

Methodology of Educational Measurement and Assessment

L. Andries van der Ark
Wilco H. M. Emons
Rob R. Meijer *Editors*

Essays on Contemporary Psychometrics

 Springer

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, Boston College, Boston, MA, USA

Editorial Board Members

Claus H. Carstensen, University of Bamberg, Bamberg, Germany

Hua-Hua Chang, Purdue University, West Lafayette, IN, USA

Hong Jiao, University of Maryland, College Park, MD, USA

David Kaplan, University of Wisconsin-Madison, Madison, USA

Jonathan Templin, The University of Iowa, Iowa city, IA, USA

Andries van der Ark, Res Inst of Child Devt & Education, University of Amsterdam,
Amsterdam, Noord-Holland, The Netherlands

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.

L. Andries van der Ark • Wilco H. M. Emons •
Rob R. Meijer
Editors

Essays on Contemporary Psychometrics

 Springer

Editors

L. Andries van der Ark
Research Institute of Child Development
and Education
University of Amsterdam
Amsterdam, The Netherlands

Wilco H. M. Emons
Department of Methodology and Statistics
Tilburg University
Tilburg, The Netherlands

Rob R. Meijer
The expertise group Psychometrics and
Statistics
University of Groningen
Groningen, The Netherlands

ISSN 2367-170X ISSN 2367-1718 (electronic)
Methodology of Educational Measurement and Assessment
ISBN 978-3-031-10369-8 ISBN 978-3-031-10370-4 (eBook)
<https://doi.org/10.1007/978-3-031-10370-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

Chapters [3], [5], [8], [16] and [19] are licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). For further details see licence information in the chapters.

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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

According to the Psychometric Society (n.d.), psychometrics is the science devoted to the advancement of quantitative measurement practices in psychology, education, and the social sciences. The Psychometric Society was established in 1935, but psychometrics is much older. The word *psychometrics* as defined by the Psychometric Society was used in Galton's (1879) essay "Psychometric Experiments" (Jones & Thissen, 2006; see also Heiser, this volume), yet quantitative measurement in education dates back at least to the Chinese state exams that started during the mid-Tang dynasty (618–907). One of the early landmarks in psychometrics was the publication of the book *Statistical Theories of Mental Test Scores* by Lord and Novick (1968). They provided a comprehensive framework of both classical and modern test theory. Their book led to rapid developments in psychometrics, a trend that was further accelerated by an increasing societal demand for accountable and responsible measurement in education. Until the end of the millennium, psychometrics was directly or indirectly based on Lord and Novick. With the start of the new millennium, psychometrics became a broader field. Faster computers allowed for something we may coin *new psychometrics*: measurement models that tap on specific attributes allowing fine-grained cognitive (or clinical) diagnosis, data-driven models for both measurement and prediction, the use of process data (e.g., response times) for formative assessment, and advanced computer-intensive estimation methods.

This book is dedicated to Klaas Sijtsma, a Dutch psychometrician and former president of the Psychometric Society who had a large contribution to both traditional psychometrics and new psychometrics. He has firm roots in traditional psychometrics. His initial research interest in the early 1980s was nonparametric item response theory, a topic he researched throughout his career. At the basis for these initial research interests was his observation that nonparametric methods received much less attention than parametric methods such as Rasch models. In his view, this lack of interest was unfounded. Being based on general assumptions, nonparametric methods provide a strong general theoretical framework for research into the properties of measurement model including parametric models. At the same time, these nonparametric methods offer practical tools for constructing sound

measurement instruments. This interplay between deepening theoretical understanding of measurement models, on the one hand, and translating the knowledge to very practical applications, on the other hand, very much characterizes Sijtsma's scientific contributions to the field of psychometrics.

Although he has published mostly on traditional psychometrics, especially on classical test theory and parametric and nonparametric item response theory, he also published many papers on new psychometrics. Examples include an early seminal paper on cognitive diagnosis models (Junker & Sijtsma, 2001); reflections on psychometrics, including the often-cited paper on Cronbach's alpha (Sijtsma, 2009); and essays on measurement foundations (Sijtsma, 1998; Sijtsma & Emons, 2013). Like his advisor Ivo Molenaar, Sijtsma stressed that psychometrics is an auxiliary science to the social and behavioral sciences. Because psychometrics is complicated, psychometricians should educate and assist social and behavioral scientists. Sijtsma wrote several scholarly textbooks, he published with social and behavioral scientists on substantive topics, and he made difficult topics accessible to social and behavioral scientists through publications in professional journals.

One textbook that had arguably a high impact in the Netherlands is Drenth and Sijtsma's (1990/2005) introductory book to test theory. This book is still used in bachelor psychology programs at many Dutch universities. Equally worth mentioning are the highly accessible and often-cited book on nonparametric item response theory (Sijtsma & Molenaar, 2003) and his latest book with Van der Ark (Sijtsma & Van der Ark, 2021), which provides a comprehensive overview of psychometrics, including both traditional and new psychometrics. Sijtsma was also one of the driving forces to come up with a thorough revision of the review system for evaluating test quality of the Dutch Committee on tests and testing (Evers et al., 2010). This system is now widely accepted as the authoritative standards for tests and testing in the Netherlands for more than a decade.

This volume, entitled *Essays on contemporary psychometrics*, provides an overview of the state of the art in both traditional psychometrics and new psychometrics. The chapters have been written by psychometricians who have collaborated closely with Klaas Sijtsma, and hence the book not only discusses traditional psychometrics but also topics from new psychometrics. In most chapters, the influence of Sijtsma's view on science is visible by the practical stance of many chapters, without ignoring the theoretical underpinnings, and by the large number of citations to Sijtsma's psychometric work. The book is divided into four parts, each with a number of chapters on a specific theme. The parts are not mutually exclusive nor exhaustive, and each chapter can be read separately. The parts are summarized as follows.

Part I. General Perspectives on Psychometrics. In the first part, Heiser discusses the roots of psychometrics before Francis Galton, and De Boeck and Gore reflect on two sides of psychometric models: psychometric models as psychological models and as measurement models. Meijer, Niessen, and Neumann discuss the underrepresentation of existing knowledge on how test scores can be used in decision-making. Veldkamp discusses how trustworthy artificial intelligence can

be integrated into the domain of psychometrics. Borsboom presents item response theory as psychometric networks.

Part II. Factor Analysis and Classical Test Theory. Hessen presents new expressions for the communality of the total score, the communality of an arbitrary item score, and the proportion of total variance explained under the one-factor model. Closed form distribution-free estimates are presented as well. Molenaar discusses two factor analysis approaches to reliability of change scores. Van Ginkel discusses methods for handling missing data in principal component analysis (PCA). Von Davier and Clauser show that using non-linear functions for equating and score transformations leads to consequences that are not commensurable with classical test theory (CTT). Ackerman, Ma, and Luecht examine how test characteristics such as test length, dimensionality, and item discrimination affect coefficient (Cronbach's) alpha and omega. Finally, Emons discusses theoretical and computational aspects of conditional standard errors of measurement.

Part III. Item Response Theory. Marsman, Bechger, and Maris discuss how to improve the efficiency of two recently published algorithms for sampling simultaneously from many conditional distributions. Hemker reflects on the use of weighted or unweighted total scores for modeling and transparent score reporting. Ligtoet investigates the inequality restrictions imposed by Mokken's model of monotone homogeneity (MH) for binary item response variables. In particular, a Bayesian test for the observable property of variables being associated is proposed for the trivariate distributions of all triplets of items. Straat, Kuijpers, Lek, and Emons explore targeted testing from a teacher perspective.

Part IV. New Psychometrics. Tijmstra and Bolsinova discuss advantages, limitations, and risks of the hierarchical model for response times from a practical measurement perspective. Ellis develops two methods for CAT based on the monotone homogeneity model. Conijn, Van Ewijk, Chen, and Van der Ark study whether validity indices can be used to detect and explain discrepancies between scores provided by multiple informants (e.g., self-reports and teacher reports) within the context of ADHD assessment. Van der Ark and Smits propose FlexCAT as a general and flexible computerized adaptive testing approach that is useful when the number of potential items to fill the item bank with is limited, resulting measurements are typically multidimensional, and both measurement and prediction are pursued. De La Torre and Santos discuss the relationship between unidimensional item response theory and higher-order cognitive diagnostic models. Finally, He, Culpepper, and Douglas propose an extension of the sparse latent class model (SLCM) for ordinal measurements, with the purpose of fully exploring the relationships between attributes and response patterns.

We would like to thank Tasos Psychogiopoulos for assisting the editors and checking all references and the reviewers for helping us to ensure the quality of this book.

Amsterdam, The Netherlands
Tilburg, The Netherlands
Groningen, The Netherlands

L. Andries van der Ark
Wilco H. M. Emons
Rob R. Meijer

References

Drenth, P. J. D., & Sijtsma, K. (2005). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen* [Test theory: Introduction to the theory of the psychological test and its applications]. Bohn Stafleu Van Loghum. (Original work published 1990)

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN review system for evaluating test quality*. Nederlands Instituut voor Psychologen. <https://www.psynip.nl/wp-content/uploads/2019/05/NIP-Brochure-Cotan-2018-correctie-1.pdf>

Galton, F. (1879). Psychometric experiments. *Brain*, 2(2), 149–162. <https://doi.org/10.1093/brain/2.2.149>

Jones, L. V., & Thissen, D. (2006). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds). *Handbook of Statistics: Vol. 26. Psychometrics* (pp. 1–27). Elsevier. [https://doi.org/10.1016/S0169-7161\(06\)26001-2](https://doi.org/10.1016/S0169-7161(06)26001-2)

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272. <https://doi.org/10.1177/01466210122032064>

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

Psychometric Society. (n.d.). *What is psychometrics*. <https://www.psychometric-society.org/what-psychometrics>

Sijtsma, K. (1998). *De data maken het model: Over het beperkte belang van meetniveaus en latente schalen* [The data make the model: About the limited importance of measurement levels and latent scales]. [Inaugural address]. Tiburg University, The Netherlands.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>

Sijtsma, K., & Emons, W. H. M. (2013). Separating models, ideas, and data to avoid a paradox: Rejoinder to Humphry. *Theory & Psychology*, 23(6), 786–796. <https://doi.org/10.1177/0959354313503724>

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage. <https://doi.org/10.4135/9781412984676>

Sijtsma, K., & Van der Ark, L. A. (2021). *Measurement models for psychological attributes*. CRC/Chapman & Hall. <https://doi.org/10.1201/9780429112447>

Contents

Part I General Perspectives on Psychometrics

1	Early Roots of Psychometrics Before Francis Galton	3
	Willem J. Heiser	
2	The Janus Face of Psychometrics	31
	Paul De Boeck and L. Robert Gore	
3	Psychological and Educational Testing and Decision-Making: The Lack of Knowledge Dissemination in Textbooks and Test Guidelines	47
	Rob R. Meijer, A. Susan M. Niessen, and Marvin Neumann	
4	Trustworthy Artificial Intelligence in Psychometrics	69
	Bernard P. Veldkamp	
5	Psychological Constructs as Organizing Principles	89
	Denny Borsboom	

