraki 1992) and the true test taker's θ, which is known.

If items are randomly selected, this means that, after each administration, each item that has not been used has an equal chance of being selected. However, with such a method of selecting items, the contents of the test are not matched to the ability of the candidate, consequently, accuracy of decisions may result as inappropriate, and testing is not efficient. Simulation studies have shown that random selection eliminates out the possible gains of adaptive testing. If items are randomly selected, there is an average gain in item pool utilization and exposure rate, but also a loss of accuracy.

Although, in CACT, the optimum item selection method makes sure that each test taker gets a different test, it still occurs that some items from the item pool are administered very frequently while others are never used or hardly ever used. The gains in efficiency go along with two following characteristics of the utilization of the item pool, which result in the following problems:

(1) Overexposure. Some items are selected with relatively high frequency, that test security is compromised;
(2) Underexposure. Some items are rarely selected that one wonders how the expenses of developing them can be justified.

Table 17.1 contains the classification accuracies, maximum exposure rate and item pool utilization for the simulation studies. Table 17.1 also shows that problems, that have been explained above, do not occur with random item selection, where all items are used. However, with random selection there is a loss in accuracy.

The RL-CACT, selecting items on the basis of ACM with $T = 10$ (maximum length of the test) and with $T = 15$, uses 65 and 71% items out of item pool, respectively. This

Table 17.1 Results of simulation studies

Item selection algorithm	Correct decisions (accuracy) (%)	Maximum exposure rate (%)	Item pool utilization (%)
Random (T = 10)	83.6	12.1	0
Random (T = 15)	85.5	16.2	0
RL (T = 10)	89.8	33.1	35
RL (T = 15)	92.1	34.3	29

is done by implementation of the ACM reward policy as explained in the previous section of this chapter.

Regarding the accuracies, ACM with both T = 10 and T = 15 performed with 89.8 and 92.1% of correct decisions, respectively. Maximum exposure rates are 33.1 and 34.3%. One can vary these rates by changing the reward for the agent, e.g. if one wants to have maximum exposure rate less than 30%, it can be done by tightening penalty for overexposing items.

17.5 Discussion

Classification CAT is formulated as a POMDP and thereby it is made accessible to RL methods. In this chapter, the main focus is on minimization the number of items (test length) needed to make a confident classification decision. This is done by training the RL agent to pick items that lead to quick and precise classification. Exposure control is maintained by using the different rewards for action selection, which takes into account the item selection history for each test taker. The proposed reward distribution penalizes the agent during the training process if it picks an item, which already has been administered for a certain test taker. Also, it penalizes the agent if the agent selects an item, for which the exposure rate limit is exceeded.

Another issue that might occur with adaptive testing is the content balancing problem. The presented approach can solve this by implementing a specific mask for the items or by giving different rewards for sequences of items. For example, if it assembles a test from an item pool with *n* domains and *M* items for each domain, and there is a constraint that exactly *m* of items from each domain have to be administered. Then we can define the following reward distribution function: give the reward *R* for the correct classification, and subtract value *d* for each unadministered domain, or *d/m* for each missing item from domain.

Any additional constraints can be taken into account by tuning the reward distribution. For instance, for different types of assessment, a different outcome might be desired, e.g. one wants higher accuracy, or higher precision or recall. This can be done by varying rewards during the training process, e.g., after classifying a small sample of test takers, a (batch) additional reward, depending on the desired outcomes, could be implemented. For instance, if in particular batch the number of type II errors exceed limits, then the reward can be reduced by a specific value.

In summary, the proposed method for adaptive item selection in computerized classification testing is efficient for tests requiring to make a confident classification decision with as few items as possible. Moreover, the proposed method can solve problems, such as overexposure and underexposure, with content-distribution constraints, through the reward allocation. However, the design of CACT with ACM model requires simulation research to ensure that the test is implemented to be as efficient as possible.

References

Astrom, K. J. (1965). Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications, 10,* 174–205.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Cesa-Bianchi, N., Gentile, C., Lugosi, G., & Neu, G. (2017). Boltzmann exploration done right. In *Advances in neural information processing systems* (pp. 6284–6293).

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*(2), 369–383.

Cho, K., van Merrienboer, B., Bahdanau, D. & Bengio, Y. (2014). *On the properties of neural machine translation: Encoder-decoder approaches*. Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.

Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling*. Presented in NIPS 2014 Deep Learning and Representation Learning Workshop.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249–261.

Eggen, T. J. H. M. (2001). *Overexposure and underexposure of items in computerized adaptive testing*. Measurement and Research Department Reports 2001-1, Arnhem, The Netherlands: CITO Groep.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713–734.

Hochreiter, S., & Schmidhuber, J. (1997). Long shortterm memory. *Neural Computation, 9*(8), 1735–1780.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences of the USA* (vol. 79 no. 8 pp. 2554–2558). April 1982.

Husken, M., & Stagge, P. (2003). Recurrent neural networks for time series classification. *Neurocomputing, 50,* 223–235.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359–375.

Lin, C.-J. & Spray, J. A. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test*. (Research Report 2000–8). Iowa City, IA: ACT, Inc.

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

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22*(3), 224–236.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35,* 229–249.

Mnih V. et al. (2016). Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237–254). New York: Academic Press.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311–327.

Spray, J. A., & Reckase, M. D. (1994, April). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the national meeting of the National Council on Measurement in Education, New Orleans LA.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21,* 405–414.

Sutton, R. S. (1984). Temporal credit assignment in reinforcement learning (Ph.D. Dissertation).

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press. ISBN 978-0-262-19398-6.

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems (NIPS), 12,* 1057–1063.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*(2), 151–166.

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 CA: Navy Personnel Research and Development Centre.

Thompson, T. (2002, April). Employing new ideas in CAT to a simulated reading test. *Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME)*, New Orleans, LA.

van der Linden, W. J. & Veldkamp, B. P. (2005). *Constraining item exposure in computerized adaptive testing with shadow tests*. Law School Admission Council Computerized Testing Report 02–03.

Weissman, A. (2004). *Mutual information item selection in multiple-category classification CAT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego CA.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning 8*, 3–4 (1992), 229–256.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part V
Technological Developments in Educational Measurement

Chapter 18
Feasibility and Value of Using a GoPro Camera and iPad to Study Teacher-Student Assessment Feedback Interactions

Fabienne van der Kleij, Lenore Adie and Joy Cumming

Abstract The importance of effective feedback for learning is widely recognised. However, much of the previous research on assessment feedback has paid limited attention to the nature of feedback interactions, and teacher and student perspectives. Innovations in technology have opened up possibilities to capture such interactions for research and professional development purposes. This study investigated the feasibility and value of using easily accessible equipment for practitioner use and research purposes. Specifically, the study focussed on use of a GoPro camera and an Apple iPad in capturing one-to-one teacher-student feedback interactions, and subsequent individual video-stimulated recall (VSR) for self-reflection. Six teacher-student pairs in three Australian secondary schools participated in the research. Evidence collected included technical data and characteristics of the devices, teacher and student experiences, and researchers' experiences and reflections. While the iPad and GoPro were both useful for capturing interactions, the iPad was more useful and practicable for VSR interviews. Findings suggest that such technology has potential for use by teachers and students to improve reflection and feedback interaction, and thus to enhance student learning. The researchers identified several technical issues that occurred in a range of settings and make suggestions to overcome these issues.

18.1 Introduction

The importance of effective feedback for learning is widely evidenced in international education research and policy (Stobart 2008; Wiliam 2011). The purpose of feedback is to provide information, to both teachers and students, to improve teaching and learning. While feedback can take many forms (Van der Kleij et al. 2015), feedback within classroom practice is generally realised through written and oral communications and interactions between teachers and students. The significance of

F. van der Kleij (✉) · L. Adie · J. Cumming
Institute for Learning Sciences & Teacher Education,
Australian Catholic University, Brisbane, Australia
e-mail: Fabienne.VanderKleij@acu.edu.au

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_18

effective feedback has been identified in research (Hattie 2009). However, research has also identified the variability of feedback effectiveness (Winstone et al. 2017). An effective feedback interaction between teachers and students requires development of some common understanding of goals and quality learning, and engagement by the students in progressing their learning (Black and Wiliam 2009). Thus, the quality of feedback interactions has become a major focus of education research and teacher professional development.

This chapter reports on a study that investigated the feasibility and value of using easily accessible equipment for practitioner use and research purposes. Specifically, we aimed to generate detailed insights into the feasibility of using a GoPro camera and an Apple iPad to capture one-to-one feedback interactions between teachers and students, used for subsequent individual video-stimulated recall (VSR) for self-reflection (Lyle 2003). We were interested in the comparison between these devices for practitioner use as well as for research purposes. Our research purpose also was to explore designs whereby teachers and students would be able to collect research evidence independent of a research team presence, using technology that could be readily available in schools. The overall purpose of such research is to develop effective but simple processes for use by both teachers and students to engage with feedback to improve learning. Important outcomes of this study are insights into possibilities and limitations of using these devices, possible ideas for improvement, and potential for upscaling to whole classroom contexts.

18.1.1 The Value of Video Feedback

Researchers have noted the advantages of using video for communicating feedback between teachers and students (e.g. Crook et al. 2012) and in peer feedback settings (Hung 2016; Lenters and Grant 2016). For example, Lenters and Grant (2016) noted the value of iPad video recordings as rich media for the communication of peer and teacher feedback. Teachers and students reported that use of these devices aided in smooth communication of feedback as intended, as meaning and intention often became distorted through only written feedback. One student in the study by Lenters and Grant noted the particular value of viewing body language in feedback communication as helping understand the intention behind feedback content.

The major advantage of using video in educational research is the opportunity to capture the full record of an interaction (Stigler et al. 2000), providing rich sources for reflection upon review. Reviewing video records of events has proven to be useful to support teacher self-reflection with the potential to encourage changes in teaching practices (Gröschner et al. 2015; Harlin 2014) and enhance teaching quality (Tripp and Rich 2012). The power of video lies in its possibilities to allow teachers to watch the recordings and view their feedback practices from a different perspective (Charteris and Smardon 2013; Harlin 2014; Jordan 2012; Van den Bergh et al. 2014).

Digital technology allows for video recordings to be immediately available for review and reflection. While the act of being videoed may cause teachers and students to initially behave differently, research has found that this effect fades within minutes (Rowe 2009) with the benefits of the technology outweighing this limitation. One method that has particular value for self-reflection, going beyond simply reviewing an interaction, is use of recorded video as a stimulus for video-stimulated recall (VSR) (Lyle 2003). This method is suitable to gain insight into teachers' and students' thinking and feelings in feedback interactions, as the video stimulus helps them recall their thoughts and reflect on their own and each other's behaviour and actions (Van der Kleij et al. 2017).

VSR has also proven to be useful for use in feedback research involving teachers (Van den Bergh et al. 2014) and students (Hargreaves 2012, 2013). While video technology has been used to capture teacher-student interactions and reflections, the use of video technology as stimulus for both *teacher* and *student* reflection on the same feedback interactions is less well documented. This is the focus of our research design, with our study examining the potential of both GoPro camera and iPad technologies for research suitability as well as potential independent use.

GoPro cameras are compact action cameras, designed to capture a variety of activities. The camera can be positioned in a broad range of ways, for example on a head mount, chest mount, handheld monopod or tripod. In classroom settings, this opens up possibilities to record learner actions and interactions from unique angles (Hummel 2015). iPads are multi-purpose tablet devices, which can serve many educational purposes, such as a platform for educative games and simulations (Murphy 2016). Although the use of tablet devices such as iPads is becoming increasingly common in classrooms, their uses for capturing teacher-student interactions through its video functionality are not well documented.

18.2 Method

18.2.1 Participants and Context

The study occurred during a six-month period across 2015 and 2016. It involved six teacher-student pairs (Pairs A–F) in Year 9 (students aged 12–13) from three Australian secondary schools, with two pairs participating in each school. All participants were volunteers. The investigation focused on feedback interactions in three subject areas in order to capture a range of potential contexts: Science, English, and Health and Physical Education (HPE). For a more detailed description of the study participants, see Van der Kleij et al. (2017).

18.2.2 Data Collection Instruments and Procedures

In this study we used a GoPro Hero 4 silver, which has a display screen at the back for ease of camera positioning and reviewing of the captured material (contrary to previous models). We used an iPad Air2 (64 GB) and used the default camera application.

Following ethics approval, school principals were contacted to request participation by two teacher-student pairs in their school. Teachers were asked to select a piece of student work to be discussed in a one-to-one feedback session with the student; this feedback conversation was the focal point of this study. The research involved teachers in different disciplines and allowed for a range of assessment formats and modes. However, the assessment design had to be based on the Australian Curriculum Achievement Standards evident in a formal marking rubric or criteria overview. The researchers did not prompt the teachers or students about the importance of feedback. Written consent was obtained from school principals, teachers, students and parents/guardians.

The data collection procedures for this study included three stages. In Stage 1, video data of 10–15 min feedback interactions were collected using a GoPro and iPad simultaneously from a similar position for comparison of usability of the two devices. Two researchers were present during the feedback sessions to support the operation of the technology.

Previously, we had investigated use of the GoPro camera as a wearable item for the teacher and student, a common use for this technology. However, occupational health and safety (OHS) requirements meant the camera could not be head-mounted and would need to be chest-mounted which may still cause OHS issues. As a chest-mounted camera would not capture interactions, and mindful of the OHS requirements, we investigated only the static camera. One of the teachers commented that replaying the video of a GoPro camera that was not in a static position when recording can result in motion sickness. Thus, using the GoPro in a static position appeared to be the most feasible option for this study.

Stage 2 involved VSR interviews with each student and teacher individually using the iPad. In other research using video-stimulated recall, the focal instances that participants have been asked to comment on have been identified by the researcher or the participant, or both (Rowe 2009). For example, in Hargreaves' (2012, 2013) study involving primary education students, the focal instances were identified by the researcher. In our study, a researcher asked the participant to pause the video at points they identified as relevant, and reflect on those instances. By interviewing both the teacher and student separately, we were able to directly compare their reflections on the feedback interaction, and the significance of different aspects of the feedback for each. The VSR sessions were audio recorded and took approximately 20–30 min. Video recordings and VSR sessions were transcribed for research analysis.

We requested for the feedback conversations and the VSR interviews to take place in a relatively quiet space, where no other students would be captured on video, for example, a small office, the library, or a quiet corner of a classroom. However, on some occasions this was not possible due to the organisation of a high school timetable and the types of feedback sessions that were recorded. As a result, the data were collected in a range of different settings. This, of course, reflects the reality of diverse settings for teacher-student interactions in schools.

Our initial intention was to have 50% of the participants watch the GoPro video on the iPad, and 50% watch the iPad video on the iPad. However, a preliminary outcome was that switching the GoPro Wi-Fi connections between the remote control and iPad was time-consuming, and the battery life of the GoPro appeared insufficient to do this. For this reason, all participants were shown only the videos taken by the iPad, on the iPad.

Stage 3 involved a short teacher questionnaire regarding their experiences using the GoPro and iPad for capturing the feedback conversation and self-reflection (See Appendix 1). The researchers also recorded written observations of the technical set-up as well as the dynamics of the teacher-student interaction, and took notes of any technical issues that emerged or issues identified by the teacher or student during the data collection process related to the use of the GoPro or iPad.

The positioning of the cameras differed across observations, depending on the data collection setting. Teachers and students were asked to sit as they normally would in a one-to-one feedback interaction. Video resolution of the GoPro can be modified; it was set at the same resolution as the iPad (1920 × 1080 pixels) to allow for direct comparison of the image quality. Videos were recorded on medium screen width in order to minimise the fish eye effect. We used a Wi-Fi remote to start and stop the recording. We used the rear camera on the iPad, placed diagonally behind the GoPro, as the aperture in the video function required the device to be farther back than the GoPro.

No additional microphones were used to enhance the quality of the sound as the aim was to test the default quality of these two video technologies. However, a special case was necessary so that the GoPro could be mounted on a tripod without covering the microphone. No external speakers were used to enhance the volume of the videos during the VSR sessions. The simplicity of the set-up and the minimalist equipment requirement were essential to test the ease to which these devices could be independently used by teachers for both research and pedagogical purposes within a variety of educational contexts.

18.2.3 Analysis

Feasibility of the GoPro camera and iPad was evaluated based on the following criteria:

(1) quality of the image;
(2) quality of the sound;
(3) battery life; and
(4) ease of use.

Two members of the research team operated the equipment for recording the feedback interactions. However, ideally teachers and students should be able to independently collect the video data. Ease of use was therefore an important criterion.

In order to analyse the feasibility of the two devices for capturing one-to-one feedback interactions we analysed three sources of data, taking account of the different contexts in which these data had been collected:

(1) *Technical data and characteristics of the two devices.* The quality of video data from the devices was compared for feasibility to capture teacher-student interactions, with the possible extension to use in whole classroom interactions. Specific attention was paid to the evaluation of visual and audio quality of the data and comparison of the GoPro camera and iPad as technological tools to inform future research. The researchers' field notes on technical matters were analysed qualitatively according to the four specified criteria to evaluate the feasibility of the GoPro camera and iPad for the particular research purpose.