Part II Factor Analysis and Classical Test Theory

6	A New Expression and Interpretation of Coefficient Omega Under the Congeneric One-Factor Model	111
	David J. Hessen	
7	A Factor Analysis Approach to Item Level Change Score Reliability	119
	Dylan Molenaar	
8	Handling Missing Data in Principal Component Analysis Using Multiple Imputation	141
	Joost R. van Ginkel	

9	Quantifying the Bias of Non-linear Equating and Score Transformations	163
	Matthias von Davier and Brian Clauser	
10	Examination of Test Characteristics' Effect on Coefficient α and Coefficient ω	181
	Terry Ackerman, Ye Ma, and Richard Luecht	
11	Methods for Estimating Conditional Standard Errors of Measurement and Some Critical Reflections	195
	Wilco H. M. Emons	
 Part III Item Response Theory		
12	Composition Algorithms for Conditional Distributions	219
	Maarten Marsman, Timo B. Bechger, and Gunter K. J. Maris	
13	To a or not to a: On the Use of the Total Score	251
	Bas T. Hemker	
14	A Bayesian Test for the Association of Binary Response Distributions	271
	Rudy Ligtvoet	
15	Efficiency and Effectiveness of Teacher-Informed Targeting Testing from Different Perspectives	283
	J. Hendrik Straat, Renske E. Kuijpers, Kimberley Lek, and Wilco H. M. Emons	
 Part IV New Psychometrics		
16	The Hierarchical Model for Response Times: Advantages, Limitations, and Risks of Its Use in Measurement Practice	307
	Jesper Tijmstra and Maria Bolsinova	
17	Computer-Adaptive Testing with Fewer Assumptions	327
	Jules L. Ellis	
18	Validity Indices for Interpreting Informant Discrepancies in ADHD Assessment	345
	Judith M. Conijn, Mengdi Chen, Hanneke van Ewijk, and L. Andries van der Ark	
19	Computerized Adaptive Testing Without IRT for Flexible Measurement and Prediction	369
	L. Andries van der Ark and Niels Smits	

20 On the Relationship Between Unidimensional Item Response Theory and Higher-Order Cognitive Diagnosis Models..... 389
Jimmy de la Torre and Kevin Carl Santos

21 A Sparse Latent Class Model for Polytomous Attributes in Cognitive Diagnostic Assessments 413
Siqi He, Steven Andrew Culpepper, and Jeff Douglas

Author Index..... 443

Subject Index 451

About the Authors

Terry Ackerman is a distinguished visiting professor at the University of Iowa in the Department of Psychological and Quantitative Foundations. His main research interests include applications of multidimensional item response theory, differential item functioning, and assessing multidimensionality. He works as a consultant to several educational testing and certification companies and has served as president of the National Council on Measurement in Education and the Psychometric Society.

Timo B. Bechger was a senior research scientist with Cito, was a lead research scientist at ACTNext, and is currently senior scientist at TCS. His research specializes in item response theory and test fairness. Timo co-founded the dexter project which aims to develop and disseminate open-source software for professional educational measurement and share the ideas behind it.

Maria Bolsinova is an assistant professor in the Department of Methodology and Statistics at Tilburg University. Her research focuses on item response theory, with a specific interest in response time modeling, Bayesian item response theory, large-scale educational assessment, and response style modeling.

Mengdi Chen is a PhD candidate at the Research Institute of Child Development and Education, University of Amsterdam. Her current research interest lies in the measurement invariance of instruments assessing teacher-student relationships in cross-cultural contexts.

Brian Clauser is a distinguished research scientist at the National Board of Medical Examiners (NBME) and served as vice president of research and director of the Center for Advanced Assessment at NBME. His research focuses on differential item functioning, standard setting, generalizability theory, and automated scoring of simulation and performance assessment.

Judith M. Conijn is a researcher at the Kohnstamm Institute, a knowledge and research center in the field of education, child rearing, and child welfare in the Netherlands. Most of her publications concern statistical methods for detecting

aberrant response patterns in questionnaire data, such as IRT-based person-fit statistics. Her recent research also focuses on various educational topics.

Steven A. Culpepper is Professor of Statistics at the University of Illinois Urbana-Champaign. He is editor of the *Journal of Educational and Behavioral Statistics*. Professor Culpepper works in latent variable modeling for applications in education, psychology, and other social sciences. He has worked in item response modeling, latent class modeling, cognitive diagnosis, and factor analysis and has made contributions to Bayesian approaches for estimation of complex models.

Paul De Boeck is Professor of Quantitative Psychology at The Ohio State University. His work mainly concerns the meaning and issues of measurement in psychology (e.g., explanatory measurement, local dependencies, differential item functioning, and IRTrees).

Jimmy de la Torre is head and professor in the Human Communication, Development, and Information Sciences Unit in the Faculty of Education at The University of Hong Kong. He is also currently a chair professor at the National Taichung University of Education in Taiwan, and an honorary professor at Universidad Autonoma de Madrid in Spain. His primary research interests are in the field of educational and psychological testing and measurement, with a particular emphasis on item response theory, cognitive diagnosis modeling, and the use of assessment to inform instruction and learning.

Jeffrey A. Douglas is Professor of Statistics at the University of Illinois Urbana-Champaign. He is currently president-elect of the Psychometric Society. Professor Douglas works in latent variable modeling with applications in medicine, clinical trials, education, and psychology.

Jules L. Ellis is an associate professor in the Behavioural Science Institute, Faculty of Social Sciences, Radboud University Nijmegen. His research focusses on nonparametric IRT, testing unidimensionality, multiple testing, and standards of reliability.

Wilco H. M. Emons is an assistant professor in psychometrics at Tilburg University in the Netherlands. His expertise includes psychometrics, longitudinal assessments, test design and validation, and multimedia-based testing. His current research interests focus on individual change assessment and innovative assessment methods in education and psychology.

L. Robert Gore is an assistant member in the Health Outcomes and Behavior Department and Department of Biostatistics and Bioinformatics at Moffitt Cancer Center, where he collaborates to provide methodology and measurement consultation to researchers in healthcare equity, mHealth, and behavioral aspects of cancer prevention and care. In his own research, he is interested in the application of mathematical models of cognition in clinical contexts. In addition, he is experienced in psychometric analyses of personnel selection tests.

Siqi He is a PhD student in the Quantitative Psychology Program at the University of Illinois Urbana-Champaign. She is interested in using latent variable modeling, machine learning methods to solve practical problems in educational, and psychological measurement.

Bas T. Hemker started in 1996 as a senior research scientist at the Research and Innovation Department of Cito (National Institute for Educational Measurement) in the Netherlands and has been working there ever since. His expertise includes psychometrics, large-scale assessment, educational measurement, test development, and evaluation of test quality. As an advisor to ministries of education on four different continents, he has contributed to the design of national assessment frameworks in several countries.

Willem J. Heiser is Professor Emeritus of Data Theory, Psychometrics, and Statistics at Leiden University (the Netherlands) and former editor-in-chief of *Psychometrika* (1995–1999) and the journal of classification (2002–2015). His current research interests are multidimensional scaling of networks, analysis and prediction of rankings, and the history of psychometrics, statistics, and psychology.

David J. Hessen is an assistant professor in the Methods and Statistics Department in the Faculty of Social Sciences at Utrecht University in the Netherlands. His research interests are in psychometrics in general and latent variable modelling in particular.

Renske E. Kuijpers has been working as a scientific researcher and psychometrician in the Department of Research & Innovation at the Dutch National Institute of Educational Measurement (Cito) since 2019. She completed her PhD at Tilburg University on the use of marginal models in a measurement context. At Cito, she is involved in several operational and research projects, including the norming of the Dutch central exams for secondary education, the norming of student monitoring systems for primary education, and several Dutch national surveys of student achievements.

Kimberley Lek is a scientific researcher in the Department of Research & Innovation at the Dutch National Institute of Educational Measurement (Cito) since 2019. She completed her PhD at Utrecht University on a study into the transitions of students from primary to secondary education and the role teachers' advice and assessments have in this transition. Currently, she participates in several projects on innovative assessment methods, with a focus on motivational aspects in assessments and accessible score reporting. Her research is very much driven by her motivation to expose inequalities in education where necessary and increase equal opportunities where possible.

Rudy Ligtoet is a postdoctoral researcher at the Institute of Empirical School Research and the Institute of Sociology & Social Psychology at the University of Cologne, Germany. His research focusses on psychometrics and the Bayesian analyses of discrete data.

Richard Luecht is Professor of Educational Research Methodology at the UNC-Greensboro, where he teaches graduate courses in applied statistics and advanced measurement. His research includes technology integration in assessment, advanced psychometric modeling and estimation, and the application of assessment engineering (AE). He has designed numerous algorithms and software programs for automated test assembly and devised a computerized adaptive multi-stage testing framework used by several large-scale testing programs.

Ye Ma is a psychometrician/research scientist at Amazon Web Services, Training and Certification Department. Her main research interests include statistical models, such as item response theory, and multidimension item response theory and its applications under psychometric settings, such as computerized adaptive testing, test reliability, and differential item functioning. She also studies the application of machine learning techniques with licensure/certification exams using various data types.

Gunter K. J. Maris was a full professor of psychological methods at the University of Amsterdam, a principal research scientist with Cito, senior director of advanced psychometrics at ACTNext, and is currently senior scientist at TCS. His research focuses on when and why learning does, or does not, happen, and when education does, or does not, work. Gunter has contributed to the founding of network psychometrics as an independent field of research, and to the advancement of online real-time rating systems for educational measurement.

Maarten Marsman is an assistant professor in the Psychological Methods group of the University of Amsterdam. His main interests are computational statistics, Bayesian statistics, and network psychometrics.

Rob R. Meijer is a full professor in the Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, the Netherlands. His research focuses on applied psychometrics, educational and personnel selection, and decision-making through tests.

Dylan Molenaar is an assistant professor at the University of Amsterdam. His research interests include psychometrics in general, and psychological measurement, test theory, factor analysis, and item response theory in particular.

Marvin Neumann is a PhD candidate in the Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, the Netherlands. His research focuses on personnel and educational selection, mechanical prediction, and the scientist-practitioner gap in assessment and selection

A. Susan M. Niessen is an assistant professor in the Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, the Netherlands. She has published in the areas of personnel and educational selection, test-use and decision-making, predictive validity, applicant perceptions, and test bias.

Kevin Carl Santos is an associate professor in the Educational Research and Evaluation Area in the College of Education at the University of the Philippines-Diliman. He is also a senior research fellow at the Assessment, Curriculum, and Technology Research Center at the same university and an honorary fellow at the University of Melbourne Graduate School of Education. His research interests include cognitive diagnosis models, item response theory, and analysis of large-scale assessment data.

Niels Smits is an associate professor at the Research Institute of Child Development and Education, University of Amsterdam, the Netherlands. His research focuses on psychological testing and methods for short assessments, and he is also interested in applied statistics, machine learning, and learning analytics.