(2) *Teacher and student experiences.* Comments made by teachers and students during the VSR regarding the experience of being video-recorded and participating in the VSR were analysed qualitatively. The data were coded using purposefully designed coding frameworks for the feedback interactions (Adie et al. 2018) and VSR interviews (Van der Kleij et al. 2017, Fig. 3). The elements of the VSR coding framework that were relevant to this paper were 'experience and reflection on being video-taped' and 'value of video reflection and insights'. In addition, teacher and student comments were considered within "a descriptive model of key conceptual influences on learners' proactive recipience of feedback" (Winstone et al. 2017, 31). This model was derived from a synthesis of the literature related to the students' role in feedback processes. The model emphasises the influence of feedback interventions on the proactive recipience of feedback, through the interplay of the feedback receiver's self-appraisal, assessment literacy, goal-setting and self-regulation, and engagement and motivation, and various interpersonal communication variables, such as the characteristics and behaviour of the feedback sender and receiver. Teacher surveys on their user experience were analysed both quantitatively and qualitatively, to take account of teachers' qualitative comments.

(3) *Researchers' experiences and reflections.* Researchers' field notes were synthesised to provide overall considerations of the comparability and success in using the two devices for the research process.

18.3 Results

18.3.1 Technical Results

The feedback interactions took place in different types of rooms that were representative of the diversity of teaching spaces that could be encountered in schools and in a research project. As a consequence, the quality of the video footage obtained differed widely across the four different settings. As the study progressed, some procedures were revised slightly as a result of learnings from data collection experiences

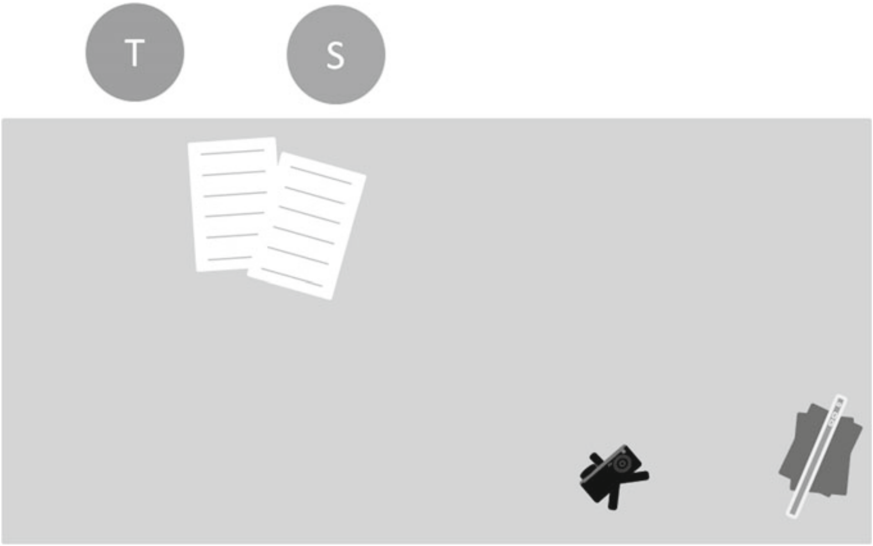


Fig. 18.1 Camera set up school 1

at the previous school. Table 18.1 provides an overview of the camera set up and characteristics of the data collection instances, along with issues identified.

The quality of the visuals was satisfactory for both devices across all recordings. In School 1 (see Fig. 18.1), an issue was that the researchers' small tablet tripod did not allow for filming at an appropriate angle, therefore a stack of books of approximately 10 cm in height was used for better camera positioning. This camera angle was still not optimal in the second video with a slightly taller teacher. After data collection in School 1 a height adjustable floor stand with a 360° flexible angle for the iPad was purchased to allow for greater flexibility for data collection in Schools 2 and 3.

The flexible floor stand was valuable as it allowed for greater flexibility for camera positioning of the iPad. The GoPro tripod was easily adjustable in height and angle, and the use of the remote helped keep the camera in position.

The audio volume of the video was acceptable for most of the GoPro recordings, but it must be noted that we used a special case that does not cover the microphone. At times, the volume of the videos recorded on the iPad was too low. This was especially the case when the iPad had to be positioned farther away from the participants, and in noisy environments and large spaces.

In School 2, videoing of the first feedback interaction took place in an HPE hall with open sides. The table tennis table of the participating teacher and student (Fig. 18.2) was slightly separated by a heavy-duty curtain from the rest of the tables on which the other students played.

This was a noisy environment for videoing; as a result the sound quality for both the GoPro and iPad videos was poor, and it was not possible to transcribe all participant dialogue. However, the teacher and student were still able to use the video in the

Table 18.1 Description of data collection: setting, positioning and identified issues

Data collection instance	Data collection setting	GoPro positioning	iPad positioning	Issues identified
School 1 conversations A and B	Large quiet meeting room of 10 m × 5 m	25 cm high tripod on table	Small tablet tripod on top of stack of books of 10 cm on table	Angle of iPad not optimal
School 2 conversation C	• Video recording: HPE hall, participating teacher and student were slightly separated from other students • VSR: staffroom adjacent to the HPE hall and sports field	58 cm high tripod on floor	125 cm high floor stand	• Noisy environment • iPad was placed 1 m farther back from the GoPro to capture both participants on video, resulted in poor audio quality
School 2 conversation D	Large and mostly quiet meeting room of 5 m × 8 m	25 cm high tripod on table	125 cm high floor stand	• Some background noise from students in hallway • Volume of video recording not loud enough during VSR
School 3 conversations E and F	Small and narrow room of 4 m × 2.5 m with windows on both long sides of the room	25 cm high tripod on table	125 cm high floor stand	• Very small room, iPad position very close to the wall in order to capture the image • Difficulties positioning camera as not to capture other students on camera through windows

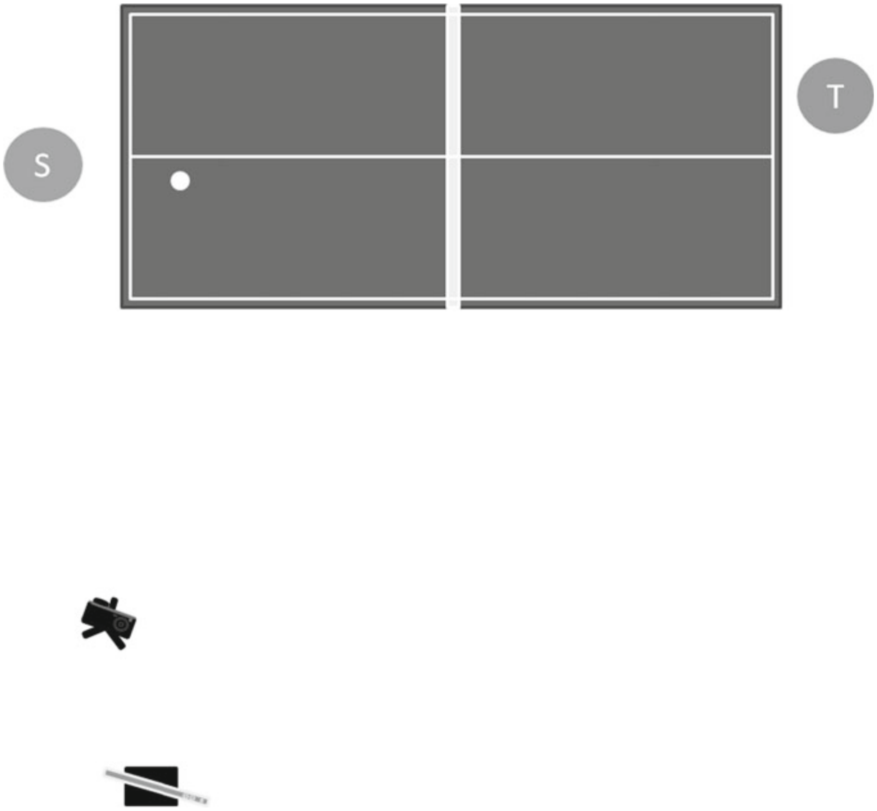


Fig. 18.2 Camera set up School 2, Conversation C

VSR sessions, which occurred in the staffroom adjacent to the hall and a sports field. While this was also a noisy environment at times, as students were playing sports outside and other teachers were also working in the office, the quality of VSR audio was satisfactory for research transcription.

The feedback interactions and VSR interviews with the second teacher-student pair in School 2 took place in a large and mostly quiet meeting room (Fig. 18.3). However, the recording did pick up the background noises when students were moving between classes and congregating outside the meeting room.

The volume of the video recording was not quite loud enough, despite conducting the videoing in a quiet room. This may have been caused by the size of the room and the distance of the iPad to the participants when recording the video.

In School 3, the feedback interactions and VSR interviews took place in a small and narrow room with windows on both sides on the long sides of the room, which posed challenges with respect to positioning the cameras. The GoPro was positioned using its tripod on the table opposite the teacher and student. The iPad had to be in the far corner of the room, very close to the wall, in order to capture the image of

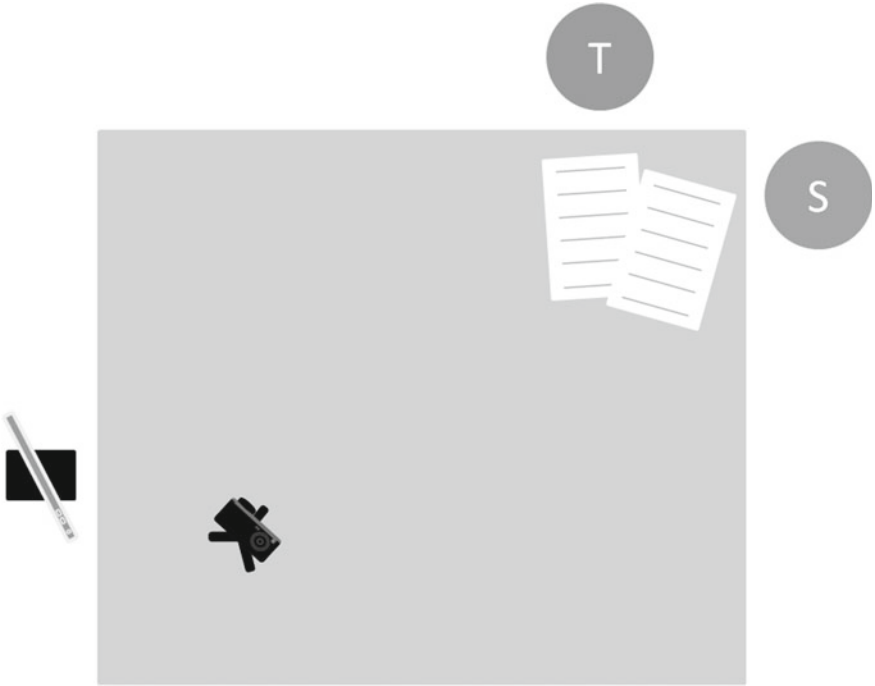


Fig. 18.3 Camera set up School 2, Conversation D

both the teacher and student (Fig. 18.4). There was some outside noise as students moved between classes during the video recording.

Battery life of the two devices was compared during data collection at School 1. Upon commencement of data collection both the iPad and GoPro were fully charged. After recording two feedback conversations (approximately 19 min in total) the GoPro battery was 1/3 full. The use of the remote control and Wi-Fi function caused the battery to run low quickly. The iPad had retained 85% of battery after filming two feedback conversations and replaying of videos to two teachers and two students. Thus, the battery life of the iPad was noticeably better than that of the GoPro.

Teacher survey responses (Appendix 1) further complemented the technical results. The survey first asked whether teachers thought the GoPro is a suitable device to capture teacher-student interactions. Two teachers were neutral, one teacher strongly agreed, one teacher disagreed, and the other two teachers did not respond to this question. The reason one teacher strongly disagreed was based on personal experience rather than this project, and due to the fact that when the GoPro is not in a static position when recording, replaying the video can result in motion sickness. The teachers who did not respond to this question had focused their attention on using the iPad, ignoring the recording on the GoPro.

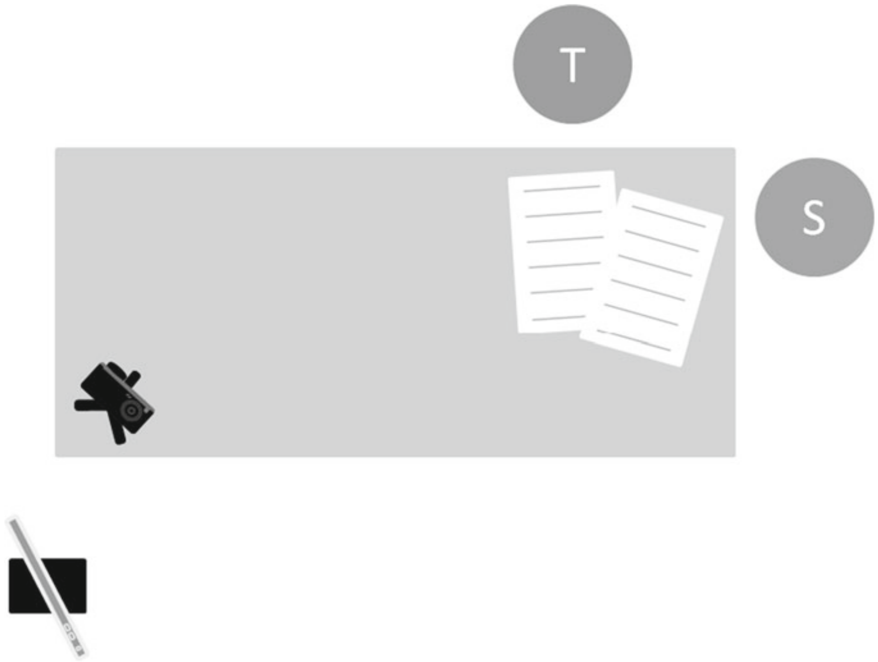


Fig. 18.4 Camera set up School 3 (in conversation E the teacher and student positions were reversed)

Five teachers strongly agreed and one teacher agreed that the iPad is a suitable device to capture teacher-student interactions (Question 2). Four teachers strongly agreed and two teachers agreed that the length of the video observation was appropriate (Question 3). The fourth question focused on the quality of the sound when replaying the video. Three teachers strongly agreed, two teachers agreed, and one teacher disagreed that the sound quality was good. The different contexts for recording may account for these differences in response. Five teachers strongly agreed and one teacher agreed that the quality of the visuals when replaying the video was good (Question 5). None of the teachers indicated having experienced any issues while being recorded by the GoPro or iPad.

18.3.2 Teacher and Student Experiences

As stated, the following data were analysed using the Winstone et al. (2017) descriptive model of feedback recipience. This model encapsulates the purpose of feedback, the interpersonal communication variables of the feedback sender and receiver, and the factors that affect recipience of feedback.

18.3.2.1 Participant Reflections on the Experience of Being Video-Recorded

The experience of being video-recorded differed for each of the participants. Some participants took a few minutes to relax and forget about the cameras, while others did not seem to be aware of them. In response to a survey question about issues experienced while being videoed and during the VSR, one teacher reported: “Only nerves and embarrassment! I’m sure I’ll get over it with more experience” (Teacher A). Feelings of nervousness appeared to be temporary in both teachers and students. The results align with findings by Rowe (2009) that while the video camera may initially distort the normal interactions between teachers and students, it is soon forgotten even in the short space of a 10–15 min feedback session. This finding adds credibility to the use of video research data of classroom interactions.

During the VSR sessions, four teachers (A, D, E and F) reflected on the experience of being video-recorded and watching themselves back on video (see also Van der Kleij et al. 2017). For example, Teacher A said:

> It was very nerve-wracking at the beginning. I didn’t expect it to be … But it was. It was like, ‘Whoa.’ I even slowed down my talking towards the end. At first I was like, ‘Blah, blah, blah, blah.’ Maybe having a camera around more often will relax you more… and get over it.

As the teacher watched the video there is an evident awareness of a tendency to talk rather than listen. While the teacher attributed this response to being videoed, there is recognition of the need to slow down ‘talking’ to communicate feedback. In addition, later in the VSR, the teacher further noted that the student’s facial expression in the video indicated confusion. It is possible that these insights may trigger a response to slow down in future feedback interactions with a student or with the class. In this instance there is a possibility that the VSR may stimulate a visual memory to slow down communication and check on receipt and understanding of feedback.

Teacher F also indicated initial nervousness about being video-recorded as he reflected on what he had learnt from watching the video:

> I’m learning that I ask very long-winded questions, but I think I was a little bit sort of—it sounds silly—but nervous at the start, like just, ‘Is she going to say anything?’ So you’re sort of trying to put it in—put it in her mouth—not put it in her mouth, but like, I don’t know, instigate it maybe a bit too much. No—yeah it’s—it’s all right, it’s good. She’s an easy student to talk to, I think the manner—my manner and everything’s okay, it’s just weird when you see yourself doing that [laughs] … I was cringing there for the first part, to be honest.

In this instance, the teacher is cognisant of his own characteristics as feedback provider and the interpersonal variables that are critical to effective dialogue. Student F also noted that Teacher F was very nervous:

> I can tell in this video Mr [teacher] is really nervous. I can tell—in the sheets he's drawing scribbles, he was, he was just drawing. I could tell he was nervous, but he also kind of like spread humorous, a humorous aura around. And that just honestly—I'm nervous as well, I don't like interviews, I just don't like it—but it kind of made me less nervous, because he's being funny, he's being himself as a teacher.