J. Hendrik Straat is a senior psychometrician in the Department of Research & Innovation at the Dutch National Institute of Educational Measurement (Cito). He completed his PhD at Tilburg University, for which he studied item-selection methods for constructing Mokken scales. Throughout his career at Cito, he was involved as psychometrician in several assessment programs including the national central exams and the end-of-primary school test. Currently, he is the lead psychometrician for the national exams and participates in various studies on targeted testing, for example, within the context of acquiring Dutch language proficiency in vocational education.

Jesper Tijmstra is an assistant professor in the department of Methodology and Statistics at Tilburg University. His research focuses on item response theory, with a specific interest in response time modeling, response style modeling, evaluating model assumptions, and nonparametric item response theory.

L. Andries van der Ark is Professor of Psychometrics at the University of Amsterdam and director of the Graduate School of Child Development and Education at the University of Amsterdam. His main research interest is the development of statistical models for test and questionnaire data, especially (nonparametric) IRT models, latent class models, and marginal models. He has also developed methods for assessing test-score reliability, assessing interrater reliability, constructing test norms, and handling missing data.

Hanneke van Ewijk is a senior researcher and behavioral scientist. Her research focuses on children with developmental difficulties such as ADHD, neuropsychological functioning, and brain development.

Joost R. van Ginkel is Assistant Professor of Methodology and Statistics in the Department of Psychology at Leiden University. His major interests include missing data, multiple imputation, and statistics in general. His research mainly concerns the development of new statistical procedures within the framework of multiple imputation.

Matthias von Davier is the J. Donald Monan, S.J. Professor in Education at the Lynch School of Education at Boston College (BC) and serves as TIMSS &

PIRLS International Study Center's executive director. His research areas include item response theory, invariance and linking, diagnostic classification and mixture models, machine and deep learning, computational statistics, model fit, and methodologies used in large scale educational surveys.

Bernard P. Veldkamp is Professor of Research Methodology and Data Analytics at the University of Twente, The Netherlands. He is head of the Learning, Data and Technology Department in the Faculty of Behavioral Management and Social Sciences. His expertise includes psychometrics, educational and psychological measurement, computer-based assessment, computerized adaptive testing, and the use of artificial intelligence in the social sciences.

Part I
General Perspectives on Psychometrics

Chapter 1

Early Roots of Psychometrics Before Francis Galton



Willem J. Heiser

Abstract Although one of the flagships of psychometrics, factor analysis, could not have been invented without Francis Galton's (1822–1911) groundbreaking concept of correlation, some other psychometric concepts had been explored already before his time. Christian Thomasius (1655–1728) pioneered personality assessment using numerical rating scales and introduced a first notion of psychometric reliability. It was Christian Wolff (1679–1754) who coined the term “psychometria” and who identified the basic difficulty of finding a suitable unit for measurement of psychological variables. Halfway the nineteenth century, Gustav Fechner (1801–1887) not only founded psychophysics but also introduced before Galton the statistical approach to the analysis of psychological data—which is so typical for psychometrics in general. He also developed some pathbreaking experimental designs for data collection, as well as the notions of a psychological scale and the psychometric function.

1.1 Introduction

When the first laboratory worldwide for both research and teaching of experimental psychology was founded in Leipzig (1879) by Wilhelm Wundt (1832–1920), it immediately attracted many students, not only from Germany and neighboring countries in Europe but also from the United States. The first of them, James McKeen Cattell (1860–1944), while discussing the topic of his dissertation with Wundt, experienced a bit of a culture shock:

The author is grateful to Denny Borsboom, Larry Hubert, Jacqueline Meulman, Stephen Stigler, and Robert Tijssen for their useful comments on an earlier draft of this chapter.

W. J. Heiser (✉)
Leiden University, Leiden, The Netherlands
e-mail: heiser@fsw.leidenuniv.nl

As a large part of the work of the laboratory was then on reaction-time experiments, it was not surprising that such a subject fell to my lot, and it was fortunate, for I had already in America begun experimental work on the time of sensori-motor processes. Wundt, however, was mainly interested in experiment for the aid it gave to introspection, and the subject assigned to me was to react as soon as I saw a light and in a second series to react as soon as I recognized its color, with a view to analyzing the factors of apperception. This I could not do, and in my second interview with Wundt I presented an outline of the work I wanted to undertake, which was the objective measurement of the time of reactions with special reference to individual differences. Wundt said that it was ‘ganz Amerikanisch’; that only psychologists could be the subjects in psychological experiments. (Cattell, 1921, p. 156)

Despite Wundt’s negative reaction, Cattell was allowed to start his project as he conceived it, with his own apparatus and in his own room, and Wundt prepared him graciously for his doctorate examination. His dissertation work in Leipzig was published as Cattell (1886).

Meanwhile in England, Francis Galton (1822–1911) had founded an anthropometric laboratory, at the occasion of the International Health Exhibition in London (1884–1885), where he measured and recorded “the chief physical characteristics of man,” including “keenness of sight, colour sense and hearing” (Galton, 1885). In total, he was able to measure 9337 ordinary persons on 17 variables. Attracted by Galton’s interest in empirically studying individual differences between people, Cattell went to London soon after leaving Leipzig and joined Galton in his research projects. That joint effort resulted in the paper *Mental tests and measurement*, of which Cattell wrote the main part, opening with the programmatic statement:

Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement. A step in this direction could be made by applying a series of mental tests and measurements to a large number of individuals. The results would be of considerable scientific value in discovering the constancy of mental processes, their interdependence, and their variation under different circumstances. (Cattell & Galton, 1890, p. 373)

Cattell continues to describe a long series of 60 tasks concerning sight, hearing, taste and smell, touch and temperature, sense of effort and movement, mental time, and memory. These were the type of tasks used by Fechner, Wundt, and Helmholtz, the pioneers of experimental psychology. However, Cattell broke with their habit of using only a few psychologists as subjects and with the high priority these German pioneers gave to finding general laws. Wundt’s low regard of individual differences continued to dominate experimental psychology for a long time. Only quite recently, we have seen several attempts to bring together classical tasks used by experimental psychologists with a serious look at individual differences. For example, Schmiedek et al. (2007) considered individual differences in reaction time and their relations to working memory capacity (WMC) and intelligence, Wilhelm et al. (2010) studied individual differences in face recognition, and Wilhelm et al. (2013) used confirmatory factor analysis to obtain a broader perspective of WMC as an individual difference construct. After more than a century of delay, these researchers are effectuating the program Cattell, Galton, and others originally had in mind.

Returning to the above quote from Cattell and Galton (1890), key terms are *constancy*, *interdependence*, and *variation*, which shows the influence of Francis Galton's statistical ideas. It is also remarkable that the type of tasks listed was far removed from what one might suppose mental testing is primarily about: ability, personality, or character. The paper has an appendix with comments by Galton, where he pays attention to exactly this aspect:

One of the most important aspects of measurement is hardly if at all alluded to here and should be emphasized. It is to obtain a general knowledge of the capacities of a man [...] In order to ascertain the best points for the purpose, the sets of measures should be compared with an independent estimate of the men's powers [...]. The sort of estimate I have in view and which I would suggest [...] is something of this kind,—'mobile, eager, energetic; well-shaped; successful at games requiring good eye and hand; sensitive; good at music and drawing. (Cattell & Galton, 1890, p. 380)

We also see a quest here for establishing the *validity* of a mental test in being able to identify important characteristics of a gentleman. However, it must be noted that these were only plans; Cattell and Galton never actually collected and analyzed these type of personality data! By contrast, there were certainly earlier attempts of mental testing of personality, and one of them (in the seventeenth century) that attempted to assess reliability will be described in Sect. 1.2 of this paper, which discusses Christian Thomasius.

A much earlier example is ability testing in ancient Greece, which represented the different facets of the ideal Greek citizen. These tests were primarily of a vocational nature, but also included athletic abilities (Doyle, 1974). Even still earlier, in ancient China:

The great Chinese philosopher and educator Confucius (551–479 B.C.) first classified people into three categories on the basis of intelligence: (1) people of 'great wisdom'; (2) people of 'average intelligence'; (3) people of 'little intelligence'. Confucius also made personality assessments of his students. (Zhang, 1988, p. 101)

Moreover, it is well-known that ancient China had a Civil Service Examination system, even though there is uncertainty about exactly how old it is (Bowman, 1989). In any event, in these earlier examples of mental testing, we do not find any notion of reliability or some form of advanced statistical analysis of the test results, which are basic elements of psychometrics. In the Dutch literature, Kouwer (1963) has taken perhaps the broadest possible historical view on the development of systems to characterize personality, but again without paying attention to actual measurement or quantification. However, for educational testing we do know when serious quantification started. Stigler (1992) and Mellenbergh (2011, p. 18) have identified that the first psychometric papers on the analysis of examination scores were published by Edgeworth (1888, 1890), in which he discussed the scaling of exams by using the normal distribution, correction for examiners bias, reliability of a single examiner, and more.

It is often stated that the modern form of intelligence testing started with the psychologist Alfred Binet and psychiatrist Theodore Simon in the period 1895–1910 (Boake, 2002), which was also the period that modern personality research

started. As a particularly interesting example of the latter, Heymans and Wiersma (1906) collected large-scale questionnaire data on personality characteristics—such as introversion-extraversion and emotionality—in 437 families, including 3 generations of each family, summarized in 90 4-way contingency tables (*cf.* Heiser, 2008). But it should be noted, as argued by Mülberger (2017), that the emergence of mental testing in this period was more widespread and gradual than just the Binet-Simon breakthrough (as is also evident from Spearman’s (1904) extensive summary of previous correlational studies of mental test data).

There is no doubt that Galton’s major contributions to psychometrics have been, as pointed out by Drenth and Sijtsma (1990, pp. 4–5), his keen interest in individual differences, the need to work with standardized research designs, and his conceptualization of regression and correlation. How Galton developed the concept of correlation has been nicely described by Stigler (2010), while Walker (1929, pp. 92–102) explained why earlier writers in the nineteenth century hovered on the verge of discovery of correlation, but did not actually uncover it.

However, to regard Galton as “the founding father of psychometrics” (e.g., Furr & Bacharach, 2008, p. 9) is perhaps one step too far, for there are earlier roots of psychometrics, at least if we take a broader view of the field and do not restrict it to mental testing. Such a broader view was sketched by Jones and Thissen (2007), and the present paper tries to add three historical lines to their paper. Apart from the already mentioned early attempt to assess reliability, we will also discuss how the name and perspective of a discipline of psychometrics was conceived by Christian Wolff in the eighteenth century. The third and most important early root is the groundwork given by Gustav Fechner’s psychophysics¹.