The student also identified the importance of the feedback provider's behaviour and how this affected her feedback recipience connecting "a humorous aura" to being "less nervous".

Teacher D identified an instance in the video where the experience of being video-recorded made her feel as though she was in a high-stakes situation:

> I wanted it to be an accurate feedback so it would be useful for him. And doing it on the spot and with him, and with you guys filming, was quite high stakes at that moment because I was like I don't want to say the wrong thing here, I don't want to be too liberal, too conservative, I want it to be accurate.

Teacher D indicated her concern with the alignment of her feedback to the curriculum standards and ensuring the feedback was useful to the student. She focussed on her manner of feedback delivery, the characteristics of her feedback message and consideration of subsequent student engagement with her feedback in the context of being videoed.

In contrast, students appeared to be less self-conscious than teachers; only two students (Students B and F) commented on the experience of being video-recorded. Although the majority of teacher participants experienced nervousness due to the presence of the cameras and/or researchers (Teachers A, C and F), only Student F verbally expressed nervousness. However, in some cases the teacher did note nervousness in a student. For instance, in response to the question "How did you find watching the video? Did you learn anything from it?", Teacher E replied: "She was giving very good feedback but she just looked more nervous on the film than what I felt she was. Just to be more aware of that … I do feel like it went well".

When replaying the video Teacher E commented on the student's affective state, which, as she stated, she had not been aware of during the feedback session. As a result of the VSR, the teacher noted that this was an aspect of her practice that she wanted to take into consideration in future feedback interactions. The teacher's response also shows her recognition of the student as feedback provider of information that can be used by the teacher to improve her practice.

18.3.2.2 Participant Reflections on the Usefulness of Using Video for Recording and Reflecting on Feedback

Replaying the video and reflecting on the feedback conversations proved to have considerable educational value. See Van der Kleij et al. (2017) for a detailed analysis and overview of the nature of teacher and student VSR comments. Student reflections mainly focused on reviewing content and strategies for improvement, and reflecting on the characteristics of the feedback message and manner of feedback delivery (Winstone et al. 2017). The nature of the teacher reflections varied more broadly, focusing for instance on characteristics and behaviour of themselves as a feedback sender and the characteristics and behaviour of the student as receiver of feedback, the alignment of feedback to assessment criteria and standards, as well as characteristics of the feedback message and manner of feedback delivery.

The teachers' and students' comments on participation in the video-recorded feedback interaction identified their ability to critique their own performance and reflect more deeply on the others' behaviours. Teachers expressed having gained insights from watching the video about their feedback practice, for example, needing to emphasise the positive, double check for student understanding, and provide time for students to identify quality elements and areas of concern within their own work. For example, Teacher A noted that her feedback was highly corrective in nature, and did not sufficiently recognise the student's strengths. Teachers B, D and F identified instances where they gave the student feedback or asked a question but did not give the student a chance to respond. Teacher C also noted that he might have spoken too much:

> It's nice to sit back and see patterns develop, from an outside view. I think there were a couple of times there where I might have spoken too much instead of like let's play roll on. But at the same time it was good to see where I jumped in and offered feedback … it's amazing to see how much change can happen when you isolate someone for ten minutes.

Teacher C has connected the importance of the context, in this case opportunity to engage in one-to-one feedback, even if only for a short (10-min) period of time and the proactive recipience of feedback to cause changes in student behaviour and learning.

Through watching the video, teachers were able to identify the student's emotions in the feedback interaction thus raising their awareness of the affective impact of their feedback. For example, Teacher A observed embarrassment in Student A, and reflected: "I wonder why she looks embarrassed. I'll ask her why did she look embarrassed—because I think because we pored over and over and over it and she's like, 'Oh yeah, I know that.' Or is it just because she didn't realise?" In this interaction, Student A did not contribute much beyond a single response as acknowledgement of the feedback (e.g. "yeah") while Teacher A provided extensive feedback on all aspects of the student's assessment response. Teacher A's VSR reflection suggests an emerging awareness that feedback involves interpersonal communication rather than a one-way transmission of information.

For students, the degree of self-reflection in VSR responses appeared to be related to the way in which the teacher had structured the feedback context (see Van der Kleij et al. 2017). Most of the students identified that they valued reviewing the feedback as it gave them the opportunity to review key points that were missed in quite densely informative conversations. For instance, Student A noted that there was a lot of information to take in during the feedback session, and reviewing the video was helpful. Student D identified that replaying the video was helpful: “I could have missed something while I was listening to her”. The students’ comments showed an awareness of the manner of feedback delivery (e.g. the quantity of feedback), the characteristics of the message (e.g. the density of information), and the interaction with their own characteristics as feedback receivers. They acknowledged that processing and acting on feedback was important but that they required further time and opportunity to revisit feedback.

Although the research involved only one-on-one feedback interactions, the teachers reported the usefulness of these conversations for their teaching as well as student learning. When asked whether she had found the experience useful, Teacher D replied: “Yes, terrifying [laughing] but very useful”. In the survey, four teachers commented on the usefulness of the experience when asked if they had any other comments. Teacher A noted: “I think it’s really worthwhile and am going to try recording my feedback again!” Teacher A recognised the value of the VSR for her reflective practice and the potential for enhancing her feedback practice. Teacher F wrote: “I thought the process was worthwhile and contained elements I could use in my professional practice”. Teacher F related her learning from the VSR to her classroom teaching practices more broadly. The teachers’ comments indicate that the VSR process supported their reflective practices that moved beyond the one-on-one feedback context to consideration of whole class interactions.

18.3.3 Researcher Experiences and Reflections

Overall, both researchers involved in the video process noted use of the technologies to be smooth and relatively problem free. However, there were specific aspects that required attention dependent on the different locations. When setting up the iPad it is important to take into account the aperture, as the image automatically zooms in when switching to video mode. For this reason, the iPad was positioned farther back than the GoPro. However, this sometimes resulted in low volume in the video recordings. Another important note was that the iPad automatically reverts to camera mode, so it needs to be set on video mode when setting up the camera, and needs to be reset on video mode shortly before recording commences.

In this study the teachers were requested to have 10–15 min feedback conversations. The actual length of the conversations varied between 6 and 18 min with an average 12 min 46 s. Our observation was that in some cases the length of the feedback session and subsequent VSR session was too long for students to review. In the lengthier feedback sessions, there was often too much information for students to remember and then act on; after watching the video, regardless of length, the students recalled no more than three main points from the feedback interaction to address in their work.

Use of the technology was observed to cause only minor technical difficulties for the participants. The video could be started by tapping a play button in the middle of the screen, but had to be paused by tapping a small pause button in the right top corner of the screen. While none of the participants required explanations on how to play and pause the video on the iPad in the VSR sessions, in a number of instances the participants accidentally touched the screen in a place that caused the video to replay from the beginning. This was a minor issue as the iPad did not show the time elapsed when replaying the video, and it then took time to retrieve the right point of play in the video to resume the review of the feedback conversation.

Another technical difficulty relevant for research involved time stamping the instances when the teacher and students paused the video to comment on the feedback interaction, as the time elapsed did not show in the video player (this feature has become available in more recent updates, but the time still only displays for about a second). It was possible to hear the video in the background of the interview, but time-stamping the video pauses was a time-consuming activity involving two researchers to check the accuracy of this record.

18.4 Discussion

This study trialled the technical and experiential feasibility of using a GoPro camera and iPad to capture one-to-one feedback interactions between teachers and students for reflection on feedback interactions, to inform the design of future research projects. Both technologies are easy to access, and iPads, in particular, with a broader range of educational uses are common in schools. The iPad, as a multi-featured device, presented several advantages as well as disadvantages. The advantages included ease of replaying video, screen size for replaying video, ease of sharing or transferring videos, battery life and familiarity to teachers and students. The main disadvantage of the iPad was that the distance required for videoing resulted in low audio quality. This could be addressed by using either a fish eye lens to enable closer camera positioning, or by using an additional microphone to capture audio. Another potential disadvantage of the iPad, although it did not occur during our data collection, could be the video recording stopping during times when connected to the Internet and a phone call occurs. The solution to this would be to have the iPad on airplane mode when recording videos. On the other hand, limitations of the GoPro as standalone technology for this type of research mainly related to the small size of the screen to replay the video. While the GoPro camera was compact and easy to mount, this form of research required a tablet or PC to facilitate VSR processes. As a practical consideration, the GoPro required spare batteries as the recording used much charge.

One critical feature for our research purposes was to know when the teacher and student paused the video in order to compare the different instances they identified as critical in the feedback conversation. However, the elapsed time did not show in the iPad's video player, so while records were made of key words spoken at the time, we had to listen carefully to the feedback conversation in the background of the interview recording to identify these times. In future research we recommend that researchers take note of video pause times.

It has been reported in previous research using VSR that some teacher participants were not willing to watch their video recordings in discussion with the researcher, because they were not satisfied with their behaviour (Rowe 2009). This was not the case in our study, although some teachers and students appeared initially uncomfortable watching themselves on video. However, their interest in reviewing the feedback conversations, especially for the teachers, appeared to overcome this reaction. Reviewing video evidence enabled students and teachers to observe reactions through facial features and body language. The visual evidence was seen to provide new information for reflection. During feedback conversations both teacher and student eyes were mostly focused on the student work or scoring rubric. Through use of VSR, the participants were able to focus on the social interaction.

The teachers' comments revealed their attention to multiple dimensions of the feedback conversation including characteristics of the feedback message, the manner of feedback delivery, their own and student characteristics and behaviour and the interaction of these variables (Winstone et al. 2017). The VSR allowed teachers and students to review both verbal and physical responses to the feedback. The teachers were critical of their own behaviour as feedback provider, especially when they identified that they had monopolised the conversation rather than allowing more time for student response, or when the characteristics of the feedback message mainly focused on correcting mistakes rather than also identifying strengths in the student's work. Replaying the video also enabled the teachers to more clearly identify the student's emotional reaction in the feedback interaction. Teachers' survey responses indicated that use of video technology independently to capture teacher-student interactions had potential to improve these interactions and enhance their feedback practice in general, because of an increased awareness of their current feedback practices. To a lesser degree, but still significant, most of the students articulated that they valued reviewing the feedback as it gave them the opportunity to hear again key points that were missed in quite densely informative conversations.

The teachers in the study commented that this form of one-to-one feedback was not common practice, though at the conclusion of the feedback session, they noted the value for student learning, and for their feedback and classroom teaching practice. Since the VSR was not difficult to set up for one-to-one feedback sessions, it would add minimal time to the feedback interaction, especially if the teacher was conducting organised conferencing sessions. The issue in this case is not related to the technology but rather the purposeful scheduling of time. Based on our findings, we hypothesise that the methods used in this study can be replicated by teachers to enhance classroom feedback practices. Engaging in VSR has the potential to enhance proactive recipience of feedback (Winstone et al. 2017) by making both teachers and students aware of their verbal and physical contributions and responses to the feedback interaction.

Extending the research from teacher-student interactions to capturing whole classroom interactions would require additional technologies. Previous video research on formative assessment in a whole classroom situation (Gotwals et al. 2015) used a camera at the back of the classroom and a cordless microphone worn by the teacher. In this type of set up the researchers reported not always being able to capture students' voices in student-student or whole class interactions when the teacher was not nearby. While the GoPro camera can be positioned at different angles, a similar issue identified by Gotwals et al. (2015) may be the inability of the GoPro microphone to pick up student conversations. Similarly, additional technology, for example, a Swivl^{TM} with several wireless markers with microphones, is required for the iPad to capture voices accurately when not in proximity to the camera. When iPads are positioned near students to capture conversations, the use of a fish eye lens for the iPad will enable participant images to be captured on screen.

18.5 Conclusion

This study trialled two technologies that are easy for schools to access regarding the quality of the sound and image and ease of use including battery life, in a range of educational contexts. The focus was teacher-student feedback interactions for the purpose of improving learning. The findings showed that while the iPad and the GoPro camera were both useful for capturing the feedback interactions, the iPad was more useful for video-stimulated recall interviews. With schools increasingly purchasing this type of technology, this finding suggests that this technology has potential for independent use by teachers and students for reflection to improve feedback practices. Through the use of these simple technologies, our project has been able to identify the relationship between teacher and student perceptions of feedback conversations. Most importantly, for our research purposes, it has potential for optimal data collection by enabling larger research data sets to be collected in a range of locations to illuminate the 'black box' (Black and Wiliam 1998) of classroom interactions and assist improvement of teaching and student learning. Additional technology is needed to upscale the use of the two devices to capture and facilitate reflection on whole-classroom interactions, but overall iPads seem more suitable for this type of use.

Appendix 1. Teacher questionnaire

To what extent do you agree or disagree with the following statements?

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1. The GoPro camera is a suitable device to capture teacher-student interactions	□	□	□	□	□
2. The iPad is a suitable device to capture teacher-student interactions	□	□	□	□	□
3. The length of the video observation was appropriate	□	□	□	□	□
4. The quality of the audio when replaying the video was good	□	□	□	□	□
5. The quality of the visuals when replaying the video was good	□	□	□	□	□

6. Did you experience any issues while being recorded by the GoPro or iPad?
7. Do you have any other comments?

References

Adie, L., Van der Kleij, F., & Cumming, J. (2018). The development and application of coding frameworks to explore dialogic feedback interactions and self-regulated learning. *British Educational Research Journal, 44,* 704–723. https://doi.org/10.1002/berj.3463.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7–74. https://doi.org/10.1080/0969595980050102.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21,* 5–31. https://doi.org/10.1007/s11092-008-9068-5.

Charteris, J., & Smardon, D. (2013). Second look—second think: A fresh look at video to support dialogic feedback in peer coaching. *Professional Development in Education, 39,* 1–18. https://doi.org/10.1080/19415257.2012.753931.

Crook, A., Mauchline, A., Maw, S., Lawson, C., Drinkwater, R., Lundqvist, K., et al. (2012). The use of video technology for providing feedback to students: Can it enhance the feedback experience for staff and students? *Computers & Education, 58,* 386–396. https://doi.org/10.1016/j.compedu.2011.08.025.

Gotwals, A. W., Philhower, J., Cisterna, D., & Bennett, S. (2015). Using video to examine formative assessment practices as measures of expertise for mathematics and science teachers. *International Journal of Science and Mathematics Education, 13,* 405–423. https://doi.org/10.1007/s10763-015-9623-8.

Gröschner, A., Seidel, T., Kiemer, K., & Pehmer, A. (2015). Through the lens of teacher professional development components: The 'Dialogic Video Cycle' as an innovative program to foster classroom dialogue. *Professional Development in Education, 41,* 729–756. https://doi.org/10.1080/19415257.2014.939692.

Hargreaves, E. (2012). Teachers' classroom feedback: Still trying to get it right. *Pedagogies: An International Journal, 7*, 1–15. https://doi.org/10.1080/1554480x.2012.630454.

Hargreaves, E. (2013). Inquiring into children's experiences of teacher feedback: reconceptualising assessment for learning. *Oxford Review of Education, 39,* 229–246. https://doi.org/10.1080/03054985.2013.787922.

Harlin, E.-M. (2014). Watching oneself teach—long-term effects of teachers' reflections on their video-recorded teaching. *Technology, Pedagogy & Education, 23,* 507–521. https://doi.org/10.1080/1475939X.2013.822413.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hummel, R. L. J. (2015). Teaching with a GoPro Camera! Simultaneously incorporate technology and learning while creating flipped classroom content. *International Conference of the Society for Information Technology & Teacher Education, 2015,* 1786–1787.

Hung, S. A. (2016). Enhancing feedback provision through multimodal video technology. *Computers & Education, 98,* 90–101. https://doi.org/10.1016/j.compedu.2016.03.009.

Jordan, L. (2012). Video for peer feedback and reflection: Embedding mainstream engagement into learning and teaching practice. *Research in Learning Technology, 20,* 16–25. https://doi.org/10.3402/rlt.v20i0.19192.

Lenters, K., & Grant, K. (2016). Feedback loops: Assembling student editors, stories, and devices for multimodal peer feedback. *Language Arts, 93,* 185–199.

Lyle, J. (2003). Stimulated recall: A report on its use in naturalistic research. *British Educational Research Journal, 29,* 861–878. https://doi.org/10.1080/0141192032000137349.

Murphy, D. (2016). A literature review: The effect of implementing technology in a high school mathematics classroom. *International Journal of Research in Education and Science (IJRES), 2*, 295–299. Retrieved from http://dergipark.gov.tr/download/article-file/231417.

Rowe, V. C. (2009). Using video-stimulated recall as a basis for interviews: Some experiences from the field. *Music Education Research, 11,* 425–437. https://doi.org/10.1080/14613800903390766.

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist, 35,* 87–100. https://doi.org/10.1207/S15326985EP3502_3.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. Abingdon: Routledge.

Tripp, T., & Rich, P. (2012). Using video to analyze one's own teaching. *British Journal of Educational Technology, 43,* 678–704. https://doi.org/10.1111/j.1467-8535.2011.01234.x.

Van den Bergh, L., Ros, A., & Beijaard, D. (2014). Improving teacher feedback during active learning: effects of a professional development program. *American Educational Research Journal, 51,* 772–809. https://doi.org/10.3102/0002831214531322.