1.2 An Early Notion of Reliability: Thomasius’ Numerical Rating System of Personality

Rating has been a method of assessing the degree of some natural characteristic by a human observer since ancient times (e.g., temperature, *cf.* Wright, 2016). However, according to McReynolds and Ludwig (1984, 1987) and Ramul (1963, p. 657), it was the German Enlightenment philosopher and jurist Christian Thomasius (1655–1728) who devised and applied the first quantitative rating scales for personality attributes of individuals. His purpose was to characterize each individual on 4 scales with 12 categories ranging from 5 to 60 in steps of 5 so that a *personality profile*

¹ As this chapter is part of a Festschrift in honor of Klaas Sijtsma, someone with a keen interest in the evolution of psychometrics (e.g., Van der Heijden and Sijtsma, 1996; Sijtsma and Junker, 2006; Sijtsma, 2016), who defends a position of psychological measurement between physics and statistics (Sijtsma, 2012), it is my hope and expectation that he welcomes these new trace lines in our history.

could be formed, and he conceived a concept of *interrater reliability* by asking several observers to rate the same person.

Thomasius' rating scales followed from his overall theory of personality, which was announced in a first programmatic publication entitled *New Discovery of a Solid Science, Most Necessary for the Community for Discerning the Secrets of the Heart of Other Men from Daily Conversation, Even Against Their Will* [English translation of Thomasius (1692a) by McReynolds and Ludwig (1984)]. The motivation for formulating this empirical approach using practical field work, interviews, and informal discussions with the common citizen was to arrive at the kind of knowledge a politician needs for effective policy making. Thomasius was convinced that a science of policy should not be legalistic, let alone philosophical. He started an autonomous discipline addressing what makes people tick (Barnard, 1971).

Further details about his rating scale system were provided by a second publication in the same year, entitled *Further Elucidation by Different Examples of the Recent Proposal for a New Science for Discerning the Nature of Other Men's Minds*. This English translation of the German title of Thomasius (1692b) is again by McReynolds and Ludwig (1984), and they also provided a translated version of the five basic postulates of his personality theory:

- I. There are four major inclinations from which all other inclinations spring. These are:
 1. Rational love [*Vernünftige Liebe*]
 2. Sensuousness [*Wollust*]
 3. Ambition [*Ehrgeiz*]
 4. Acquisitiveness [*Geldgeiz*]
- II. All human beings are characterized by these inclinations and all possess some part of each of them.
- III. At all times one of the four inclinations is dominant in a person.
- IV. The difference among persons in human inclinations must be recognized not only from the dominant inclination but also from the proportion of the other three.
- V. One can appropriately assign 60 points to the strongest inclination and 5 points to the weakest (or at times more) and then judge the remaining two in accordance with the difference between the 60 points and the value of the lowest inclination. (Thomasius, 1692b, p. 239)

The four inclinations indicated in postulate I are the basis of the four rating scales that are to be used to rate any individual on the basis of conversations with the rater (a trained observer). According to McReynolds and Ludwig (1984), "All kinds of data went into the rating determinations—educational, occupational, and familial information about the subject; reports of his daily habits; interpersonal styles; behaviors that the individual found pleasurable; and so on." They also comment that Thomasius' description of *Rational love* comes close to what we would now call *Altruism*, that *Sensuousness* is concerned with *Hedonic tone* (seeking pleasure and avoiding pain), that *Ambition* must be understood as *Social ambition*, and that *Acquisitiveness* not only relates to a *Passion for money* but also to *Stinginess* and *Envy*.

Regarding the numerical rating categories on the four attributes, it is noteworthy that they are seen as proportions (postulate II) and kept within the range 5–60

(postulate V). It is plausible that this particular choice of values for the rating categories was inspired by the usual scale markers in time measurement (60 min in an hour, 12 months in a year). Furthermore, only the dominant attribute gets the maximal score of 60 (postulate III), and the other attributes need to be seen in proportion to the dominant one. Due to the aim to compare patterns of attribute proportions between individuals (postulate IV), the whole approach seems to fit into what two and a half centuries later has been called *Q-methodology* (Stephenson, 1936, 1953; Cattell, 1952)—a small, but basic part of psychometrics.

Thomasius included a section in *Further Elucidation* called “About the Test of Certainty of This Science,” beginning as follows: “Just as in mathematics, where there is no better way to check to see if one has calculated correctly than to repeat the process two or three times in order to find out if the sum is the same, I have thought that in the discovery of other truths, regardless of what the discipline it may be, this method might be the best way of checking” (quoted in McReynolds and Ludwig (1984)). He then gives an example of a single individual who was rated by himself and by two students who had been trained well in the method of scoring. It turned out that the three patterns were very close, a sign of considerable *interrater reliability*. Even without recourse to a numerical reliability coefficient, the expression of the patterns in quantitative terms obviously facilitated comparisons enormously. Note that this form of interrater reliability is different from the more usual form at present, in which for each attribute separately the ratings of different raters across individuals would be compared.

What was the impact of Thomasius’ quantitative methodology? The short answer is: by the end of the eighteenth century, he was not taken seriously anymore. According to Barnard (1971), “To his German contemporaries and near-contemporaries Thomasius was something of an idol. Nineteenth-century intellectuals—Hegelians in particular—generally dismissed him as an unsystematic, facile eclectic, and only the present century has witnessed a moderate revival of interest in him, though scarcely beyond the confines of Germany.” However, there does seem to be a renewed interest in the Anglo-Saxon world for his “desacralization of philosophy” (Hunter, 2000). Moreover, it is the irony of history that after Thomasius was forgotten, *Q-methodology* is thriving presently in political science and communication science (cf. Brown, 1991, 1993; McKeown & Thomas, 2013).

1.3 Qualities of the Soul Can Be Measured: Wolff’s Proposal of Psychometria

Christian Wolff (1679–1754) was “arguably the most eminent German philosopher between Leibniz and Kant, and an important figure in the development of thought about the state and its tasks as well as about the national economy” (Drechsler, 1997). But he was also vitally important for the *Sciences of the Soul* (Vidal, 2011)

and in particular laid the groundwork for their methodology, which he coined *Psychometria* (Ramul, 1960; Feuerhahn, 2004)).

In the beginning of his career, he had chosen to specialize in mathematics, obtaining his doctorate in 1703 at the University of Leipzig, where he was soon invited to become a staff member of the first scholarly journal in Germany, the *Acta Eruditorum Lipsiensium*. Apart from mathematics, he soon expanded into other areas within the *Faculty of Arts*, then including all fields of learning except *Divinity*, *Law*, and *Medicine* (Drechsler, 1997).

Due to the Great Northern War between an alliance of Denmark-Norway, Saxony, and Russia against the Swedish empire, Wolff decided to leave Leipzig in 1706, and he accepted an offer of the University of Halle, where he became Professor of Mathematics, upon recommendation of no less than Gottfried Wilhelm Leibniz (1646–1716). Drechsler relates:

Wolff greatly enjoyed teaching [...] and also began lecturing in what we would today call Philosophy. [...] He was also by then a prolific and celebrated author, and was thus unanimously elected as Fellow of the Royal Society in London. [...] Embarrassed by the fact that a Prussian subject had thus been honored abroad but not at home, the Berlin Academy subsequently made him a member as well. [...]

In 1723, however, Wolff had to flee from Halle in one of the most celebrated dramas in the academy in the eighteenth century. The incident which caused the drama was his farewell address as *Prorector* in 1721. [...] In it, Wolff described the Chinese philosophy and ethics, namely Confucianism, as rather admirable and really as largely in agreement with his own moral principles. Indeed, his lecture submitted *proof that one could find moral truths through the powers of reason of natural Man without the help of divine revelation*. [...] If one follows Wolff's argument, there remains little place for Christian mission; to the contrary, it seems that one could actually learn a few things from the Chinese. (Drechsler, 1997, p. 112)

It became a scandal of immense proportions, where opinion leaders adhering to a strict form of Protestantism forced the King of Prussia to accuse him of gross impiety and to order him to leave the city of Halle and all other Prussian lands immediately. Fortunately, Wolff could use his influential network to escape and obtained the *Papin's chair of Mathematics and Physics*, as well as the *chair of Philosophy* at the University of Marburg, just a week later.

It was in Marburg that Wolff published major works about philosophy and psychology, including *psychometria*. Concerning philosophy, "He reemphasized Leibniz's conviction that mathematics has a role in philosophy. As he wrote in his *Discursus praeliminaris de philosophia in genere* (Wolff, 1963, original work published in 1728), philosophy must use mathematical knowledge. For in philosophy we wish to have complete certitude [...] [and] in many cases, complete certitude depends on mathematical knowledge and demonstrations" (Leary, 1980, p. 155).

Concerning psychology, Wolff's point of departure was Leibniz' doctrine that *Intensity*, *Continuity*, *Variation*, and *Covariation* apply equally well in the material as in the mental realm. What now follows is a summary of Leary (1980, pp. 154–155). With respect to *Intensity*, the concept of *force* in physics corresponds to the concept of *clarity of ideas* in psychology. The principle of *Continuity* states that all

differences in nature are different in degree rather than in kind, whether we consider *physical motion* or *mental consciousness*. *Variation* refers to the principle that every material object and every mental idea undergoes continuous change in the degree of its intensity. Material objects change in *momentum*, and mental concepts change in *amount of consciousness*. *Covariation* refers to the principle that change in one part of the system leads to a (reverse) change in some other part of it. For the material world, it implies that an *increasing force* in one body corresponds to a *decreasing force* in another, while for the mental world it implies that an *increase of clarity* in one idea corresponds to a *decrease of clarity* in another.

In his important work *Psychologia Empirica, methodo scientifica pertractata* Wolff (1962a, original work published in 1732), explained that

The art of discovery (*ars inveniendi*), which involves deducing unknown truths from already known ones, can proceed either *a priori* or *a posteriori*. In the latter case, which is the only one of interest to empirical psychology, findings are based on observation or experimentation (*ex experimentis*). Both are forms of ‘experience’ (*experientia*), that is, of knowledge acquired by paying attention to our perceptions. Observation involves no voluntary alteration of nature, experimentation (*experimentum*), by contrast, requires it. Watching the sky cloud over is an observation, whereas pumping air from a pneumatic machine is experimentation. The *ars observandi* used by physicists, doctors, and above all astronomers, is the proper method of empirical psychology. *Ars experimentandi*, on the other hand, is used only by physicists—even if, Wolff suggested, it could be applied to the whole of philosophy and even to natural theology. (Vidal, 2011, pp. 128–129; footnotes and references to the source omitted here)

In the same work, Wolff also formulated his mathematical law about the magnitudes of pleasure and displeasure: “Pleasure is proportional to the perfections of which we are conscious, as well as to the certainty of our judgments concerning these perfections.” In a footnote he added: “These theorems belong to psychometry, which conveys a mathematical knowledge of the human mind and continues to remain a desideratum. It should teach us how to measure the magnitudes of perfection and imperfection and also the certainty of a judgment, and insofar determine [both measures]” (Ramul, 1960, p. 256). As recently noted by Mei (2021, p. 91), Wolff’s psychometria is “a form of methodological mediation that implies the ability to measure the *effects* of the soul rather than its *substance*. In other words, psychometria allows us to take into scientific consideration the possibility of a first form of the *naturalization* or *mathematization* of the mind.”

More specifically, in his *Philosophia prima, sive Ontologia, methodo scientifica pertractata, qua omnis cognitionis humanae principia continentur*, Wolff (1962b, original work published in 1736) gave a number of examples of how psychometrics could proceed to measure the qualities of the soul. What now follows is a summary of Mei (2021, pp. 91–96), who also gives source references. As a preliminary, Wolff states that each quality is measurable (and calls it a “common prejudice” that not all qualities are measurable). For instance, *density of fluids* is a quality and can be measured with an aerometer, *temperature* can be measured with a thermometer, and the *gravity of air* can be measured with a barometer. Moreover, qualities have a degree, and therefore we have the possibility to establish the *size of the degree*, which he regards as an *imaginary notion* (recall Leibniz’s principle of the continuity

of nature). He mentions *degree of speed* and the notion of *substance* as other examples of imaginary notions. Wolff mentions the following three possibilities for psychometric measurement.

1. Measuring *duration* and *clearness* of psychic phenomena.

Thoughts are not immediate and some time is required to allow human thought to proceed. The term *time* refers to continuous processes and *duration* to the simultaneous existence of several successive things. Time can be represented through the imaginary notion of a straight line consisting of a continuous series of points, so that there is an analogy between time and number. Furthermore, perceptions can be partial or composite. If ideas belonging to a visible object and its corresponding word become clearer over time, it is because the movement of the material ideas is faster. A composite perception consists of several partial perceptions, and if the partial perceptions become clearer, then the corresponding composite ones are perceived distinctly. The greater the number of the particular, clear perceptions, the greater the degree of the distinctiveness of the subsequent composite perception. So it is duration and/or the number of required perceptions that allows measurement of psychic phenomena.

2. Measuring the *intensity* of psychic phenomena: Memory and the imagination.

According to Wolff, if something is distinctly perceived, it is also easier to retain in memory. Therefore, the quality of memory admits different degrees that may vary from individual to individual. We can identify this degree of quality by looking at the time spent holding on to an idea in the mind or to the number of acts by which the reproduced ideas are delivered to memory or with which they are held in memory. Therefore, people with a “great” memory can reproduce the ideas of many things, like those who can remember the whole Bible and can quote each part of it in the right order or those with a “long” memory who can remember a long series of things or events. Imagination also has different degrees, to the extent that it reproduces the ideas of many things, while memory recognizes ideas reproduced. There are individual differences in the quality of the soul, due to a possible diversity of nerve fibers. Body and soul are closely connected and interdependent, which implies that psychometria enables measurement of the *effects* of the soul, instead of measuring the soul as *substance*.

3. *Measuring degrees of attention and individual differences.*

A major pillar of Wolff’s psychometria is that degrees are the “quantities of qualities.” Also, every time we talk about degrees, we do not refer to objects, individuals, or activities, but to *relations* between them. For example, we say that this line is three or five times as thick as another one. Likewise, for intellectual qualities we can say that one person’s ability is larger than someone else’s. Degrees of attention can be greater or smaller depending on (a) how much the sense organs are involved in perception (which can be measured by their arousal), (b) how long mental content is preserved or extinguished, (c) how many different things a person can pay attention to simultaneously, (d) the selectivity with which a person pays

attention to some objects but not others, and finally (*e*) whether someone typically pays attention to actual objects or to imaginary objects. These examples show that levels of attention and individual qualities can be measured by counting relational data.

What happened to Wolff's program of empirical psychology and psychometria? First and foremost, he did not collect any data himself to see how his methodological ideas would work, and neither did his contemporaries. But there were several authors in the eighteenth century who also dealt theoretically with the question of mental measurement (for an overview, see Ramul, 1960 and Vidal, 2011). One of them was Gottlieb Friedrich Hagen (1710–1769), a philosophy teacher at the Bayreuth gymnasium, who was a follower of Wolff and had a position as Adjunct Professor in the faculty of philosophy at Halle in the period 1731–1737. Like Wolff, he wanted his work to have the universal applicability of mathematics while also being socially useful. As Vidal notes:

He imagined psychological experiments [...] that would alter the soul, for example, by scaring people and then observing their reactions. Such experiments could contribute significantly to self-knowledge [...]. Hagen also conceived a *dynametria* to measure the faculties (*dunamis*) of the soul, again within the framework of a sort of quantitative casuistry. He argued that, like the mechanical faculties of the body, the representative faculties of the soul are finite in number; since they vary considerably from individual to individual, they may be compared quantitatively. (Vidal, 2011, p. 130)

Ramul (1960) concluded his pioneering essay by noting that although measurement of mental phenomena attracted the attention of several eighteenth-century scholars, much of what the individual authors had to say were their personal “ideas” with little continuity in their development, except for some of Wolff's students. No one carried out any actual measurements. By contrast, he says, “by far the larger part of the psychological measurements known to us from that time [...] have been carried out not by psychologists (or philosophers) but by naturalists [who studied such elementary phenomena as visual acuity, the size of the blind spot, and the duration of visual afterimages]. And thus the program of [...] psychometry remained wholly on paper in the eighteenth century” (Ramul, 1960, p. 264). Although Wolff's psychometria did influence Immanuel Kant (1724–1804) and Johann Friedrich Herbart (1776–1841) on a conceptual level (Leary, 1980; Sturm, 2006), the definite start of psychometrics had to wait until the second half of the nineteenth century.

1.4 Birth of Experimental Design and Psychological Scaling Methodology: Fechner's Psychophysical Paradigm

Psychophysics is the brainchild of Gustav Theodor Fechner (1801–1889), physicist and philosopher with important contributions to psychology, psychometrics, and statistics. Born in Gross Särchen (a small village in the German region of Saxony) as the son of a clergyman, he started studying medicine in 1817 at the University of

Leipzig and earned a baccalaureate in 1822. But he had a lot of other interests: “at about the same time he began writing a series of sometimes mystical philosophical pieces on the identity of mind and matter, a practice that was to last throughout the rest of his life” (Stigler, 1986, p. 242). He did not finish medicine, however, and got more interested in physics. To earn some money, he began translating the textbook of Jean-Baptiste Biot (1774–1862) on experimental physics from French into German and started lecturing in 1824. Then he published a paper on the galvanic battery (Fechner, 1831) that was inspired by the pathbreaking experimental work of Georg Simon Ohm (1787–1854) on the laws of electricity published in 1827. It made his reputation as a physicist, and he was elected as extraordinary professor of physics at Leipzig in 1831, and in 1834 he was promoted to full professor of physics at the same university.

Stigler (1986) has emphasized the lasting influence of Ohm’s work on the young Fechner, because his 1831 paper already:

bore the hallmark of Fechner’s later work. Even though it made no use of probability in its analysis, it was an extensive, painstakingly-detailed account of a series of multifactor experiments. Everything that could be varied was varied; everything that could be measured was measured; everything that could be recorded was recorded. And in all this mass of detail (the record of the experimental results alone covers about 200 pages) he did not lose sight of overall objectives. (Stigler, 1986, p. 243)

Several biographies of Fechner have associated his turn to psychophysics to a period of illness and personal crisis in his early forties after he had ruined his eyesight by doing experiments in subjective color perception, looking often at the sun through colored glass. He recovered when he entered his garden not wearing the mask that covered his eyes for many years. Overwhelmed by how beautiful everything looked, especially the flowers, it seemed to him “like a glimpse beyond the boundary of human experience” (the last quote is from Fechner’s autobiographical notes as cited in Murray, 2021, pp. 76–79). But Stigler is not impressed:

As appealing as such stories are as devices for raising the origin of scientific ideas to the level of heroic myth, they do not seem to be essential to an understanding of Fechner’s intellectual development. The urge to experiment, the interest in physics and both mind and body, and an ambition to influence human thought—all the essential ingredients were already in place in the 1820s. The Fechner who by 1855 had begun the extensive experimentation that led to his *Elemente der Psychophysik* was essentially the same Fechner who had devoted two full years to the study of electrical current in 1829–1831. (Stigler, 1986, p. 243)

For additional intellectual influences of earlier scientists on Fechner, these go back “nearly a hundred years to the measurement of sensitivity and of the discriminatory capacity of the senses as accomplished by physiologists and other natural philosophers” (Boring, 1961). At this point, we cannot elaborate on that story because of our focus on psychometrics, but we have to give a brief introduction to Fechner’s Law and how to check it experimentally.

In the next summary, we use Fechner’s notation, as given in the short excerpt from the 1860 *Elemente der Psychophysik* reproduced in Miller (1964, ch. 4). It is well-known that *Weber’s law* states that the sensation difference between two

stimuli remains constant when the relative stimulus difference, or the increase in one stimulus, remains constant. Let the stimulus which is increased be called β and the small increase $d\beta$, where the letter d is to be considered simply as a sign that $d\beta$ is a small increment of β . The relative stimulus increase therefore is $d\beta/\beta$. Choosing d so that two sensations are “just noticeably different” (*jnd*), Fechner took the *jnd* as the unit of sensation, which could be counted to form *magnitudes of sensation*. Let the sensation that is dependent upon the stimulus be called γ and the small increment of sensation be $d\gamma$. Now, Weber’s law is usually stated as $d\beta/\beta = \text{constant}$. By invoking the assumption that the change in sensation $d\gamma$ is equal for all *jnds*, Fechner could transform Weber’s law into

$$d\gamma = \kappa \frac{d\beta}{\beta}, \quad (\text{fundamental formula})$$

where κ is a constant dependent on the units for γ and β . Fechner’s next step was to consider the fundamental formula as a differential equation and integrate it. The result is

$$\gamma = \kappa \log \frac{\beta}{b}. \quad (\text{measurement formula})$$

Here b is the threshold value of the stimulus β , a value at which the stimulus is no longer detectable, called the *stimulus limen* L or RL (from the German *Reiz Limen*), corresponding to $\gamma = 0$. The scale of γ is then the number of *jnds* that a sensation is above zero. Finally, Fechner made one more step by regarding b as the unit for the measurement scale of the stimulus β , so that the measurement formula simplifies to what he called the *metric formula*:

$$\gamma = \kappa \log \beta, \quad (\text{metric formula})$$

the form usually found in the textbooks, where the metric formula is usually called *Fechner’s law*. Fechner himself preferred to keep using the name *Weber’s law*, out of respect for his physiology professor in medical school, Ernst Heinrich Weber (1795–1878) himself (for more on Weber’s importance as a pioneer of quantitative psychology, especially his experiments on the sensitivity of the touch sense, see Murray, 2021, ch. 3). But according to Stigler (1986, p. 243), such emphasis on Weber might be misleading in the light of the fact that Fechner was so well acquainted with the early work of Ohm who developed a similar logarithmic relationship between the loss of force in a current and the length of a wire.