Van der Kleij, F. M., Adie, L. E., & Cumming, J. J. (2017). Using video technology to enable student voice in assessment feedback. *British Journal of Educational Technology, 48,* 1092–1105. https://doi.org/10.1111/bjet.12536.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85,* 475–511. https://doi.org/10.3102/0034654314564881.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37,* 3–11. https://doi.org/10.1016/j.stueduc.2011.03.001.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52,* 17–37. https://doi.org/10.1080/00461520.2016.1207538.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 19
Game-Based Spoken Interaction Assessment in Special Need Children

Jos Keuning, Sanneke Schouwstra, Femke Scheltinga and Marleen van der Lubbe

Abstract The purpose of the present study was to explore whether it is possible to collect high-quality data about children's spoken interaction skills using the *Fischerspiel* board game as an entertaining, non-threatening means to evoke conversations between children in special elementary education. The game was administered to a total of 681 eleven- and twelve-year-old children with varying educational needs. The quality of the conversations between the children was evaluated with a specially designed observation form. The observation forms were filled in by trained test leaders and four independent expert raters. Video recordings showed that almost all children were willing to participate in the game, even the children who usually barely speak in class. Moreover, the game provided more than sufficient data to assess different dimensions of spoken interaction skills. Analyses further showed that the observation form functioned well and provided reliable scores. A group effect was nevertheless present and two test leaders deviated largely from the expert raters. These test leaders may have been insufficiently equipped to cope with the task. Application of Automatic Speech Recognition (ASR) technology in a (computer-based) spoken interaction assessment might ease the task and increase rating quality.

19.1 Introduction

Oral language performance is a good predictor of various school outcomes such as mathematics (Chow and Jacobs 2016; Fuchs et al. 2005) and reading and writing (Dickinson et al. 2010; Hart and Risley 2003; Kent et al. 2014; Nation and Snowling 2004; Dougherty 2014). Deficits in oral language skills may also underlie difficulties

J. Keuning (✉) · S. Schouwstra
Cito, Arnhem, The Netherlands
e-mail: jos.keuning@cito.nl

F. Scheltinga
University of Amsterdam, Amsterdam, The Netherlands

M. van der Lubbe
Inspectie van het Onderwijs, Utrecht, The Netherlands

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_19

in text comprehension and writing (Shanahan 2006; Naucler and Magnusson 2002). At school, language is necessary to understand instruction, to communicate about content and to demonstrate understanding. Moreover, language ability is related to social behavior and poor language skills affect social behavior negatively (Chow and Wehby 2018). Within the educational context, oral language is generally referred to as listening and speaking. Listening is as a receptive skill: it is the process of recognizing words and understanding messages expressed by others. Speaking is a productive skill: it is the process of transmitting ideas and information orally in a variety of situations. Both skills are used alternately in spoken interaction. Conversation partners must not only be able to convey and understand the content, but in addition, they must also be able to manage, produce and interpret verbal and nonverbal communication. While there is much experience in monitoring the listening and speaking skills of children, it is less clear how to monitor the children's spoken interaction skills. The present study focuses on an even more unexplored area: spoken interaction assessment in special need children.

19.1.1 Measuring Spoken Interaction Skills

Spoken interaction includes three content domains, namely Language form, Language content and Language usage (Bloom and Lahey 1978; Lahey 1988). Language form refers to the grammar of language, that is, to the phonological, morphological and syntactic rules that determine how sounds, words and sentences are formed, respectively. Language content refers to the ability to make up an own (non-)fictional story with a clear line of thought. Finally, in Language usage it is about the understanding and use of communicative functions and conversation rules such as expressing and interpreting emotion, making conversation, maintaining a topic of conversation, taking turns or asking others for information (McLaughlin 1998). When measuring spoken interaction skills it is important that all three content domains are covered. In addition, O'Malley and Pierce (1996) state that a spoken interaction assessment should: (a) test a child's competence as authentically as possible, (b) include an evidence model which shows how responses are scored, analyzed and interpreted, (c) take a limited amount of time to administer. In practice, it is difficult to meet all aforementioned criteria at the same time. Moreover, an extra complicating factor in assessing spoken interaction is the influence of the context and conversation partner. An interplay does exist between individual language ability and contextual factors (Chapelle 1998). Furthermore, the course of a conversation is determined in interaction and co-construction (Kramsch 1986).

Taylor and Galazci (2011) suggested to control context and to use detailed scripts to challenge children at their own level in a meaningful and standardized manner. These suggestions were implemented in a recent large-scale assessment in the Netherlands by having three children jointly conduct a conversation assignment (Van Langen et al. 2017). The assignment was conducted in the context of a national charity action: the children had to make up a goal and activity with which they could

Table 19.1 Spoken interaction assignments for one-to-one settings

1. Short situation sketch with an interaction between child and test leader + Reasonably standardized administration and scoring + Structured and therefore in line with the child's needs + Alignment on conversation partners and Taking turns are slightly covered – Children can suffice with short responses – It requires a lot from the test leader (or teacher) to respond in a standardized manner; outcomes are a bit uncertain
2. Role-play + Alignment on conversation partners and Taking turns are covered – According to teachers not suited for many special need children – Administration and scoring not standardized
3. Cartoon storytelling + Reasonably standardized administration and scoring + Elicits somewhat longer stories in most children – Alignment on conversation partners and Taking turns are not well covered – Appeals to the child's imagination and creativity – The performances in mainstream and special need education are often similar; this is not what is expected, validity might be an issue
4. Referential communication task + Reasonably standardized administration and scoring + Alignment on conversation partners and Taking turns are slightly covered – Appeals to specific concepts or words (e.g., colors, shapes, front/back/left/right etc.) – The child only gives an instruction, storytelling is not really part of the assignment

raise money. In order to finance the preparation of their idea they were awarded a (fictitious) amount of 50 euro. The children had to agree on the project planning and the corresponding financial framework. The topics to be discussed were shown on a paper for the children as a guide during the assignment. In addition, each child received his own conversation guide as input for the conversation. The conversation guide contained information that was specifically intended for that child. For example, each child was expected to have his or her own contribution at various points in the conversation. The conversation guides encouraged each child to actively contribute to the conversation at minimally two moments. Although not every child received information on all discussion topics, they always did have the opportunity to contribute. The child could, for example, give an opinion, give a reaction to the opinion of others, join this opinion or make a compromise. In this manner, the conversation could proceed as naturally as possible.

The charity assignment functioned well when administered to mainstream eleven- and twelve-year-old children (see Van Langen et al. 2017), but it is unlikely that the assignment is also appropriate for special need children. According to special education teachers, the context is, for instance, too abstract for children with lower cognitive abilities. Many other spoken interaction assignments are possible, all with specific advantages and disadvantages. Tables 19.1, 19.2 and 19.3 show which assignments can be administered in a one-to-one setting, a small-group setting or a classroom setting, respectively.

Table 19.2 Spoken interaction assignments for small-group settings

1. Group storytelling (e.g. about the weekend) + Alignment on conversation partners and Taking turns are covered dependent on the role of the child in the group − Some children may contribute a lot while others do not; the group affects the chance an individual child gets to show what he or she can do − Administration and scoring are not standardized
2. Group assignment (e.g. development of a project plan) + Alignment on conversation partners and Taking turns are covered dependent on the role of the child in the group − Some children may contribute a lot while others do not; the group affects the chance an individual child gets to show what he or she can do − Administration and scoring are not standardized − Not well suited for special need children

Table 19.3 Spoken interaction assignments for classroom settings

1. Lecture or book discussion + The child can show extensively what he or she can do − Alignment on conversation partners and Taking turns are not well covered − Administration not standardized
2. Observation in a natural situation such as a circle discussion + A familiar situation for the children − Administration and scoring are not standardized − Some children may contribute a lot while others do not − Not all children may feel safe in a circle discussion

19.1.2 Assessment in Special Elementary Education

At this moment, it is unclear which assignment should be preferred in which situation. The assignment must, however, account for the specific characteristics of the children in special elementary education. Special elementary education is meant for children who need orthopedagogical and ortho-didactical support. In the Netherlands, almost all special need children experience multiple problems in relation to learning, behavior or development. The children show problematic internalizing or externalizing behavior, for example, or have a communication problem or intellectual disability (see, for example, Ledoux et al. 2012). It is the accumulation of problems that makes it difficult, and sometimes impossible, for these children to follow education in a regular setting. A few considerations apply when developing an assignment or test for special need children. First, the children's special needs are generally better taken into account in practical assignments than in paper-and-pencil multiple-choice tests. Second, due to the limited attention span of many special education children, assignments should be short and varied. Third, assessments should engage the children and give them a feeling of success as many already encountered multiple failure experiences during their school career. Finally, the children in special education show great diversity and some children require (additional) adjustments to assessment practices in order to demonstrate what they know and can do. A protocol with an overview of allowable adjustments is required to cater for the characteristics of the children being assessed.

An assessment in special elementary education also requires specific choices with regard to context, layout and use of language (see, for example, Cito 2010). Contexts must be meaningful and connect to the children's experiences, for instance, and potentially 'provocative' contexts such as a football match must be avoided at all times. Images must have a high contrast and be available to the children in black-and-white or enlarged format. Moreover, the images must be 'real' and support the assignment; talking animals or purple frogs are not appropriate, for example. Finally, language should be as concrete as possible. Negative sentences, imagery and emotionally charged sentences such as 'stop doing that' should be avoided, just as I-sentences, long compound sentences and complex cause-and-effect relationships. Although such guidelines also apply to children in regular elementary education to some degree, they are particularly important in special elementary education. Especially the children with special educational needs will perform best in familiar situations without pressure. In general, the children will naturally apply many of their listening and speaking skills in free circumstances. Data about the children's skills can then be collected best by observation; it does not bother the children with an assessment and they can show their skills in a familiar setting without 'explicitly knowing' that their skills are being monitored and documented.

19.1.2.1 The Present Study

In recent literature, game-based assessment has been developed as a more fun and accessible way for children to assess their knowledge and skills. Games are well suited for assessment purposes as they naturally present children with a stream of choices during gameplay. All these choices can be recorded and it is also possible to record how the children arrived at their choice (Stieger and Reips 2010). This allows game-based assessments to capture information that often cannot be captured by traditional paper-and-pencil assessments (Shute and Ventura 2013; Landers 2014). Moreover, it is relatively easy to create authentic and familiar situations as children play games every day. However, especially in special elementary education it should not be a battle against each other; solving the game together should be the objective. Against this background, a game for the assessment of spoken interaction skills was developed. The game was based on the *Fischerspiel*. This is essentially a board game which will be described in more detail below. An observation form was further developed in order to assess the children's spoken interaction skills during gameplay.

Observation as a measurement strategy has, despite numerous advantages, certain unique limitations (Michaels 1983), such as imposed limitations on the types of behavior observed, problems with the category systems, observer bias and interferences. In this study it was examined whether such limitations occurred in the games-based assessment for spoken interaction. Different aspects of the game, the observation form and the test leader were evaluated. The first objective was to evaluate whether particular game characteristics were a source of invalidity. Messick (1995) distinguished two sources of invalidity: underrepresentation and irrelevant variance. When an assessment fails to include important aspects of the skill the assessment suffers from underrepresentation. When an assessment contains excess variance associated with other aspects than the skill of interest the assessment is hampered by irrelevant variance. Both sources of invalidity were studied: the variety in the children's conversations was mapped, and in addition, it was examined whether the group and turn-taking affected the spoken interaction skills that individual children showed. The quality of the observation form was evaluated next. It was examined whether the assessment covered the relevant aspects of spoken interaction, and moreover, scale dimensionality was examined via exploratory factor analysis. Finally, the third important aspect of the assessment was considered: the test leader. The quality and reliability of the test leaders' evaluations were mapped by comparing the test leader ratings to an expert rating. The overall quality of the ratings was assessed and it was attempted to identify extreme or deviating ratings.

19.2 Method

19.2.1 Participants

A total of 681 eleven- and twelve-year-old children from 33 different special education schools in the Netherlands participated in the study. A two-fold stratification procedure was used to select the schools. Region was used as explicit stratification criterion: all Dutch special education schools were classified by region (North, East, South and West) and then a separate sample was drawn for each of the regions, so that the relative share of each region in the sample was representative of the relative share in the population of Dutch special education schools. School size was used as implicit stratification criterion: within each region the schools were organized from small to large and then, after generating a random start point, every *k*th school on the list was selected, so that both smaller and larger schools were included in the sample. No exclusion criteria were used for drawing the school sample. Within each school all children in the final grade (eleven- and twelve-year-olds) were expected to participate. Children with specific language impairment, hearing problems, selective mutism or aphasia were excluded from the study and also the children who lived in the Netherlands for less than 2 years were not eligible to participate. The sample consisted of 423 boys (62%) and 258 girls (38%) at varying educational levels. It was expected that after elementary school about 7% of the children would move on to General Secondary Education or higher. The other children would expected to move on to either Preparatory Vocational Education (49%) or Special Secondary Education (44%). These percentages are in line with the Dutch special education school population. More boys than girls attend Dutch special education and only a very small percentage moves on to the higher levels of Dutch secondary education.

19.2.2 Materials

The children's spoken interaction skills were assessed with an existing board game from Germany; the *Fischerspiel*. The game is played on a board with an island with several harbors and a sea. Each player has his or her own harbor and a colored boat to transport fish from the sea to the harbor. The aim of the game is to work together to bring all fish to the harbor before the wind reaches strength 12. When a player gets a turn, he throws a special die and consults his fellow players to determine who can best use the thrown number to get a fish and bring it to one of the harbors on the island. Players win together if all the fish are on the island. There is also a wind symbol on the die, however, and rolling the wind symbol increases the strength of the wind by 1. When the wind reaches strength 12, all boats sink and the game is lost. The quality of the conversations between players was evaluated with a specially designed observation form. The form included seventeen performance aspects. Each performance aspect was presented with three indicators: poor basic proficiency (0); fair proficiency (1) and good basic proficiency (2). Below the seventeen indicators of a good basic proficiency level are presented:

1. The child's conversations with the group are meaningful and relevant.
2. The child regularly takes the initiative to start, continue or stop a conversation.
3. The child usually takes the floor in an appropriate way.
4. The child integrates contributions from the group into his own contribution when relevant.
5. The child takes the initiative to achieve a joint communication goal by involving the group in the conversation.
6. The child makes his way of thinking understandable.
7. The child consistently uses language that fits the situation.
8. The non-verbal behavior of the child strengthens his verbal message.
9. The child shows adequate active listening behavior.
10. The child consistently gives appropriate verbal and nonverbal responses.
11. The child's contribution shows sufficient variation in word use
12. The child's vocabulary is sufficient to hold a conversation.
13. The child speaks fairly fluently with only occasional hitch, false starts or reformulation.
14. The child's pronunciation, articulation and intonation make the child's speech intelligible, despite a possible accent.
15. The child conjugates verbs correctly.
16. The child uses (combinations with) nouns correctly.
17. The child generally constructs correct simple, complex and compound sentences.

The observation form was a reflection of the Dutch reference framework for spoken language (Meijerink 2009). At the basic level of spoken language proficiency (1F) it is, for instance, expected that the child recognizes conversation situations and can use appropriate routines to give instruction or exchange information. At the highest level (4F) it is expected that the child is able to participate in casual, formal, and extended conversations on practical and academic topics. Language levels 1F and 2F apply to (special) elementary education.

19.2.3 Procedure

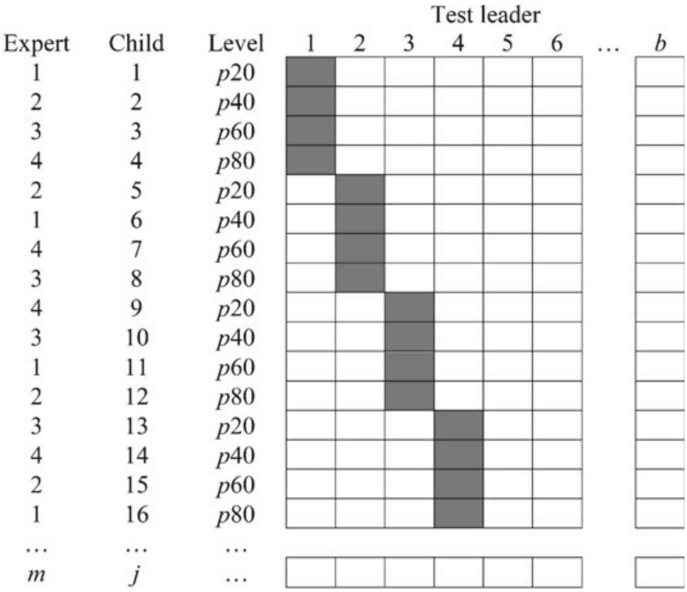
Administration of the *Fischerspiel* board game took place in a small and quiet, relatively stimulus-free room. The game was played in groups of three to four children. The groups were assembled randomly, but if a combination of children was inconvenient according to the teacher, a small change in the composition of the group was allowed. A quarrel during the break could, for example, be a reason to place a child in another group. Each administration was supervised by a test leader. The test leader did not participate in the game but acted as coach. The test leader monitored the course of the game and ensured that all the children got an equal number of turns and felt safe. In addition to a coaching role, the test leader also fulfilled the role of assessor during the administration. The observation form was filled in for each child separately after three rounds of the game. Try-outs showed three playing rounds to be more than sufficient to get an idea of the children's spoken interaction skills. Moreover, the children generally could play the game independently after three rounds, giving the test leader time to fill in the forms. In order to ensure that the test leaders could conduct the assessment task as reliably as possible, the following four measures were taken:

1. Each performance indicator was elaborated with one or more examples.
2. The test leaders received an extensive training on the use of the assessment form.
3. Each administration was recorded on video, so that the test leader had the possibility to complete or check the assessment afterwards.
4. Questions about the assessment and dilemmas could be presented to other test leaders in a WhatsApp group.