Fechner’s law generated substantial objections and a lot of discussion among the psychologists of his time and later (for brief overviews, see Boring, 1961; Stevens, 1961; Zudini, 2011, pp. 82–84). Also, it might be noted by the reader, as did George Miller, that

Fechner’s law relating subjective sensation to objective stimulation is exactly the same as D. Bernoulli’s law relating subjective utility to objective money. But Fechner’s law was

immediately strengthened by his proposals for psychometric methods of measurement, whereas methods for measuring the subjective magnitudes that Bernoulli was talking about were not developed until the middle of the twentieth century. A theory is good, but a theory plus measurements is a great deal better. (Miller, 1964, p. 99)

As a matter of fact, Fechner did know that Daniel Bernoulli (1700–1782) came up with the concept of diminishing marginal utility and suggested a logarithmic function for it. In particular, in *Elemente der Psychophysik*, he quoted Bernoulli's treatise *Specimen theoriae novae de mensura sortis* published in 1738, where Bernoulli writes: "Certainly the value must not be estimated from the price of the thing, but from the advantage acquired therefrom. The price is estimated by the thing itself; the advantage, by the state of the persons involved. Thus, without doubt, the gain of 1000 ducats is far more important for poor persons than for rich persons, although the amount is the same for both. [...] Thus, it is indeed exceedingly probable that any small advantage adds to the ultimate good in reciprocal proportion to their status of the people involved" (quoted in Fechner, 1860, p. 197). But immediately after this quote, Fechner remarks: "He bases his differential formula [...] and his logarithmic formula [...] on these considerations. We later base the same on Weber's law in a more general way" (Fechner, 1860, pp. 197–198). A theory is good, but a theory motivated by verified empirical regularities is even better! He also points out the role of Laplace, who developed Bernoulli's idea further in his *Théorie analytique des probabilités* (1812), and to Poisson, who mentioned and accepted it in his *Recherche sur la probabilité des jugements en matière criminelle et en matière civile, précédés des règles générales du calcul de probabilités* (1837).

In checking the logarithmic law for sense data, values of the physical stimulus R (from German *Reiz*) were taken as its strength β in a measure valid for the chosen domain (weight, touch, brightness, pitch, and so on) and given in terms of b as the unit of measurement (*cf.* the measurement formula). Let us elaborate what Fechner's proposals were for finding scale values γ of the psychological response S (from German *Sensation*). [Confusingly, the initials of the German terms are the reverse of the English terms Stimulus and Response!] There are three basic methods and two additional ones, which are regularly described in the classical texts of Guilford (1936), Brown and Thomson (1940), and (partly) Bock and Jones (1968). They all refer to Titchener (1905) as the basic source. Here is a brief description, where we follow the distinction suggested by Brown and Thomson (1940) to distinguish between names for *methods of experimenting* in order to collect data (*experimental design* in modern terms) and *processes of calculation* after the data have been collected (*analysis methods*). The basic psychophysical methods are as follows.

1. *Method of reproduction or adjustment.* This experimental design is one of the oldest and most fundamental of psychophysical methods. According to Titchener (1905, p. 160), it is "a free gift to psychophysics from the exact sciences of physics and astronomy." Fechner introduced it in *Elemente der Psychophysik* with tactual and visual measurements. In his own words, from the 1882 revision of the *Elemente*, English translation by Guilford (1936, p. 25):

A certain distance, e.g., between compass points or parallel threads, is presented. This I call the normal distance. I am to make another distance, the error distance, as nearly equal to this as it can be made by eye. First of all, starting from an error distance that is too large or too small, I adjust it roughly, in an irresponsible sort of way, to apparent equality with the normal. Then I consider whether or not it really corresponds to sensible equality, and I shift the boundary of the error distance, thread or compass point, to and fro—until I seem, with a definitive adjustment, to have touched equality as closely as I may. (Fechner, 1882, p. 105)

In this case, the stimulus was an interval. More generally, the task for the subject is to *adjust* or *reproduce* a variable stimulus V , so that it appears subjectively equal to a given standard or *comparison stimulus* C . Anyway, the task is repeated a large number of times, so that we get a distribution of numerical adjustments. Method of analysis for this design is called the *method of average error*, meaning that we take the arithmetic mean of the observed scale values of the reproduced V s. This choice was driven by the time-honored decomposition $Observation = Truth + Error$, used by astronomers in the 1820s, who had no doubt in their mind that they were “after something real, definite, objective, something with an independent reality outside of their observations, a genuinely Platonic reality inherited from the then-unshakable edifice of Newtonian theory” (Stigler (1992, pp. 61–62). It is also the basis of *classical test theory*, pioneered by Spearman (1910), and of *signal detection theory* (Link, 1994).

2. *Method of limits or method of minimal changes*. Primary use of this experimental design is the determination of sensory thresholds. For the stimulus limen RL , the experimenter decreases a variable stimulus V in small steps until it is no longer detected. For the *difference limen* DL , we have a pair of stimuli, V (a *variable* stimulus) and C (a *constant* or *standard* stimulus). V is first made equal to or slightly smaller than C and then decreased in small steps until the observer calls it *just noticeably smaller* than C . If there are N repetitions of the procedure, the simplest analysis method used is to calculate again the mean of the midpoints between C and the last V . According to Guilford (1936, p. 115), the original *method of just noticeable differences*, which was already used by Weber in 1829 to measure *jnds* in passive pressure and lifted weights, presupposed that a human observer can recognize a *jnd* when he sees one. Weber would follow the procedure described above and was ready when the observer reported that he perceived a *jnd*. Fechner recommended a change in the method that was an improvement and has been permanently adopted. The change is to also start from positions of *extreme inequality*; now, the sequence results in the new notion of a *just not noticeable difference* (a *jnnd*), which is usually slightly smaller than the *jnd*. One then takes the average of the *jnd* and the *jnnd* as the true limen. The occurrence of different limiting values for these two starting positions suggests the presence of a *perceptual hysteresis* effect. As a matter of fact, Hock and Schöner (2010) have recently considered several possible mechanisms for such effects, detectable by a *modified* method of limits. There were already more variations in experimental design earlier, for which Urban (1907) is a good source.

3. *The constant method (or method of constant stimuli) and the method of right and wrong cases.* These methods can be used for determination of stimulus limens (*RLs*), differential limens (*DLs*), and equal sense distances, as well as the determination of other psychological scale values outside the strict realm of psychophysics. It is regarded as the most satisfactory of all Fechnerian scaling methods. The experimenter selects in a pilot study a limited number of stimuli, usually four to seven, that are going to be *constant* during the experiment. Let us call them $C_j, j = 1, \dots, n_c$. Next, an additional stimulus T is selected as the *target*, somewhere on the physical continuum depending on the specific purpose of the experiment. For example, if the target is a stimulus limen *RL*, T is typically defined as the physical stimulus that has a probability equal to 0.5 of producing a response, which corresponds to a scaled value of $\gamma = 0$ on the psychological scale. Each constant stimulus C_j is then paired with the target T , and these pairs are presented either simultaneously or successively to the observer in prearranged or random order. The observer has to tell which of the two is “greater than” or “above” the other one, or the reverse. The presentation of each pair is repeated a large number of times, say n_j times, and the observations can be summarized in n_c relative frequencies p_j . For a differential limen *DL*, the C_j and T_j are in fact *pairs of stimuli*, and this case is called the *method of constant stimulus differences*, in which we have *pairs of pairs*, which are compared in terms of the magnitude of their sense differences.

The major analysis method developed by Fechner was called the *method of right and wrong cases*. This ingenious method finds the scale value of the target as the point on the physical scale that is the *median* of the discrete distribution of the comparison stimuli. It is by definition the location for which half of the judgments “ C_j greater than T ” are right (those to the right of the median), while the other half of the judgments “ C_j greater than T ” are wrong (those to the left of the median). Now, how do you find the median of a discrete distribution? Several simple methods were used to determine the median, such as linear interpolation, but Fechner came up with a new, pathbreaking procedure. Since each C_j has a relative frequency p_j of “greater than” judgments up to that point, these observed relative frequencies will tend to be *monotonically increasing*, within the interval 0.0–1.0. Fechner proposed to fit a cumulative distribution function to the data, in particular the *normal ogive*. Given that choice, it is easy to find the median as the inflection point of the curve (the scale value β corresponding to a probability of 0.5 on the y -axis), where the curve’s positive acceleration changes into negative acceleration. Note that to use the inflection point of the normal ogive to find the scale value of a particular stimulus is essentially the same as the definition of the item difficulty parameter in *item response theory* (IRT), as used by Lord (1952).

Fechner used the original parameterization of the cumulative normal curve that Gauss had used in his first publication on least squares in 1809 (see Stigler, 1986, pp. 140–143). Gauss expressed the argument of the exponent as $-h^2\Delta^2$, with Δ the usual error term and h a *precision* parameter, which indicates the steepness of the normal ogive, or the *sensitivity* of the observer. Thus, the relation of h with

the standard deviation that is now commonly used is $h = 1/\sigma\sqrt{2}$. Fechner had a justification for the hypothesis of the normal ogive (which has become known as the *phi-gamma function*, and hence the *phi-gamma hypothesis*). It was suggested to him by his Leipzig University colleague August Ferdinand Möbius (1790–1868). Asked to judge whether one stimulus is “greater than” another one, the observer would form a mental estimate of each stimulus, making a normally distributed error, and reports the difference between the two mental estimates (Stigler, 1986, p. 247). In this way, Fechner could measure not only (possibly different) limens for each individual observer but also individual differences in their sensitivity or precision, associated with smaller standard deviations in their mental estimates. The normal ogive or phi-gamma function in the context of psychophysics has been called the *psychometric function* by Urban (1910), in analogy with the *biometric function*, which models a binary outcome (e.g., dying) as a function of some predictor (e.g., age).

Fechner’s method for fitting psychometric functions was simple *Gaussian least squares*, which in the early nineteenth century had become a standard analysis method for astronomers and geometers, but for psychologists it was an important innovation (Fechner, 1859). Nevertheless, Müller (1878, 1879) argued that proportions near 0.5 should be weighted more than proportions deviating from 0.5 in either direction, because the standard errors of proportions are a function of the mean. A further justification for using *weighted least squares* with weights $n_j/p_j(1-p_j)$ was provided by Urban (1908, 1910). Hence they were called *Müller-Urban weights*—and still mentioned as a term in the current APA Dictionary of Psychology. This weighted procedure is known as *probit analysis*.

There are two more classical experimental designs that aim at finding scale values for psychophysical or psychological stimuli to which Fechner contributed only partly. They were proposed with special interest in scaling *supraliminal* stimuli (relatively far apart), in which case a psychological *S*-scale of sensations cannot be formed by counting *jnds*.

4. *The method of equal appearing intervals (or method of equal sense distances).*

In its original form, the *method of equal sense distances* required the observer to *bisect* a given distance on a specific psychological continuum. For example, “given two sound intensities, R_1 and R_3 , the latter being of greater intensity than the former, O [the observer] had the problem of finding a stimulus R_2 such that the interval $R_1 - R_2$ equaled $R_2 - R_3$ ” (Guilford, 1936, p. 143). This task is the simplest one for obtaining equal sense distances, but it can already be used for testing Weber’s law in cases where the stimuli are supraliminal. The reasoning is that if we define the small but supraliminal increments $\Delta R_1 = R_2 - R_1$ and $\Delta R_2 = R_3 - R_2$, then it is easy to show that Weber’s law implies that R_2 must be the geometric mean of R_1 and R_3 . This prediction can be tested on observations obtained in a bisection experiment (Guilford, 1936, p. 144). For a classic application to tonal intervals, the reader is referred to Pratt (1928).