The administration of the *Fischerspiel* board game took approximately 30 min, depending on the course of the game. To prevent potential group effects and effects of turn-taking order the following was done:

(a) Children were randomly assigned into groups.
(b) There was a starting round and each child had several turns; at a certain moment it is unlikely that the children still know who exactly started.
(c) The child who had the turn always had to take the initiative, but other players had the possibility to respond; there was no fixed order.

After completion of the administrations, a selection of children were re-assessed by a subject-area expert. The re-assessment was conducted in an incomplete design which was specifically developed to efficiently detect aberrant rating behavior. The design assumed that there were b, $b = 1,\ldots, B$, test leaders and m, $m = 1,\ldots, M$, expert assessors. From each test leader b a total of J children were selected (1 per



			Test leader							
Expert	Child	Level	1	2	3	4	5	6	…	*b*
1	1	*p*20	▨							
2	2	*p*40	▨							
3	3	*p*60	▨							
4	4	*p*80	▨							
2	5	*p*20		▨						
1	6	*p*40		▨						
4	7	*p*60		▨						
3	8	*p*80		▨						
4	9	*p*20			▨					
3	10	*p*40			▨					
1	11	*p*60			▨					
2	12	*p*80			▨					
3	13	*p*20				▨				
4	14	*p*40				▨				
2	15	*p*60				▨				
1	16	*p*80				▨				
…	…	…								
m	*j*	…								

Fig. 19.1 Schematic representation of the design used to examine rater reliability

group) and each expert assessor *m* re-assessed $B \times J$ children. The children *j* were selected on the basis of their percentile rank in order to ensure that both low and high ability children were re-assessed. A total of 16 test leaders was involved in this study, and four different subject-area experts all re-assessed one child per test leader. This means that a total of 64 children (16 × 4) were re-assessed by one of the four subject-area experts. Figure 19.1 gives a schematic representation of the design. As soon as the re-assessments were conducted, difference scores were calculated for each performance indicator by subtracting the expert rating score form the test leader rating score. The difference scores were then the basic observation in the analysis.

19.2.4 Statistical Analyses

Analyses within the framework of Classical Test Theory were conducted to answer the first research question. First the distribution of total scores was examined and then for each of the seventeen performance aspects (items) the *p*-value and r_{it}-value was computed. The *p*-value was computed as the ratio between the mean score and the maximum achievable score. Values between 0.500 and 0.700 can be considered optimal (Crocker and Algina 1986; Feldt 1993), but lower (>0.100) and higher values (<0.900) might be acceptable dependent on item type and purpose of the test. The r_{it}-value is the correlation between the item scores and total scores. Values below 0.190 indicate that the item does not discriminate well, values between 0.200 and

0.290 indicate sufficient discrimination, and values of 0.300 and above indicate good discrimination (Ebel and Frisbie 1991). Although the Classical Test Theory analyses can easily be conducted, the manner of administration may distort results. Separate analyses were therefore conducted to examine whether group or turn order effects were present or not. Classical Test Theory analyses were conducted separately for the first, second, third and fourth group member, and by means of a three-level regression analysis the proportion of variance explained by group membership was estimated.

To answer the second research question, the matrix of polychoric correlations between the seventeen performance aspects was visually inspected by means of a correlogram. After inspection of the correlogram, an exploratory Principal Axis Factor Analysis with Varimax rotation was conducted. In order to choose the number of factors well-reasoned, we started with a parallel analysis as proposed by Horn (1965): a simulation-based method in which essentially a random dataset is generated with the same number of items and exactly the same score range. The eigenvalues of the items in this random simulated dataset are then compared with the eigenvalues of the items in the actual dataset. All factors with an eigenvalue larger than the random (simulated) eigenvalues were retained in the factor analysis. The viability of the factor solution was assessed in light of the Dutch reference framework for spoken language and the conceptual framework by Bloom and Lahey (1978) and Lahey (1988).

The third research question was answered by examining the reliability and quality of the rating scores. Reliability was estimated in terms of the Greatest Lower Bound and Guttman's Lambda2 (Sijtsma 2009; Ten Berge and Sočan 2004). Coefficients higher than 0.800 were considered to be good and coefficients below 0.700 to be insufficient. The differences between the expert rating scores and the test leader rating scores were used to evaluate the quality of the rating. The lack of agreement with the norm (i.e., the expert rating scores) was mapped for each of the test leaders by computing the Mean Absolute Error (MAE):

$$MAE_b = \frac{\sum_j \sum_i \left| s_{bji} - s_{mji} \right|}{N_j},$$

where s_{bji} and s_{mij} are the rating scores of test leader b and expert m, respectively, for child j on item i and N_j the number of ratings on child j by test leader b and expert m. The Median Absolute Deviation (MAD) was used as measure for detecting aberrant rating behavior. To optimally account for a possible asymmetric distribution of **MAE**, the median absolute deviation from the median was based on all points greater than or equal to the median: $MAD = Mdn(|\mathbf{Y} - Mdn(MAE)|)$, where $\mathbf{Y} = \{MAE_b \in \mathbf{MAE} : MAE_b \geq Mdn(\mathbf{MAE})\}$. Given this distance, a test leader b was marked as outlier if:

$$\frac{MAE_b - Mdn(\mathbf{MAE})}{MAD} > 2.5$$

Threshold value 2.5 was suggested by Leys et al. (2013) but other values are possible. The overall quality of the ratings was finally assessed by computing

Cohen's weighted kappa coefficient (κ) and Gower's similarity coefficient (G_{xy}). Cohen's kappa was interpreted as follows: κ < 0.200 poor; 0.200 < κ < 0.400 fair; 0.400 < κ < 0.600 moderate; 0.600 < κ < 0.800 good; κ ≥ 0.800 excellent. Gower's similarity coefficient was considered low if $G_{xy} < 0.650$, acceptable if $0.650 \leq G_{xy} < 0.800$, and high if $G_{xy} > 0.800$.

19.3 Results

The distribution of total scores is visually presented on the left-hand side in Fig. 19.2. As can be seen, the score distribution was slightly skewed to the left (−0.814) and had fatter tails than a normal distribution (3.155); the sample mean was 25.790 with a standard deviation of 6.260. On the average, children obtained about three quarters of the total number of points that could maximally be achieved (25.790 ÷ 34). Further analysis showed that the seventeen items functioned quite similarly. As can be seen from Table 19.4, the *p*-values varied from 0.596 to 0.916 and the r_{it}-values were all higher than 0.300. Although the size of the r_{it}-values is related to the manner in which the items were scored (i.e., a three-point Likert scale instead of dichotomous correct-incorrect scores), the r_{it}-values indicated the items to discriminate very well. Columns c1, c1, and c2 show the percentage of children with scores 0, 1 and 2 respectively. The number of children with a zero score was remarkably low for some items, especially for those related to fluency, comprehensibility or grammar mastery. Table 19.4 nevertheless shows that the items in the observation form were all appropriate for distinguishing children with weaker spoken interaction skills from children with better spoken interaction skills. There were no reasons to drop items from the observation form or to merge score categories.

The Classical Test Theory analyses were repeated for the first, second, third and fourth group member separately in order to examine whether turn order effects were

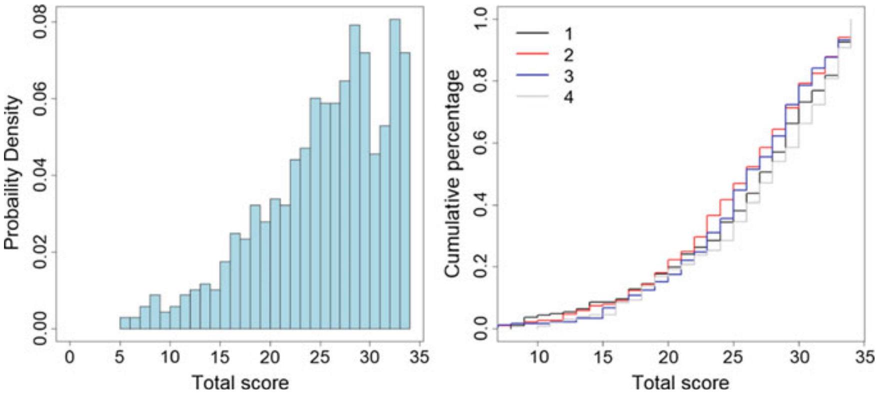


Fig. 19.2 Score distribution for the entire sample (left) and each group member (right)

Table 19.4 Results analyses at item level

Item	c0	c1	c2	p	r_{it}
1. The child's conversations with the group are meaningful and relevant	0.157	0.493	0.349	0.596	0.725
2. The child regularly takes the initiative to start, continue or stop a conversation	0.157	0.441	0.402	0.623	0.575
3. The child usually takes the floor in an appropriate way	0.098	0.276	0.626	0.764	0.539
4. The child integrates contributions from the group into his own contribution when relevant	0.072	0.266	0.662	0.795	0.538
5. The child takes the initiative to achieve a joint communication goal by involving the group in the conversation	0.164	0.446	0.389	0.612	0.638
6. The child makes his way of thinking understandable	0.116	0.382	0.502	0.693	0.741
7. The child consistently uses language that fits the situation	0.035	0.305	0.659	0.812	0.613
8. The non-verbal behavior of the child strengthens his verbal message	0.072	0.606	0.322	0.625	0.520
9. The child shows adequate active listening behavior	0.106	0.461	0.433	0.664	0.679
10. The child consistently gives appropriate verbal and nonverbal responses	0.110	0.285	0.605	0.747	0.574
11. The child's contribution shows sufficient variation in word use	0.070	0.256	0.674	0.802	0.729
12. The child's vocabulary is sufficient to hold a conversation	0.031	0.223	0.746	0.858	0.698
13. The child speaks fairly fluently with only occasional hitch, false starts or reformulation	0.040	0.270	0.690	0.825	0.584
14. The child's pronunciation, articulation and intonation make the child's speech intelligible, despite a possible accent	0.040	0.197	0.764	0.862	0.589
15. The child conjugates verbs correctly	0.013	0.186	0.800	0.894	0.534
16. The child uses (combinations with) nouns correctly	0.006	0.157	0.837	0.916	0.532
17. The child generally constructs correct simple, complex and compound sentences	0.016	0.348	0.636	0.810	0.617

present or not. The right-hand side of Fig. 19.2 shows the empirical cumulative distributions for the different group members. The cumulative distributions were not exactly the same but in light of the sample sizes and the unsystematic ordering of the distributions there was also no reason to conclude that children were disadvantaged if they were player two, three or four. A multilevel regression analysis with children nested in schools, score as dependent variable and group member number as predictor confirmed this conclusion: (member 2-1) $\beta = -0.757$, $z = -1.260$; (member 3-

1) $\beta = -0.272$, $z = -0.440$; and (member 4-1) $\beta = 0.699$, $z = -1.040$. Also, the four analyses at item level showed similar results for the four group members. For example, the p-values differed only 0.052 points on average and the r_{it}-values maximally differed 0.160. A three-level regression analysis without predictors further showed that children within groups were more similar than children across groups. The proportion of explained variance at group level was 0.127. A group effect was thus present in the data, and therefore, there may be occasion to account for group in some analyses. That was not well possible in the present study, however, due to the very small number of groups per school.

Dimensionality was investigated next by presenting the polychoric inter-item correlations in a correlogram. In Fig. 19.3, all correlations are represented by means of a color: darker blue means a higher positive correlation and darker red means a larger negative correlation. As can be seen, all items were positively correlated to each other. The theoretical dimensions of spoken language, however, cannot easily be found. A parallel analysis was therefore conducted in order to determine the number of factors to retain from factor analysis. The results showed that three factors should be retained. This suggestion was adopted. The rotated (pattern) matrix with loadings below 0.300 suppressed is reported in Table 19.5. H2 and U2 represent the communality and specific variance, respectively, of the standardized loadings obtained from the correlation matrix. The communalities were at or above 0.400, except for one just below that value, indicating shared variance with other items. The primary factor loadings were generally above 0.600 and the gap between primary factor loadings and each of the cross-loadings was almost always at least 0.200. Almost no cross-loading was above 0.300, moreover, further indicating that the structure with three underlying factors has a satisfactory fit. Together, the three factors explained 68% of the variance in the items, with factors 1–3 contributing 27, 21 and 20%, respectively. The three factors are in keeping with the three theoretical dimensions of spoken interaction, namely Language form, Language usage and Language content. Some complex cross-factor loadings were nevertheless also present. Especially items 4 (The child integrates contributions from the group into his own contribution when relevant) and 9 (The child shows adequate active listening behavior) did not contribute to one specific factor. Whereas these items theoretically most likely appeal to the social component of conversations, the factor analysis clearly suggested that these items also appeal to substantive quality.

Finally, the reliability and quality of the rating scores was examined. The Greatest Lower Bound was equal to 0.952 and Guttman's Lambda2 was equal to 0.899. These values indicate a very high reliability, but with these values it is not guaranteed that the assessments were also adequate. Therefore, the quality of the rating scores was examined next by comparing the test leader rating scores to an expert rating. Figure 19.4 shows the Mean Absolute Error (MAE) for each test leader. The bottom grey dotted line is the median of the MAE's for the test leaders.The top grey dotted line is the median of the MAE's for an infinitely large number of random assessments. As can be seen, the rating scores for test leaders 1, 4 and 9 were very similar to the ratings scores of the subject-area experts. The rating scores of test leaders 2, 6 and 12, on the other hand, were quite different from the expert rating scores. The MAE

Fig. 19.3 Correlogram of the matrix with polychoric inter-item correlations

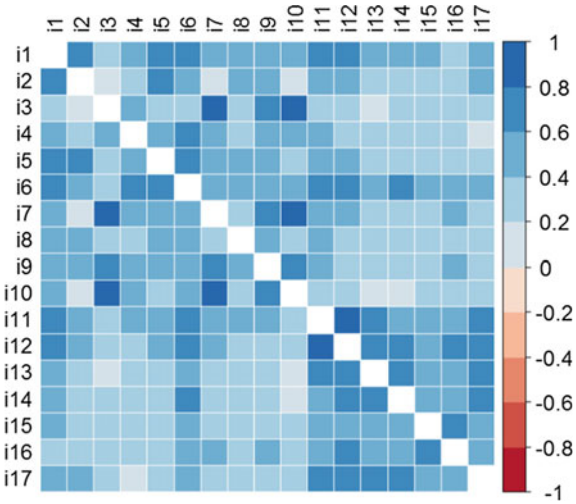


Table 19.5 Results three factor principal axis factor analysis

Item	Factor 1: Language form	Factor 2: Language usage	Factor 3: Language content	H2	U2
3. The child usually takes the floor in an appropriate way		0.910		0.870	0.130
4. The child integrates contributions from the group into his own contribution when relevant		0.470	0.450	0.450	0.550
7. The child consistently uses language that fits the situation		0.840		0.810	0.190
9. The child shows adequate active listening behavior		0.610	0.530	0.700	0.300
10. The child consistently gives appropriate verbal and nonverbal responses		0.900		0.860	0.140
1. The child's conversations with the group are meaningful and relevant.	0.440		0.660	0.700	0.300
2. The child regularly takes the initiative to start, continue or stop a conversation			0.830	0.760	0.240
5. The child takes the initiative to achieve a joint communication goal by involving the group in the conversation			0.840	0.790	0.210
6. The child makes his way of thinking understandable.	0.490		0.620	0.710	0.290
8. The non-verbal behavior of the child strengthens his verbal message			0.550	0.390	0.610

(continued)

Table 19.5 (continued)

Item	Factor 1: Language form	Factor 2: Language usage	Factor 3: Language content	H2	U2
11. The child's contribution shows sufficient variation in word use	0.660		0.480	0.720	0.280
12. The child's vocabulary is sufficient to hold a conversation	0.810		0.340	0.800	0.200
13. The child speaks fairly fluently with only occasional hitch, false starts or reformulation	0.800			0.680	0.320
14. The student's pronunciation, articulation and intonation make the student's speech intelligible, despite a possible accent	0.770			0.660	0.340
15. The child conjugates verbs correctly	0.700			0.540	0.460
16. The child uses (combinations with) nouns correctly	0.740			0.630	0.370
17. The child generally constructs correct simple, complex and compound sentences	0.720			0.620	0.380

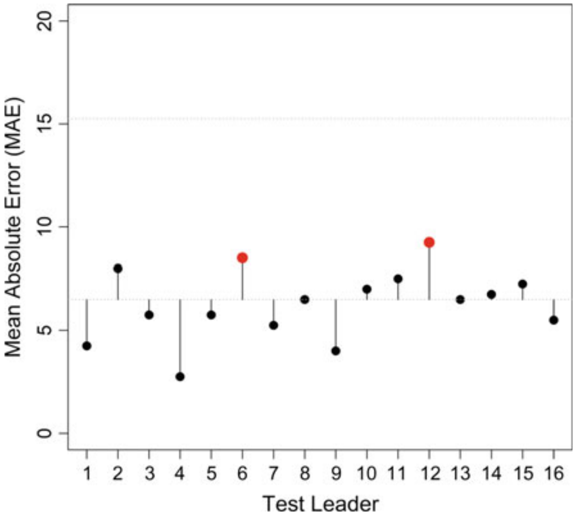


Fig. 19.4 Mean absolute error per test leader

for test leaders 6 and 12 was even so large that their rating behavior can be considered aberrant in comparison with the other 12 test leaders; the MAD for these test leaders was larger than 2.5. Cohen's weighted kappa coefficient (κ) and Gower's similarity coefficient (G_{xy}) together indicated a fair overall rating quality: $\kappa = 0.307$ and $G_{xy} = 0.815$, where absolute agreement is remarkably higher than relative agreement. On average it did not statistically matter whether the assessment was done by a test leader ($M = 27.078$, $SD = 4.487$) or a subject-area expert ($M = 26.328$, $SD = 5.252$); $t(126) = -0.869$, $p = 0.387$.