In the more general *method of equal appearing intervals*, the observer is asked to sort a relatively large set of n stimuli into a relatively small set of m piles, or *classes*, separated by *equal sense distances*. The stimuli sorted within a pile should have high psychological *similarity* (e.g., they should sound about equally intense). The observations can be collected in a frequency table of n rows (stimuli in the order of their scale value, if known) by m columns (classes labeled with consecutive integers), and the pattern that one would expect is that of a discrete bivariate normal distribution with negative correlation: i.e., high frequencies in the upper left corner of the table, extending along the diagonal to the lower right corner, tapering off toward the upper-right and the lower-left corners. We can now define the sensation scale S by allocating equal intervals between the classes, with the consecutive integers as scale values. That allows us to use the *method of right and wrong cases* to calculate S -values for the stimuli. For each row, we first transform the frequencies into cumulative frequencies and next fit the psychometric function to smooth them, with the equally spaced S -values on the x -axis. Then the median can be found as the inflection point of the curve and defines the scale value of the row stimulus. Fechner's law can be checked by plotting these against the physical R -values. This analysis method was suggested by Thurstone (1929), as a correction on the approach taken earlier in the notorious *Sanford weight experiment* on lifted weights (Sanford, 1898; also see Titchener, 1905, pp. 82–85; Murray, 2021, pp. 86–89), in which the average physical scale value in each pile was considered as the adjusted R -value and plotted against the equally spaced class intervals on the S -scale to check Fechner's law. Thurstone (1929) illustrated his corrected procedure with an example of 96 cards filled with irregularly spaced dots, where the stimulus magnitude was the number of dots on the card, and they were sorted in 10 piles. It showed convincingly that Fechner's law could be verified to hold for supraliminal stimuli.

5. *The method of choice and the method of paired comparisons*. Only the first steps in the development of these methods will be briefly described. Fechner was the first to study systematically the aesthetic properties of the so-called golden section in his treatise *Zur experimentellen Ästhetik* (Fechner, 1871). Among other methods, he proposed the *method of choice* (die *Wahlmethode*), in which an observer must choose one stimulus among k alternatives (see Guilford, 1936, pp. 222–225, and Green, 1996, for more detailed accounts of this development). For $k = 2$, the observer has to choose one stimulus out of a pair, and if this basic element is repeated for more pairs of m stimuli (not necessarily all of them), we arrive at the *method of paired comparisons*. An early application of this method was in the construction of a handwriting scale by Thorndike (1910), but a good method to analyze paired comparisons was still to be desired at that time.

With its methodological innovations, experimental designs and analysis methods, Fechner's work prepared the ground not only for experimental psychology but also for psychometrics, statistics (Sheynin, 2004), and even probability theory: the prominent applied mathematician Von Mises (1912) referred to Fechner's posthumously published work *Kollektivmasslehre* (1897) as one of the inspirations

that later brought him to introduce *randomness* as a basic concept in the theory of probability (Von Plato, 1994, pp. 182–183). Several improvements in the design of the constant method were introduced in a large-scale weight-lifting experiment by Peirce & Jastrow (1885). They wanted to measure *jnds* as precisely as possible, because of their skepticism about the existence of difference limens. The most important improvement was to determine the order of presentation of the pairs of weights by randomization, using two packs of playing cards (Stigler, 1978). They found that the sensitivity of the subjects was far below Fechner's threshold and concluded that there was no evidence for a difference limen (Stigler, 1992). In connection with this experiment, and similar ones in early experimental and educational psychology, Dehue (1997) has defended the claim that randomized designs were introduced by psychologists before Ronald Fisher introduced them in his classic handbook *The Design of Experiments* (1935). This claim was challenged by Hall (2007), who placed Fisher's rationale for the promotion of randomization in the tradition of agricultural field experiments starting in the middle of the nineteenth century. But it would lead us outside the scope of this chapter to pursue this priority issue further.

The impact of Fechner's psychophysics on Wundt and his doctoral students has been very large, as most of their experiments involved his methodology, except for Wundt's notion that psychologists should be the observer or subject (and not an arbitrary person). For only trained psychologists could use *introspection* to report their *apperception*, which is an unconscious process that interprets raw sense data in relation to past experiences. Prominent among Wundt's students were psychologists from the United States (in total 33 of them), including James McKeen Cattell (1860–1944) and Edward B. Titchener (1867–1927), who already made their appearance in this paper. Especially Cattell at Columbia University was a strong advocate for the statistical turn in American psychology in the period 1890–1915, part of a more general rise of statistical methodology in anthropology, sociology, and economics (cf. Camic & Xie, 1994). One of Cattell's students, Edward L. Thorndike (1874–1949), became the founder of modern educational psychology and educational testing at Teachers College of Columbia University and was one of the founding fathers of the Psychometric Society. From England, Charles Spearman (1863–1945) was also a Wundt PhD student, but he was more influenced by Francis Galton and the rise of mental testing in the last decade of the nineteenth century. Galton himself has praised Fechner in a letter from 1875, for having laid “the foundation of a new science [...] [in which a] mass of work by Arago, Herschel and various astronomers fall in as a part of the wide generalizations of Fechner, and much criticism and recognition of him will be found in Helmholtz” (Sheynin, 2004).

The formulation of the concept of the *psychometric function* was for sure Fechner's greatest contribution to psychometrics. It was recognized in standard textbooks, not only in the specialized ones already mentioned but also in more general popular texts about statistics. For instance, Truman Kelley's textbook *Statistical Method* (Kelley, 1924) has a section on psychophysical methods, in particular the method of right and wrong cases. But the psychometric function was also a topic for further clarification and research (e.g., Boring, 1917; Thomson,

1919; Urban, 1933). By far the most clarifications, generalizations, and extensions came from Louis Leon Thurstone (1887–1955) during the start of his career.

In our discussion of the method of equal appearing intervals, the psychometric function was used to find S -values of the stimuli as the median of a discrete distribution defined on equally spaced intervals on the sensation scale. It should be noted that nowhere in Thurstone's (1929) procedure he needed to use R -values (physical magnitudes) to derive the S -values. More generally, Thurstone (1927a) had already sketched a new framework for psychophysics, in which he introduced the concept of normally distributed *discriminal processes* with which an organism differentiates stimuli by calculating *discriminal differences*, which leads to the normal ogive psychometric function. It reminds us of the already mentioned suggestion made by Möbius to Fechner for justifying the phi-gamma hypothesis. In this suggestion, Möbius assumed that *mental estimates* of each stimulus were made, with a normally distributed error, and supposed that the differences between two mental estimates are reported by the observer. Indeed, Thurstone (1927a) fully developed the same idea and showed that Weber's law and Fechner's law are independent of each other and also that equally often noticed differences are not necessarily equal on the psychological continuum.

Within the same general framework, Thurstone (1927b) formulated his famous *Law of comparative judgment*. We already briefly met the experimental design to which this law applies in our discussion of the 5th Fechnerian method, i.e., the method of choice and paired comparisons. In a paired comparison design, m pairs of stimuli out of a set $\{S_1, \dots, S_n\}$ are formed, where m can be the total number of possible pairs $\frac{1}{2}n(n-1)$, or some subset of it. The subject is asked to compare the stimuli in a pair on any psychological attribute of interest and make a choice which one *dominates* the other. The dominance judgment may be *personal preference*, for example, when the stimuli are odors and the subject has to indicate which one smells more pleasant. Or the judgment may be an expression of *social or moral values*, for example, when the stimuli are crimes or offenses and the subject has to indicate which one is more serious. Upon replication of these judgments for one subject or across a number of different subjects, relative frequencies can be determined for each pair, and the Law of Comparative Judgment forms the basis for a statistical analysis that finds scale values for the stimuli on a psychological continuum, or *Thurstone scale*.

The assumption that each stimulus generates a normally distributed discriminial process in the mind of the subject(s) leads to a model in which the probability of making a choice of S_i over S_j is equal to the normal ogive of the difference in the means μ_i and μ_j , divided by the discriminial dispersions σ_{ij} , which are related to the standard deviations and the correlation between the two processes. The most often encountered case in which one assumes that $\sigma_{ij} = \sigma$ is called *Case V*. It is in fact equivalent to a simple probit model in terms of the differences in the means and without interaction terms. An authoritative treatment of Thurstonian scaling and some of its extensions is Bock and Jones (1968). An overview including modeling the discriminial dispersions σ_{ij} by multidimensional scaling was given by Heiser and De Leeuw (1981), while Takane (1989) and Maydeu-Olivares (2001)

gave formulations of Thurstonian models in terms of the analysis of covariance. Böckenholt (2004) proposed solutions for the arbitrary location of the origin in a Thurstone scale.

Out of the general model, a whole series of other psychological scaling methods emerged, such as the *method of successive intervals* (Saffir, 1937), *method of graded dichotomies* (Attneave, 1949), and the *law of categorical judgment* (Torgerson, 1958). In these methods, there are pairs of stimuli by judgment categories (or category boundaries), instead of pairs of stimuli by stimuli. If we consider pairs of ability test items by persons and collect dominance responses for them, we obtain an *Item Response Theory* (IRT) model with a normal ogive *item characteristic curve* (a psychometric function that gives the probability of a correct response for a person with a score somewhere on the *S*-scale of *ability*, where the inflection point of the curve is called the *item difficulty* parameter). A theoretical development of such an IRT model formed the basis of the new test theory by Lord (1952). In fact, Thurstone (1925) had already formulated the basic idea and had illustrated it with a set of data with Binet-type questions, collected by Cyril Burt on 3000 London school children. A more detailed treatment and comparison with current IRT methods, including the reasons for switching from the cumulative normal to the logistic function, can be found in Bock (1997). For an enthusiastic review of Thurstone's general scaling framework, see Lumsden (1980), who concludes his paper by saying: "During the 1920s Thurstone stole fire from the gods. (As a punishment they chained him to factor analysis.)" (Lumsden, 1980, p. 7). Indeed, Thurstone's work in the beginning of his career expanded the scope of psychophysical scaling enormously by allowing the inclusion of non-physical stimuli, which made it the early root of two main branches of psychometrics, the scaling branch and the IRT branch.

1.5 Conclusions and Discussion

We have seen that Christian Thomasius had already more or less done in 1692 with his rating system of personality what James McKeen Cattell and Francis Galton had in mind in 1890 with their outline of mental testing—except of course that Thomasius was not in the position to calculate the interrater correlations between the 12-point rating scales that he and his two students had collected on one particular individual. Anyway, as pointed out by Jones and Thissen (2007, p. 5), the proposal to use sensory reactions and motor skills as a way of assessing mental ability was invalidated by a study of a graduate student at Columbia, who found that, for each of Cattell's proposed tasks, the correlations with class grades were essentially zero (Wissler, 1901)—which effectively ended this approach to mental testing.

Unfortunately, we must also conclude that Thomasius, with his proposal to study different personality profiles of political leaders, was simply too far ahead of his time, because political psychology started to have an interest in this topic only somewhere in the 1970s (Simonton, 2014). Ironically, Galton and Cattell became

best known for their interest in quantitative studies of another subpopulation of the human race: men of science.