19.4 Conclusions and Discussion

In the present study, the *Fischerspiel* board game was used as an entertaining, non-threatening means to evoke conversations between children in special elementary education. Spoken interaction was observed and rated using a newly developed observation form with seventeen performance aspects. It was first examined whether the conversations during the game were sufficiently varied to assess the children's spoken interaction skills. In addition, it was examined whether particular characteristics of the board game were a source of invalidity. When the board game would have elicited very limited or only highly similar conversations, irrespective of the children's skills level, the assessment would, for instance, fail to reveal the differences in skill between children. Sufficient variation was present, however, and the different performance indicators also turned out to function well. The *p*-values were in an acceptable range and all performance indicators had a good discrimination. The performance indicators thus discerned well between children with poor spoken interaction skills and children with good spoken interaction skills. It can therefore be concluded that the board game elicited varied conversations between children and that all aspects of basic spoken interaction proficiency (1F) were observable and assessable. Thus, we can conclude that the assessment did not suffer from underrepresentation of the target skill.

Whether the board game imposed limits that cause (skill) irrelevant variance was evaluated next. Turn taking is one potential source of irrelevant variance, as it might cause differences between children even if their true performance level is equal, but in this study, the order in which children took turns (i.e., first, second, third or fourth in the row) did not significantly affect performance. However, a group effect was found. Analyses showed that the children's performance within groups was more similar than the children's performance across groups. This finding is consistent with several studies on paired test settings where a so-called interlocutor effect was evidenced quite often. Many studies, that is, reported that low skilled children performed better when paired with high skilled children (see, for example, IlSun 2017). More research is needed to study whether the ability of the group members indeed causes differences. At the same time, one should be cautious in using a single paired or group setting in (high-stake) assessments for individual decisions. In high-stakes assessment each child should preferably play with different groups during the observation. For a survey, a group effect might be less problematic as findings are aggregated across groups.

After studying the characteristics of the game the quality of the observation form and the influence of the test leader was considered. The performance aspects provided reliable scores, but the high reliability might in part be caused by the aforementioned group effect. It might be that the test leaders were not able to observe differences between the children within one group really well. A high similarity of the evaluations within groups automatically yields a higher reliability coefficient. As could be expected from theory (see Bloom and Lahey 1978; Lahey 1988), the performance aspects did appeal to three different dimensions: Language form, Language usage and Language content. The first dimension contained all performance aspects that related to grammar, intelligibility and vocabulary. The second dimension contained the performance aspects that related to interaction and alignment on conversation partners. The third dimension contained the performance aspects that related to the quality of the conversation. Finally, the agreement between the expert rating scores and the test leader rating scores turned out to be reasonable. However, two out of sixteen test leaders displayed evaluations that were quite different from the experts' evaluations. The used methodology allows for an early detection of aberrant behavior and with such a methodology a timely intervention is also possible. The test leader could receive extra training, for instance, or some evaluations could be conducted again. This, however was not feasible in the present study, as the experts evaluated the children's spoken interaction skills afterwards from videos. A computer-based assessment in which the test leader and expert can do the evaluation at the same time would speed up the process and potentially prevent test leader effects.

To conclude, the *Fischerspiel* board game proved to be a promising entertaining and non-threatening way of assessing children's spoken interaction skills. Play is important for learning (Mellou 1994), play can be used for learning (so-called serious games, Abt 1970), and play is informative about learning (Otsuka and Jay 2017). Special need children were the target group in the present study, and given the learning obstacles these children encounter, it was crucial to develop an assessment that was practical, short and varied, and would give a feeling of success. The *Fischerspiel* met these criteria, but clearly, also children in regular education could use an assessment with such characteristics. The application of the game as an assessment instrument should therefore also be studied in regular education. Problems associated with observation should then receive particular attention. A computer or online version of the *Fischerspiel* board game might help to overcome some problems. It is then easier to have children playing the game in different groups, the observation can be conducted more unobtrusively and aberrant rating behavior can be detected much faster. Another potential advantage is that in computer games automatic speech recognition technology (ASR) might be used to aid the evaluation. For instance, Ganzeboom et al. (2016) recently developed a serious ASR-based game for speech quality. Such developments are very promising and should therefore certainly be considered when further developing (observer-bias free) assessments for spoken interaction skills. Until then, games like the *Fischerspiel* are a nice alternative.

References

Abt, C. (1970). *Serious games*. New York: Viking Press.

Bloom, L., & Lahey, M. (1978). *Language development and language disorders*. New York: Wiley.

Chapelle, C. A. (1998). Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning & Technology, 2,* 22–34.

Chow, J. C., & Jacobs, M. (2016). The role of language in fraction performance: A synthesis of literature. *Learning and Individual Differences, 47,* 252–257.

Chow, J. C., & Wehby, J. H. (2018). Associations between language and problem behavior: A systematic review and correlational meta-analysis. *Educational Psychology Review, 30,* 61–82.

Cito. (2010). *Checklist toetsconstructie Speciaal (basis)onderwijs*. Arnhem: Cito.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Dickinson, D., Golinkoff, R. M., & Hirsh-Pasek, K. (2010). Speaking out for language: Why language is central to reading development. *Educational Researcher, 4,* 305–310.

Dougherty, C. (2014). Starting off strong: The importance of early learning. *American Educator, 38,* 14–18.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NY: Prentice Hall.

Feldt, L. S. (1993). The relationship between the distribution of item difficulties and test reliabilility. *Applied Psychological Measurement, 6,* 37–49.

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97,* 493–513.

Ganzeboom, M., Yılmaz, E., Cucchiarini, C. & Strik, H. (2016). An ASR-based interactive game for speech therapy. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, San Francisco, CA, USA, Sept 2016.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap. *American Educator, 27,* 4–9.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

IlSun, H. (2017). The effects of paired partners' proficiency levels on test-takers' speaking test performance. *Journal of Research in Curriculum & Instruction, 21*(2), 156–169.

Kent, S., Wanzek, J., Petscher, Y., Al Otaiba, S., & Kim, Y. (2014). Writing fluency and quality in kindergarten and first grade: The role of attention, reading, transcription, and oral language. *Reading and Writing: An Interdisciplinary Journal, 27,* 1163–1188.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70,* 366–372.

Lahey, M. (1988). *Language disorders and language development*. New York: MacMillan.

Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & Gaming, 45,* 752–768.

Ledoux, G., Roeleveld, J., Langen, A., & van Smeets, E. (2012). *Cool Speciaal. Inhoudelijk rapport* (Rapport 884, projectnummer 20476). Amsterdam: Kohnstamm Instituut.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49,* 764–766.

McLaughlin, S. (1998). *Introduction to language development*. Londen: Singular Publishing Group.

Meijerink, (2009). *Referentiekader taal en rekenen*. Enschede: SLO.

Mellou, E. (1994). Play theories: A contemporary review. *Early Child Development and Care, 102*(1), 91–100.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry to score meaning. *American Psychologist, 50,* 741–749.

Michaels, J. (1983). Systematic observation as a measurement strategy. *Sociological Focus, 16*(3), 217–226.

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading, 27,* 342–356.

Naucler, K., & Magnusson, E. (2002). How do preschool language problems affect language abilities in adolescence? In F. Windsor & M. L. Kelly (Eds.), *Investigations in clinical phonetics and linguistics* (pp. 99–114). Mahwah, NJ: Erlbaum.

O'Malley, J. M., & Pierce, L. V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. New York: Addison-Wesley.

Otsuka, K., & Jay, T. (2017). Understanding and supporting block play: Video observation research on preschoolers' block play to identify features associated with the development of abstract thinking. *Early Child Development and Care, 187*(5–6), 990–1003.

Shanahan, T. (2006). Relations among oral language, reading and writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 171–183). New York: The Guilford Press.

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120.

Stieger, S., & Reips, U. D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior, 26,* 1488–1495.

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking*. Cambridge, UK: Cambridge University Press.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613–625.

Van Langen, A., Van Druten-Frietman, L., Wolbers, M., Teunissen, C., Strating, H., Dood, C., et al. (2017). *Peilingsonderzoek Mondelinge Taalvaardigheid in het basisonderwijs*. Nijmegen: KBA.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 20
The Theory and Practice of Educational Data Forensics

Sebastiaan de Klerk, Sanette van Noord and Christiaan J. van Ommering

Abstract Although the testing community has been trying to prevent test fraud through multiple practices and methods (e.g., applying strict security practices or testing protocols), test fraud is still a ubiquitous problem. Exact numbers are unknown, but self-report studies show that up to 85% of students admit to committing test fraud at least once during their school career (Lin and Wen, Higher Educ 54:85–97, 2007; Hughes and McCabe, Can J Higher Educ 36:1–12, 2006; Berkhout et al., Studie and Werk 2011. SEO Economisch Onderzoek, Amsterdam, 2011). Research on the statistical detection of test fraud, also called educational data forensics (EDF), already exists since the 1920s (Bird, School Soc 25:261–262, 1927), but the body of research started growing considerably since the 1970s (e.g., Angoff, J Am Stat Assoc 69:44–49, 1974). Nowadays, many methods and models are presented in the literature. Two of those models are the Guttman error model (Guttman, Am Soci Rev 9(2):139–150, 1944; Meijer, Appl Psychol Measur 18(4):311–314, 1994) and the log-normal response time model (Van der Linden, Psychometrika 80(3):689–706, 2006). In the first part of this chapter, both models will be discussed. In the second part of this chapter, an empirical study on the functioning of the Guttman- and response time model will be presented. In the final part of the chapter, the design, development and validation of a protocol on the use of EDF will be presented.

20.1 Educational Data Forensics

Educational data forensics (EDF) is the statistical analysis of test takers' response data with the aim of the detecting aberrances that potentially indicate test fraud (Wollack and Fremer 2013). In this chapter EDF is discussed from several angles. We will start with a short historical context and introduction to the most used EDF methods and models, followed by a discussion on how EDF can be used in practice. In part two and three of this chapter, we will present two empirical studies on EDF.

S. de Klerk (✉) · S. van Noord · C. J. van Ommering
eX:plain, Amersfoort, The Netherlands
e-mail: s.dklerk@explain.nl

© The Author(s) 2019

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_20

20.1.1 Historical Context, Theory, and Practice of EDF

Test fraud can be defined as deliberately influencing the test process with the aim of improving the results on one or more tests. This puts different methods of committing test fraud within a broad spectrum. Less harmful methods like cheat sheets or test peeking on the one hand of the spectrum and more harmful methods like item theft or identity fraud on the other hand of the spectrum. The administration of a test yields test taker response data that can be used to detect potential test fraud. The statistical methodology used to this end is part of the EDF domain. Although most of the current methods are exclusively applicable to computer-based test (CBT), the first methods discussed in the literature were used to detect test fraud on paper-based tests (PBT). For example, Bird (1927) already studied the number of identical mistakes in the exam papers of pairs of test takers (i.e., test takers who were seated next to each other in the exam room). He reasoned that if the number of identical mistakes exceeded 'the chance threshold', then it was very likely that test fraud had been committed. However, the threshold is difficult to determine, especially for pairs of essays. In the following decades, the multiple-choice test made its entrance in the field of testing. Response data from mc-tests enabled researchers to use more advanced EDF methods on detecting collaboration between two or more test takers or so-called 'copying behavior'. Saupe (1960) was one of the first researchers that published an empirical model that could be used to determine the likelihood that two test takers collaborated on their mc-test. The statistical procedures on copying behavior in mc-tests were further elaborated and discussed by Bellezza and Bellezza (1989). More recently, Van der Linden and Lewis (2015) introduced Bayesian statistics, opposing the traditional hypothesis testing approach, for detecting copying behavior on tests. Since 1927 many statistical indices on copying behavior or similarity analysis have been presented in the literature. All these indices revolve around the rationale that two or more test takers have a strong similarity in their answers, both correct and incorrect.

Obviously, the higher the correspondence between answers, the more likely it is that test fraud has been committed. Yet, research is still not consonant about when there is enough correspondence to irrefutably prove that test fraud has been committed. It is questionable if this will ever be possible at all. There are many potential causes to aberrations in response patterns (Reise and Due 1991), and there is always a chance that a test taker produces a very unlikely yet honestly produced response pattern. Furthermore, similarity analysis models may become obsolete very quickly. The introduction of CBT has also increased the use of randomly composed tests (e.g., on the fly testing or computerized adaptive testing), which do not allow test takers to copy answers because each test taker is presented a unique sample of questions from the item bank. Of course, randomized testing contributes to the prevention of test fraud, but it does not make it impossible. Other methods of test fraud (e.g., the use of technological devices, preknowledge, or collusion) can still occur. The fact that similarity analysis lacks the (statistical) power to detect all forms of test fraud and because new testing methods (e.g., CAT) do not allow similarity analyses to be performed has led researchers to investigate other ways of detecting test fraud.

For example, statistical methods to assess the congruence between an item response pattern and a specified item response theory (IRT) model started to proliferate since the 1990s (Reise and Due 1991). These models are also called 'person fit' models. The general idea is that response patterns can be defined a priori using these models. Aberrations to these patterns that are determined a posteriori may point to potential test fraud. A well-known person fit model is the Guttman error model. Guttman (1944) invented an item response model in which there is always a chance of 1 that you can demonstrate an ability when you possess that ability, and a chance 0 that you cannot demonstrate an ability when you do not possess that ability. An aberration to such a pattern is called a 'Guttman error' (Meijer 1994). Simply put, in a testing context, a Guttman error indicates that you have answered an item correctly that is beyond your ability level. Of course, this happens rather often—coincidentally a test taker knows the answer to a difficult question or guesses correctly a couple of times. Nonetheless, a strong aberration in the number of Guttman errors that a test taker produced during the test, potentially indicates that test fraud has been committed. The standard Guttman error is the point of departure for many different EDF indices (Van Noord 2018).

Still other EDF methods rely on the analysis of the number of erasures or revisions that a test taker made during the test (McClintock 2015; Van der Linden and Jeon 2012). Interestingly, these methods can be used on both PBT and CBT. In PBT, highly sensitive scanners can discriminate reliably between erasure marks and marks made to indicate the test takers' answers (McClintock 2015). In CBT, when programmed correctly, the test administration software can log whether a test taker (or someone else) changed their answers to the test. To an increasing extent, test organizations are routinely conducting erasure analyses on their tests. The number of wrong-to-right (WTR) changes seem to be most promising in detecting potential test fraud (Maynes 2013). Qualls (2001) was one of the first researchers that empirically investigated how often erasures occur during a test, and what can be regarded as an aberrant number of erasures. Traditionally, and this also holds for the number of Guttman errors or the number of identical mistakes (similarity analysis), researchers check for test takers that are beyond three standard deviations from the mean in their number of erasures and label these test takers as aberrant. As already mentioned before, more recent and advanced Bayesian models are opposing this view, also in erasure analysis (Van der Linden and Jeon 2012; Van der Linden and Lewis 2015). Several comprehensive test fraud cases in the United States have been revealed through erasure analysis (see for example https://nyti.ms/2k6osyd). Remarkably, it was discovered that teachers had changed the test takers' answers to make themselves and their schools look better.

Response time analysis is also used to detect potential test fraud (Van der Linden 2006). However, these analyses can only be performed in CBT, as the system must log a test takers' response time. There are several ways to investigate the authenticity of response times, but the central idea is that fraudulent test takers show aberrances in their response time. Fast response times may indicate preknowledge of the test items and slow response times could point to the use or misuse of (unauthorized) resources, such as a smartphone or a modified calculator that can help in answering test items (Van der Linden 2006; Van der Linden and Jeon 2012; Van der Linden and Lewis 2015). Basically, except for extreme aberrations, response times are difficult to use in detecting test fraud. This is because there is high variance in the response time of test takers, both within and between test takers. Between test takers, because some test takers are just naturally slower or faster than others, and within test takers because of intrapersonal factors such as concentration, indecisiveness, etc. Therefore, simply using an overall test taker mean and then labeling test takers three SD under or above the mean for several items will not work. Van der Linden (2006) presents a log normal response time model for the analysis of response times. In his log normal model both the test taker's working speed and the characteristics of the test time are accounted for, when determining whether a test taker displays aberrancies. For example, a fast response time is not necessarily conspicuous when it concerns a test taker who is naturally very fast and is answering to a concise and closed-ended 'factual' question.

Above, the historical context of EDF and the most used methods have been discussed. The counter side of these methods is that they are in many cases difficult to apply, especially by practitioners. Often, big data log files need to be analyzed, preferably through automatized algorithms. Easy to use and hands-on software does not exist for EDF. Researchers can resort to *R* (2018) packages, but most test agencies do not employ researchers or people that know how to work with *R*. We have, therefore, been working on developing a hands-on software application that can automatically analyze large data sets for potential test fraud. We have called the software *EDF Monitor*. *EDF Monitor* can, through a transformation database and a communication protocol, be connected to every test administration program. Test takers' response data are uploaded into the software daily through an automatic job. The analysis is run nightly, and aberrant test takers are 'flagged' automatically, so that the practitioner can use the information the following day. *EDF Monitor* currently encompasses nine algorithms, and we are planning to expand. The algorithms are based on similarity analyses, the Guttman error person fit model, and the log normal response time model. In the future, erasure analysis will also be part of *EDF Monitor*. Of course, we were interested in the empirical functioning of *EDF Monitor* and the analyses. We have therefore performed a study on the detection rate (i.e., the number of true positives) and the reliability (i.e., the number of false positives) of the software.

20.2 An Empirical Study on the Reliability of *EDF Monitor*

The main objectives of the study were to establish the highest detection rate using different and manually set combinations of the algorithms in the software (i.e., the true positive rate), and the reliability of the algorithms (i.e., the true negative rate). The focus of the study was on the functioning of the Guttman error based indices, the log normal response time model, and a combination of both. In total a combination of six indices were researched. The first three indices were the standard Guttman error model (G^*), and the Guttman error relative to the distance of the vector of item responses. In the standard Guttman error model, a Guttman error exists in a vector of item responses that are ordered from easy to difficult, when a more difficult item is answered correctly after an easier item has been answered incorrectly. In the adapted Guttman error models, that take relative distances into account, a Guttman error only counts as an error when a specific number of positions in the vector have been skipped or when the distance in p-value (derived from Classical Test Theory) has crossed a pre-specified p-value threshold. The fourth index is the standard Guttman error model combined with the log normal response time model. The fifth and sixth indices are the adapted Guttman error indices combined with the log normal response time model.

20.2.1 Method

The study was conducted using mixed methods, with a research design that covers a confirmatory and exploratory phase. In the first phase the control group and the experimental group were compared, and in the second phase the different experimental conditions were compared. In the experimental conditions, the participants were instructed to commit test fraud in several ways: (1) using a smart device (smartphone), (2) internal collaboration (proctor leaves exam room), (3) proctor assistance, (4) using cheat sheets (notes on paper), and (5) pre-knowledge. Participants in the control condition were regular test takers which were put under enhanced surveillance (i.e., increased number of proctors).

Participants

The control group consisted of 37 participants (age: $M = 39.30$ years, $SD = 13.70$; 14 females). The experimental group consisted of 80 participants, distributed over the different conditions: smartphone ($n = 18$; age: $M = 17.83$ years, $SD = 2.85$; 7 males), internal collaboration ($n = 16$; age: $M = 17.38$ years, $SD = 1.02$; 8 males), proctor assistance ($n = 21$; age: $M = 16.95$ years, $SD = 1.69$; 7 females), cheat sheet ($n = 8$; age: $M = 17.38$ years, $SD = 1.19$; all males), and pre-knowledge ($n = 17$; age: $M = 51.25$ years, $SD = 3.58$; 5 times male). Due to value of the test items, participants in the fifth condition (i.e., pre-knowledge) could not be actual test takers. For test security reasons it would be highly undesirable to have regular test takers gain pre-knowledge of the test items of an exam. Therefore, four co-workers with deep knowledge of the content of the item bank were selected to take five trial exams each, on the condition that they were familiar with the item bank.

Materials

Randomly drawn items from the Basic Competence Legal Knowledge for Extraordinary Detective Officers item bank (linear on the fly) were used (230 items). The unique tests consisted of 50 items and had an average p-value of 0.69, an average

RIT-value of 0.28 and an average reliability of $\alpha = 0.76$. All tests were trial exams and test takers would not be certified when they passed the exam.

Procedure

The experiment was run in several individual and group sessions of approximately 90 min. Groups consisted of participants in the same experimental condition or the control condition. Participants in different conditions were never combined in one exam room. Only the participants in the fifth condition (e.g., pre-knowledge) were run in individual sessions for practical reasons, as the group setting had no additional value. The participants in this condition were also not proctored, all others were by two or more proctors, including a researcher. All participants were informed about the research goals, method, and consequences of participation, after which the participants signed the informed consent form. Participants in the control group were strictly monitored, to ensure nobody in this group cheated. Respondents in the experimental groups were instructed to cheat in a specific way to attain the best grade possible on the trial exam.

In the first condition (i.e., smartphone) the participants were told to consult their smartphone during a period of three minutes. This moment was announced after the first 15 min of the exam had passed. The violators were not allowed to talk out loud or make other noises during this time. In the second condition (i.e., internal collaboration) the group of violators was left alone by their proctors for three minutes, allowing them to consult each other for answers. The violators were notified of the occurrence and instructed to use their time wisely. They were not told how long the proctor would be gone, as this would also be unknown to them, were this an event to occur in reality. In the third condition (i.e., proctor assistance) the violators were instructed to consult the proctor at least once when they struggled with an item. The proctor, a teacher from the same educational institution as the test takers, was instructed to provide more information than usually allowed, or, if this information did not help, the actual answer to the question. Since the participants were to wait their turn to ask a question, they were encouraged to mark the respecting item and continue with the exam while the proctor took turns. In the fourth condition (i.e., cheat sheets) the violators were asked to bring a cheat sheet to the exam. The proctor was instructed to ignore the sheets. In the fifth condition (i.e., pre-knowledge) the violators completed the exam without interference of a proctor or researcher.

Data analysis

First, it is tested whether the mean score for the six indices in the cheating conditions is significantly higher than in the non-cheating condition. This is evaluated using independent *t*-tests. Secondly, the optimal true positive detection ratio and true negative ratio are established. In addition, and for an explorative purpose, for each condition the detection ratio was determined through a Kruskal-Wallis ANOVA, combined with Mann-Whitney U and Wilcoxon Signed Ranking tests.

Table 20.1 Descriptive and results from the initial data forensics analyses

Index	Not cheating			Cheating					
	(N = 37)			(N = 80)					
	M	SD		M	SD		t(115)		p
G^*	0.08	1.16		0.88	1.08		−3.65		<001
			Mean rank			Mean rank	U	z	
G^*_{d20}	0.52	1.41	49.07	0.99	1.51	63.59	1112.50	−2.16	0.031
$G^*_{d0.20}$	−0.40	0.91	48.96	0.02	1.10	63.64	1108.50	−2.19	0.029
G^*_{rt}	0.01	1.04	39.34	1.60	2.15	68.09	752.50	−4.28	<0.001
G^*_{rtd20}	0.41	1.10	46.64	1.87	2.75	64.72	1022.50	−2.71	0.007
$G^*_{rtd0.20}$	−0.34	0.75	46.22	0.75	2.29	64.91	1007.00	−2.84	0.005

Note In this table p is reported two-tailed, and z is corrected for ties

20.2.2 Results

First, regarding the comparison between the non-cheating and cheating conditions, results showed that the scores on all six indices were higher in all five cheating conditions as compared to the control condition (see Table 20.1).

Secondly, regarding the optimal true positive and true negative ratio, we found that a *combination* of combined indices leads to highest ratios. That is, when the $G{*}_{rt}$ index (the standard Guttman error model combined with the log normal response time model), was combined with the G^*_{rtd12} (the adapted Guttman error model with 12 skipped positions in the item response vector that was ordered from the easiest item to the most difficult item), the $G^*_{rtd0.21}$ (the adapted Guttman error model for a p-value threshold of 0.21 combined with the log normal response time model), and the G^*_{rtd20}, it leads to the optimal true positive and true negative ratio for this data set. The true positive ratio was in this case 37.5%. That is, of the 80 students in the cheating conditions, 30 were indeed detected as cheaters. 50 of 80 students who were in one of the cheating conditions were not detected, which is translated in a false negative ratio of 62.5%. The true negative ratio was 97.3%—or a false positive ratio of 2.7%. That is, of the 37 students in the non-cheating condition only one was wrongly detected as a cheater (a type I error).

Using the combination of combined indices discussed above, our explorative analyses showed that especially cheating through proctor assistance and pre-knowledge could best be detected (see Table 20.2). Yet, four Kruskal-Wallis ANOVAs indicated that there were no significant differences between the cheating conditions for analysis with the G^*_{rt} index, H (corrected for ties) $= 8.030$, $df = 4$, $N = 80$, $p = 0.090$, Cohen's $f = 0.336$, the G^*_{rtd12} index, H (corrected for ties) $= 8.453$, $df = 4$, $N = 80$, $p = 0.076$, Cohen's $f = 0.346$, the $G^*_{rtd0.21}$ index, H (corrected for ties) $= 3.571$, $df = 4$, $N = 80$, $p = 0.467$, Cohen's $f = 0.218$, and the G^*_{rtd20} index, H (corrected for ties) $= 8.006$, $df = 4$, $N = 80$, $p = 0.091$, Cohen's $f = 0.336$.

Table 20.2 Mann Whitney U statistics determining the validity of the indices for every experimental group

Experimental index	Group	Control condition	U	z	p	r
	Mean rank	Mean rank				
Smart phone						
G^*_{rt}	36.19	24.01	185.50	−2.66	0.008	0.36
G^*_{rtd12}	36.25	23.99	184.50	−2.69	0.007	0.36
$G^*_{rtd0.21}$	30.75	26.66	283.50	−0.93	0.353	0.13
G^*_{rtd20}	33.25	25.45	238.50	−1.72	0.085	0.23
Internal collaboration						
G^*_{rt}	31.31	25.14	227.00	−1.35	0.178	0.19
G^*_{rtd12}	30.75	25.38	236.00	−1.18	0.239	0.16
$G^*_{rtd0.21}$	27.59	26.74	286.50	−0.19	0.849	0.03
G^*_{rtd20}	26.56	27.19	289.00	−0.14	0.890	0.02
Proctor assistance						
G^*_{rt}	41.98	22.42	126.50	−4.25	<0.001	0.56
G^*_{rtd12}	42.60	22.07	113.50	−4.49	<0.001	0.59
$G^*_{rtd0.21}$	35.88	25.88	254.50	2.24	0.025	0.29
G^*_{rtd20}	39.26	23.96	183.50	−3.36	0.001	0.44
Cheatsheet						
G^*_{rt}	31.75	21.11	78.00	−2.09	0.037	0.31
G^*_{rtd12}	32.63	20.92	71.00	−2.31	0.021	0.34
$G^*_{rtd0.21}$	28.44	21.82	104.50	−1.34	0.201	0.20
G^*_{rtd20}	27.25	22.08	114.00	−1 03	0.327	0.15
Pre-knowledge						
G^*_{rt}	38.03	22.66	135.50	−3.35	0.001	0.46
G^*_{rtd12}	33.68	24.66	209.50	−1.97	0.049	0.27
$G^*_{rtd0.21}$	34.68	24.20	192.50	−2.36	0.018	0.32
G^*_{rtd20}	35.21	23.96	183.50	−2.47	0.013	0.34

Note In this table z is corrected for ties, except for the Cheatsheet experimental group, and p is two-tailed

20.2.3 Discussion and Conclusion

The aim of this study was to assess the detection strength of several data forensics indices. The response data of a control group of highly supervised examinees and an experimental group of instructed cheaters were analyzed for that purpose. We found that a combination of the Guttman error model, two adaptations to the Guttman error model, and the log normal response time model yielded the highest detection rate (37.5%) combined with the lowest type I error rate (2.7%). It could therefore be said that the strength of these indices lies in their combination. Or, response time analysis is a better behavioral indicator when combined with the Guttman error model than when the Guttman error model is used separately. Although there were no significant differences in the indices between the five methods of cheating, a trend seems to lean towards proctor assistance and pre-knowledge to be best detectable with this combination of indices. The goal of a larger and improved follow-up study is to investigate whether this is indeed the case.

20.3 The Design and Development of the *EDF Protocol*

This research was focused on developing standards covering the entire process of examination to limit the chances of security risks (e.g., optimizing the prevention of exam fraud, and its detection by means of data forensics). Accordingly, the corresponding research question was:

1. Which standards regarding preventing and detecting exam fraud in the process of examination need to be included into the *EDF Protocol*?

In addition, practitioners should be able to act on indications of exam fraud based on these standards, this study therefore also answered a second research question:

2. Which conditions must be considered during development of the *EDF Protocol* to support practitioners in detecting possible gaps in the security of their examination process?

20.3.1 Method

The *EDF Protocol* was constructed and validated in five consecutive steps: (1) a literature search relating relevant standards and criteria on security of the examination process, and also prevention and detection of exam misconduct; (2) development of the *EDF Protocol* prototype; (3) validation of the prototype standards and criteria through semi-structured interviews with content experts; (4) adjustment of the prototype towards a final version of the *EDF Protocol*; and (5) empirical testing of the protocol by putting the protocol to practice.

Step 1—Literature search

For the first step the PRISMA framework described by Moher et al. (2009) was used for conducting the literature review. To compile an evidence base for the development of the *EDF Protocol*, three major databases were searched: Scopus, Web of Science, and Google Scholar.

For the main topic of the study several search terms were used (see Table 20.3). Boolean search operators were also used during this step (e.g., AND, OR, NOT,

Table 20.3 Search terms used in the literature search

Keywords	Related/more specific/broader
Test security	Educat*, prevention, detection, standards, fraud, cheating
Data forensics	Educat*, fraud, cheating

and *). The initial search findings were reduced by excluding duplicates. Hereafter, the articles were first screened on title, and secondly the abstract. Articles were included in the study if the main topic of the paper or chapter related to security of examination, or if the paper or chapter provided a structured set of guidelines or standards on security of examination. This method not only summarized existing literature, but also aimed to generalize and transfer findings for policy making and practice (Cassell et al. 2006). Prior to the development of the EDF prototype, an overview was made of the most important findings from the literature review. These insights were used to develop an EDF prototype.

Step 2—Developing an EDF-Protocol prototype

The insights gathered in the literature search were used in the development of the first set of standards of the prototype (Part A), as well as a corresponding grading system. The intention was to make the standards (concerning prevention of misconduct during the process of examination) as complete as possible before starting the interviews.

The development of part B (i.e. the standards and criteria for detection of misconduct by means of using data forensics) took more time and effort. Although there is a considerable amount of scientific literature on the possibilities of using data forensics, research is mostly focused on case- or comparative studies, and thus often lacking proper directions for practical implementation. The intention with this part of the prototype was therefore to enter the interviews more open minded, hence gain insight on what the content experts deem to be included or excluded in terms of data forensic standards.

During this step a deliberate choice was made for a distinction between a set of standards for prevention and a set of standards for detection (by means of data forensics) because these goals do not always coincide in practice.

Step 3—Validating the EDF-Protocol standards
Participants

The prototype was validated by means of seven semi-structured interviews. All approached experts have practical and theoretical experience on the subject. These interviews were held with content experts from different backgrounds, amongst them psychometricians, policy makers and practitioners in the field of test security or education. To keep development of the prototype and validation of the content separate steps, the participating experts were not involved during the development of the prototype.

Procedure and materials

Systematic expert interviews offer the possibility to identify strengths and weaknesses in the content (McKenney and Reeves 2012; Piercy 2015). This method is a valuable source of data collection, particularly when establishing the correctness (e.g., validating) of the content of a product (Wools et al. 2011). The interview format consists of four categories; category one focused on general questions concerning the protocol ($n = 7$), category two focused on questions concerning the protocol content ($n = 4$), category three related to the grading of the protocol ($n = 5$), and category four focused on the data forensic standards ($n = 5$). An example of an interview question would be: "The goal of the protocol is to provide a good check whether the process of examination is secure. Do you think this is feasible in the current form?"

At the start of the interview, each respondent was asked for consent verbally. This means that they were asked whether the interview could be recorded and whether the input from the interview could be used to validate the content of the prototype. It was also agreed in advance with the participants that they would receive the transcript of the interview, to be completely transparent about the input that was collected.

After the interviews, all the recordings have been converted to verbatim transcripts to keep statements in their proper context. Cues and codes were written in the margin of the transcript to indicate a reference to a specific question or part of the prototype. Subsequently, text fragments were summarized based on the interview categories ($n = 4$). The selection of usable statements was done on an individual basis by the author.

Step 4—Adjustment of the EDF prototype and final protocol

In the fourth step, the statements from the experts were used to transform the prototype into a final version of the *EDF Protocol*.

Step 5—Implementation of the EDF Protocol

In the fifth step of this design research the final *EDF Protocol* was used to determine if there was a possible security risk within an exam program.

20.3.2 Results

Step 1—Literature search

The literature search was split into two main topics. Firstly, the search for literature on 'Test Security', and secondly the search for 'Data Forensics' related literature. The literature search was carried out in June 2018. As was described in the method section, the PRISMA framework was used in this step (Moher et al. 2009).

The first major topic was 'Test Security'. The key search term was based on the research question, namely test security. To broaden or specify the search, the following search terms were also used: prevention, detection, standards, fraud and cheating. Not all search terms provided usable information. Figures 20.1 and 20.2 show the steps of the search processes for both queries.

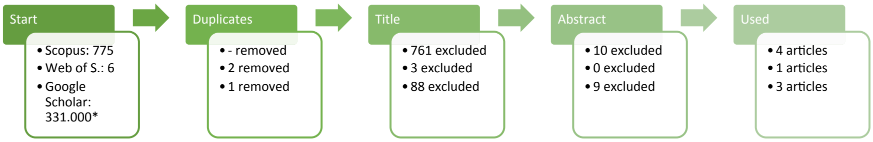


Fig. 20.1 PRISMA flow chart of the search process in the query of test security

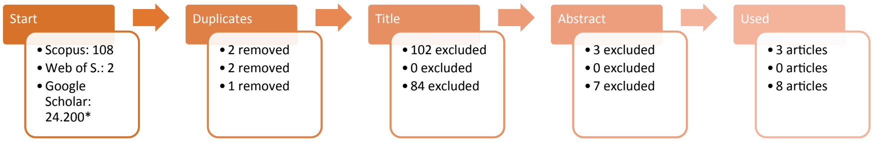


Fig. 20.2 PRISMA flow chart of the search process in the query of data forensics

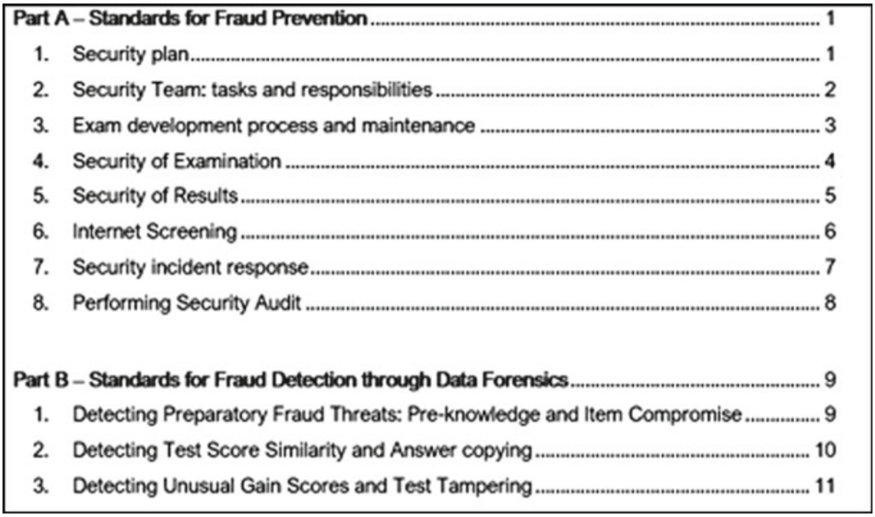


Fig. 20.3 Body of content of the EDF protocol prototype

The second major topic of this step was focused on gathering literature on data forensics. For this topic, the main keyword, data forensics, directly relates to the main research question. Again, to broaden or specify the search at certain points, the following search terms were also used: educat* standards, fraud and cheating. Figure 20.3 shows the steps of the search process.

Because the literature search did not yield the desired results, a snowballing approach, presented by Wohlin (2014), was used to find more relevant literature on the topic of test security. Because of this method, scanning the reference lists of the articles and handbooks that were found during the initial literature search provided new information on studies in the field of data forensics and test security (n = 20).

Step 2—Developing an EDF Protocol prototype

Based on the literature review and reading through similar protocols, manuals and handbooks, two main areas for development were identified. First, an area concerning standards and criteria with a focus on preventing misconduct during the process of examination (Part A). Second, an area with a set of standards concerning the detection of misconduct after examination by means of data forensics (Part B). The EDF prototype's body of content is presented in Fig. 20.3. The prototype standards each relate to the most commonly used and accepted guidelines on test security: The Security Guidelines by NVE (Dutch Association for Assessment) and Caveon (2016), the guidelines by SURF (ICT cooperation for education and research in the Netherlands, 2014), and the Standards for Educational and Psychological Testing (AERA et al. 2014) as well as other literature to support the inclusion of these criteria in the prototype.

Step 3—Validating the EDF-Protocol standards

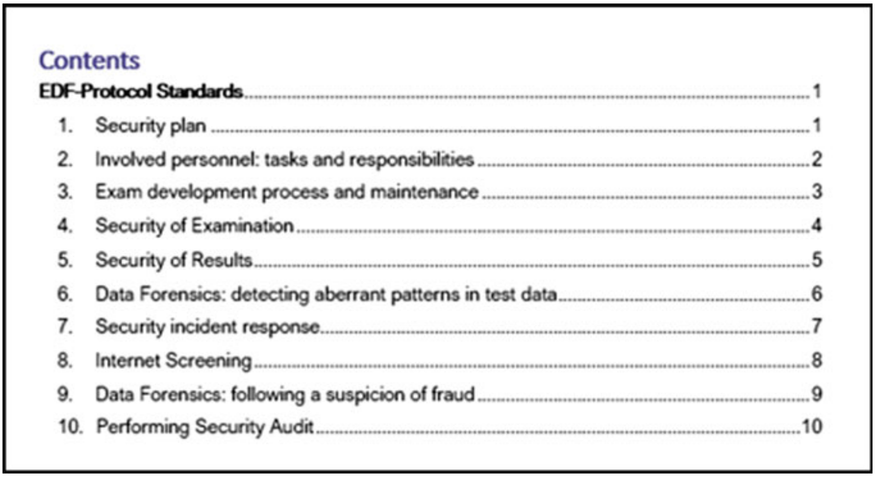
The content of the prototype was validated by means of seven semi-structured expert interviews. The interview is divided into four categories; category one focused on general questions concerning the protocol ($n = 7$), category two focused on questions concerning the protocol content ($n = 4$), category three related to the grading of the protocol ($n = 5$), and category four focused on the data forensic standards ($n = 6$). The third step of the research yielded several valuable statements made by the content experts, which have been incorporated into the final version of the *EDF Protocol*. The way these statements are embedded in the protocol and the arguments for inclusion are described in step four.

Step 4—Adjustment of the EDF prototype and final protocol

The interview statements were summarized into three categories. The first category describes adjustments based on statements referring to the protocol in general (e.g., "include possible evidence in the protocol"). The second category include adjustments referring to the content (e.g., "include awareness in the protocol"). The third category include grading adjustments (e.g., "add the option 'not applicable"). The EDF-protocols' body of content is shown in Fig. 20.4.

General protocol adjustments

The first adjustment removes the distinction between part A and part B. After statements from several content experts, the three data forensics standards have been revised into two standards, and hereafter included within part A. Thus, the result is a set of ten standards concerning security of the examination process. The first data forensics standard (standard 6), describes several criteria around detecting aberrant patterns in test data. The second data forensics standard (standard 9) include criteria aimed for handling a suspicion of fraud or misconduct. Subsequently, these two data forensics standards now have the same grading system as the other standards. These adjustments have been made to make the *EDF Protocol* more streamlined in general and the content more consistent.



Contents

EDF-Protocol Standards 1

1. Security plan 1
2. Involved personnel: tasks and responsibilities 2
3. Exam development process and maintenance 3
4. Security of Examination 4
5. Security of Results 5
6. Data Forensics: detecting aberrant patterns in test data 6
7. Security incident response 7
8. Internet Screening 8
9. Data Forensics: following a suspicion of fraud 9
10. Performing Security Audit 10

Fig. 20.4 Body of content EDF protocol

The second adjustment was the introduction of an evidence table for each standard. This adjustment was based on two potential advantages. First, this table offers the opportunity to gather concrete insights per standard on how each criterion is currently dealt with. Secondly, the provided evidence gives the opportunity to initiate a discussion. For example, to check for potential security risks, or to determine if a change in practice is needed. The third general adjustment was a change in the order of the standards. They have been adjusted to make the standards more logically reflect the process of examination in a chronological way.

Content adjustments

Standard two has been revised based on several expert statements. Firstly, the name ‘Security team’ raised questions, and was considered too big or too vague. The image created with this standard was that a separate team should be responsible for securing the exam process. However, this was not intended with this standard. However, the aim for this standard was to support awareness and to offer guidance in assessing the responsibility and integrity of all involved personnel within the process of examination. Accordingly, the name of standard two was revised into ‘Involved personnel: tasks and responsibilities’. Also, the description of the four criteria have been revised to support security awareness.

Another clearly voiced point of feedback in some interviews was the lack of a standard concerning the assessor of exams or tests. The significance of including assessor integrity in the protocol was made very clear; however, instead of devoting an entire standard to the assessor, several criteria have been revised, and new criteria were developed to meet the statements made in this area (e.g., standard 2: criteria 2, 3 and 4, standard 4: criteria 5, and standard 5: criteria 4). This choice is based on the fact that the integrity of all personnel involved was already included in the revised second standard.

Finally, several adjustments have been made in terms of naming the criteria. Reasons for these adjustments were not always found in the interview transcripts but were for example based on the fact that the original naming of some criteria did not fully represent a criterion. In one case, however, two content experts rightly pointed to the fact that criteria one (Proctoring) and four (Use of materials) of standard four, of the prototype, aimed to measure the same. Namely, the use of unauthorized materials, therefore the latter (use of materials) was excluded.

Grading adjustments

In all interviews, on various topics, several experts stated that drawing conclusions by means of the rubrics could be risky, especially considering the impact these conclusions might have. In the prototype, the impact of the assessment was not clearly reflected in the criteria when considering assessing a diversity of exam programs. Therefore, several adjustments have been made to make the protocol even more manageable in terms of grading. First the rubrics have been revised. In the prototype, all levels of grading (e.g., insufficient, sufficient and good) had a description. In order to make the protocol more manageable, only a clear description of the 'sufficient' level was now included in the rubric. The descriptions of the other levels have become fixed, namely: (1) Insufficient: the described criteria are not met; (2) Good: the criteria are amply met/demonstrates how this is acted upon. Because they now have a fixed character they are excluded from the rubrics and included as a note under each standard.

Secondly, a new grading option was introduced, the option 'Not applicable' has been included. This adjustment is based on comments from experts whom stated, 'I understand that you've included this criterion, but for me this would not apply'. In the prototype, there was no way of indicating applicability of certain criterion. Thirdly, a minor change was made in terms of usability. In the prototype the awarding of a score was open. This could be done, for example, by filling in an 'X' by hand. In the final version blocks have been added, when clicking a block an 'X' will automatically be applied. This makes the protocol slightly more user-friendly and more intuitive.

Step 5—Implementation of the EDF Protocol

During the fifth step, the *EDF Protocol* was used to evaluate and measure possible security risks within an exam program. In the scope of the current study, this step has been taken to determine the actual practical value of the protocol. A consultation with the manager of the exam program was organized to implement the *EDF Protocol.* The application of the protocol in the exam program was the final validation strategy for the content of the protocol. In doing so, the application of the protocol has demonstrated that it is functioning as intended, and therefore this step confirmed its added value for practice. The effectiveness of the protocol can best be described by presenting the results, hence the validation process will be discussed together with the findings and recommendations.

To summarize, 6 out of 10 standards were assessed with a '*medium/high security risk*'. Although this is not an ideal score for the exam program, it does show that the protocol can flag security gaps in the examination process and due to the open nature of the criteria it was also possible to provide several concrete recommendations to limit the chances of security risks in the future. In addition, the remaining 4 out of 10 standards were assessed with a '*low security risk*'. This indicated that the standards were developed in such a way that proper security measures also get rewarded by the protocol. Although exam fraud can never be fully banned, these findings advocate the current content of the protocol, since it seemingly provides standards covering the entire process of examination.

20.3.3 Discussion and Conclusion

This design research started on the premise of developing a set of standards, enabling practitioners to prevent and detect possible fraud during the process of examination. In the end, the research provided a set of standards aimed at achieving a well-secured exam process as well as increasing awareness in doing so.

By means of the five design steps carried out in this study, the main research question is unambiguously answered by stating that the *EDF Protocol* provides enough direction and guidance in securing the entire process of examination. To summarize these standards: (1) Security plan, (2) Tasks and responsibilities, (3) Exam development and maintenance, (4) Security of examination, (5) Security of results, (6) Data forensics I, (7) Incident response, (8) Internet screening, (9) Data forensics II, (10) Security audit. Continuous application of the protocol in the future must determine whether the current set of standards and underlying criteria is sufficient. To illustrate, within this study the protocol was used for an exam program that did not have a security plan. Although this was well illustrated by applying the protocol, which emphasizes the usability of the protocol, we do not yet know how the protocol responds to a well-secured exam program in terms of evaluating and measuring the possible security risks.

To answer the second research question, during development, several conditions have been considered to provided practitioners with the ability to act on indications of exam fraud based on these standards. By adding an 'evidence-table' for each standard, organizations are given the opportunity to provide concrete insights per standard on how each criterion is currently dealt with, meaning they can now include their own practice in the protocol. Secondly, it provides the foundation for an internal discussion. By doing so, security awareness is being encouraged on a personal level, and at a policy level, again, the foundation is laid for a well secure exam program. Also, the implementation of the protocol results in a 'protocol report', including findings for each standard as well tailor-made recommendation (e.g., short term or long term). A deliberate choice was made not to include a set of fixed recommendations into the protocol, on the contrary, these recommendations are now the result of implementation. In doing so the protocol can be used more widely in various exam programs, without compromising or limiting the quality of implementing the *EDF Protocol* for individual exam programs.

20.4 General Discussion

Two empirical studies on educational data forensics have been presented in this chapter. The first study was an empirical study in which the effectiveness of a combination of EDF indices was tested. The main finding was that the combination of Guttman error model indices and log normal response time model indices were able to differentiate between non-cheating and cheating students. Using the right combination and the optimal cut-off scores for the indices, we were able to detect 37.5% of the cheating students, at a false positive ratio of 2.7% (the type I error). Finally, the findings indicated that especially preknowledge and the assistance of the proctor during the test were best detectable.

The second study had a more qualitative character. The goal of the study was to establish an educational data forensics protocol, consisting of evidence-based standards, that could be used by both researchers and practitioners. Through five design and development steps we were able to develop an *EDF Protocol* that consisted of ten standards, and a grading system that can be used to evaluate potential test security risks and breaches.

To conclude, we see EDF more as a continuous improvement process for test organizations, rather than a set of techniques or methodologies that are deployed ad hoc. Test fraud continues to advance, and improvements in test security should keep pace. In that sense, the *EDF Monitor* and the *EDF Protocol* can work in tandem. Potential breaches detected through the protocol can point to where to look in the data for test fraud, using the monitor. Conversely, the origin of detected test fraud with the monitor can be found by applying the protocol. Although test fraud will probably never be fully eradicated, the use of monitoring and security protocols allows test organizations to protect against misconduct.

References

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69,* 44–49.

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16*(3), 151–155.

Berkhout, E. E., Van der Werff, S. G., & Smid, T. H. (2011). Studie & Werk 2011. Amsterdam: SEO Economisch Onderzoek. Retrieved from http://www.seo.nl/fileadmin/site/rapporten/2011/2011-29_Studie_en_Werk_2011.pdf.

Bird, C. (1927). The detection of cheating in objective examinations. *School & Society, 25,* 261–262.

Cassell, C., Denyer, D., & Tranfield, D. (2006). Using qualitative research synthesis to build an actionable knowledge base. *Management Decision, 44*(2), 213–227.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139–150.

Hughes, J. M. C., & McCabe, D. L. (2006). Academic misconduct within higher education in Canada. *The Canadian Journal of Higher Education, 36,* 1–12.

Lin, C. H. S., & Wen, L. Y. M. (2007). Academic dishonesty in higher education: A nationwide study in Taiwan. *Higher Education, 54,* 85–97.

Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 187–214). New York, NY: Routledge.

McClintock, J. C. (2015). Erasure analyses: Reducing the number of false positives. *Applied Measurement in Education, 28*(1), 14–32.

McKenney, S., & Reeves, T. C. (2012). *Conducting educational design research*. New York, NY: Routledge Education.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*(4), 311–314.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269.

Piercy, K. W. (2015). Analysis of semi-structured interview data. *Department of Family, Consumer, & Human Development,* 1–16.

Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practices, 20*(1), 9–16.

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*(3), 217–226.

Saupe, J. L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement, 20*(3), 475–489.

Van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204.

Van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics, 37*(1), 180–199.

Van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika, 80*(3), 689–706.

Van Noord, S. (2018). *Using data forensics to detect cheating in randomized computer-based multiple-choice testing* (Unpublished master thesis).

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. London, UK: ACM.

Wollack, J. A., & Fremer, J. J. (2013). Introduction: The test security threat. In A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 1–13). New York, NY: Routledge.

Wools, S., Sanders, P. F., Eggen, T. J. H. M., Baartman, L. K. J., & Roelofs, E. C. (2011). Evaluatie van een beoordelingssysteem voor de kwaliteit van competentie-assessments [Testing an evaluation system for performance tests]. *Pedagogische Studiën, 88,* 23–40.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.