For Francis Galton, that project started already with the publication in 1869 of his notorious book *Hereditary Genius*, in which he tried to demonstrate the genetic transmission of intelligence, drawing on data culled from biographies and biographical dictionaries of scientists, eminent military leaders, philosophers, lawyers, and artists (Galton, 1869). The French botanist Alphonse de Candolle, who had read *Hereditary Genius*, responded acutely with the publication of *De Candolle* (1873), a book in which he offered an elaborate statistical study of the lives of outstanding scientists (members of the Academies of Science from Paris, Berlin, and London, including their foreign members). As noted by Ruth Schwartz Cowan: “He found that a very high proportion had come from countries or cities that possessed a moderate climate, a democratic government, a tolerant religious establishment, and an important trade centre. He concluded that Galton was wrong and that environmental factors did indeed play a crucial role in the production of outstanding men” (Cowan, 1970, p. ix).

Galton’s immediate response was to produce a similar study called *English Men of Science: Their Nature and Nurture* (Galton, 1874), in which he aimed to show—not surprisingly—the dominance of nature over nurture. This time his data were autobiographical replies to a long questionnaire, sent out to 180 eminent scientists (fellows of the Royal Society, and the like), of whom 100 were selected for statistical treatment. One conclusion was that a strong and innate taste for science is a prevailing characteristic among scientific men and another that they had fewer children than their parents (Godin, 2007, p. 696). It served Galton well in pursuing his political program of eugenics.

Cattell followed Galton with several projects of measuring eminent scientists. In 1895, he acquired the weekly journal *Science*, established in 1883, which had run into financial difficulties. He used it as a vehicle for reporting the results of his statistical studies on science, based on an extending directory of researchers, called *American Men of Science*. He started with 4131 entries in 1906 (Cattell, 1906a, b), which accumulated to 34,000 entries in 1944. The directory included their background characteristics, fields of study, estimates of scientific merit, measures of productivity, mobility, and so on. In addition, as documented in Webster (1985), he also developed a system of academic quality rankings on the level of institutions instead of individuals (Cattell, 1910).

Because of these projects in the measurement of science, Galton and Cattell are now regarded as *pioneers of scientometrics* (Godin, 2007), as is De Candolle (Szabó, 1985). Furthermore, Godin (2006) has described how and why systematic counting of publications, citations, and acknowledgements (the output side of science, a branch of scientometrics known as *bibliometrics*) originated with several other psychologists, following in Cattell’s footsteps.

What can be said about the lasting influence of Christian Wolff? He was certainly correct in thinking that the duration and clearness of thoughts could be empirically studied, as well as intensity, memory, attention, and individual differences. He, too, was ahead of his time, for these topics had to wait more than a hundred years

before they became incorporated in the empirical research programs of people like Wundt, Müller, Helmholtz, Ebbinghaus, and their students in the second half of the nineteenth century. In evaluating Wolff's impact, Vidal (2011, p. 111) remarks: "When the Aristotelian frameworks disintegrated, by the 1720s at the latest, psychology became the science of the human mind. In university circles, it was Christian Wolff who gave this shift its most systematic form. Hegel mentions in his *Lectures on the History of Philosophy* that Wolff gave the discipline a systematic structure which had served as a standard 'down to the present day', that is, until the 1820s."

Empirical psychology had to wait until Fechner laid out the psychophysical paradigm for measuring sensation. Wundt and contemporaries incorporated Fechner's pioneering work on experimental design and measurement of sensation and at the same time started to criticize him and to come up with alternatives (Murray, 2021, Ch. 5–8; Zudini, 2011). A much discussed criticism was that mental phenomena would *in principle* not be accessible for quantification, called the *quantity objection*. Michell (2006) phrased the denial of this criticism as the *psychometricians' fallacy*. For nuanced discussions of the quantity objection, see Hornstein (1988) and Sturm (2006). We have already noted the strong influence of psychophysics in the early twentieth century on experimental psychology, psychometrics, and educational psychology in the USA, including the upcoming testing movement in both Europe and the USA. However, after World War II its influence was waning, partly because of the upsurge of Stevens's "new psychophysics" in experimental psychology, based on magnitude estimation (Bolanowski & Gescheider, 1991), and partly because of the reorientation in item response theory to the logistic psychometric function.

But the mathematical psychologists have kept the fire burning! For example, Luce (1959) has critically examined what different forms a functional law like Fechner's law can have, dependent upon the scale levels of the independent and dependent variable. Here, it should be noted that Rozeboom (1962) has shown that Luce's conclusions were too strong, because they were based on a principle that is dubious at best. On the positive side, Dzhafarov and Colonius (2011) have argued that a lot of criticisms on Fechner's work are based on misinterpretation (partly due to Fechner's own expository and terminological shortcomings) and that Fechner's law can be derived without the notion of a *jnd*. In addition, they indicated that if we replace the term *difference sensation* with a more modern-sounding term *subjective dissimilarity*, then this change of perspective leads to the conclusion that Fechner's theory has the *additivity property* of a *unidimensional distance* (Dzhafarov & Colonius, 2011, p. 129). They also give examples of generalizations to multidimensional Riemannian geometry.

Finally, Zudini (2011) has shown convincingly that Fechner's system satisfies the conditions posed by the principles of *classical measurement* in Book V of Euclid's *Elements*, which poses a theory of *proportions between magnitudes* (Euclid, 1956). It appears that time is coming for someone to write a unifying book entitled *Elements of Psychometrics*.

References

- Attneave, F. (1949). A method of graded dichotomies for the scaling of judgments. *Psychological Review*, 56(6), 334–340. <https://doi.org/10.1037/h0063110>
- Barnard, F. M. (1971). The “Practical Philosophy” of Christian Thomasius. *Journal of the History of Ideas*, 32(2), 221–246. <https://doi.org/10.2307/2708278>
- Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, 24(3), 383–405. <https://doi.org/10.1076/jcen.24.3.383.981>
- Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, 16(4), 21–33. <https://doi.org/10.1111/j.1745-3992.1997.tb00605.x>
- Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. Holden-Day.
- Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods*, 9(4), 453–465. <https://doi.org/10.1037/1082-989X.9.4.453>
- Bolanowski, S. J., & Gescheider, G. A. (1991). *Ratio scaling of psychological magnitude: In honor of the memory of S.S. Stevens*. Lawrence Erlbaum.
- Boring, E. G. (1917). A chart of the psychometric function. *The American Journal of Psychology*, 28(4), 465–470. <https://doi.org/10.2307/1413891>
- Boring, E. G. (1961). The beginning and growth of measurement in psychology. *ISIS*, 52(2), 238–257. <https://doi.org/10.1086/349471>
- Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist*, 44(3), 576–578. <https://doi.org/10.1037/0003-066X.44.3.576.b>
- Brown, S. R. (1991). William Stephenson (1902–1989): Obituary. *American Psychologist*, 46(3), 244. <https://doi.org/10.1037/0003-066X.46.3.244>
- Brown, S. R. (1993). A primer on Q-methodology. *Operant Subjectivity*, 16(3/4), 91–138. <https://doi.org/10.15133/j.os.1993.002>
- Brown, W., & Thomson, G. H. (1940). *The essentials of mental measurement*. Cambridge University Press.
- Camic, C., & Xie, Y. (1994). The statistical turn in American Social Science: Columbia University, 1890 to 1915. *American Sociological Review*, 59(5), 773–805. <https://doi.org/10.2307/2096447>
- Cattell, J. M. K. (1886). Psychometrische Untersuchungen, I. Apparate und Methoden [Psychometric Studies, I. Equipment and Methods]. *Philosophische Studien*, 3, 305–335.
- Cattell, J. M. K. (1906a). A statistical study of American men of Science: The selection of a group of one thousand scientific men. *Science*, 24(621), 658–665. <https://doi.org/10.1126/science.24.621.658>
- Cattell, J. M. K. (1906b). A statistical study of American men of science: II. The measurement of scientific merit. *Science*, 24(622), 699–707. <https://www.jstor.org/stable/1634085>
- Cattell, J. M. K. (1910). A further statistical study of American men of science. *Science*, 32(827), 633–648. <https://www.jstor.org/stable/1635729>
- Cattell, J. M. K. (1921). In memory of Wilhelm Wundt, by his American students, section II. *Psychological Review*, 28(3), 155–159. <https://doi.org/10.1037/h0073437>
- Cattell, R. B. (1952). The three basic factor-analytic research designs—Their interrelations and derivatives. *Psychological Bulletin*, 49(5), 499–520. <https://doi.org/10.1037/h0054245>
- Cattell, J. M. K., & Galton, F. (1890). Mental tests and measurements. *Mind*, 15(59), 373–381. <https://www.jstor.org/stable/i339158>
- Cowan, R. S. (1970). *Introduction to the second edition of Galton (1874)*. F. Cass.
- De Candolle, A. (1873). *Histoire des Sciences et des Savants depuis Deux Siècles: Suivie D’Autres Études sur les Sujets Scientifiques en Particulier sur la Sélection dans L’Espèce Humaine* [History of science and scholars for two centuries: Followed by other studies on scientific subjects, in particular on selection in the Human species]. H. Georg, Libraire-Editeur.
- Dehue, T. (1997). Deception, efficiency, and random groups: Psychology and the gradual origination of the random group design. *ISIS*, 88(4), 653–673. <https://doi.org/10.1086/383850>

- Doyle, K. O. (1974). Theory and practice of ability testing in ancient Greece. *Journal of the History of the Behavioral Sciences*, 10(2), 202–212. [https://doi.org/10.1002/1520-6696\(197404\)10:2<202::AID-JHBS2300100208>3.0.CO;2-Q](https://doi.org/10.1002/1520-6696(197404)10:2<202::AID-JHBS2300100208>3.0.CO;2-Q)
- Drechsler, W. (1997). Christian Wolff (1679–1754): A biographical essay. *European Journal of Law and Economics*, 4(4), 111–128. <https://doi.org/10.1023/A:1008682025945>
- Drenth, P. J. D., & Sijtsma, K. (1990). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen* [Test theory. Introduction to the theory of the psychological test and its applications]. Bohn Stafleu Van Loghum.
- Dzhafarov, E. N., & Colonius, H. (2011). The Fechnerian idea. *American Journal of Psychology*, 124(2), 127–140. <https://doi.org/10.5406/amerjpsyc.124.2.0127>
- Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, 51(3), 599–635. <https://www.jstor.org/stable/2339898>
- Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, 53(4), 599–635. <https://www.jstor.org/stable/2979547>
- Euclid, (1956). *The Thirteen Books of Euclid's Elements, Vol. 2 (Books III-IX)* (L. Thomas Heath, Trans. & Ed.). Dover Publications.
- Fechner, G. T. (1831). *Maßbestimmungen über die Galvanische Kette* [Dimensional determinations via the galvanic chain]. Brockhaus.
- Fechner, G. T. (1859). Über ein wichtiges psychophysisches Grundgesetz und dessen Beziehung zur Schätzung der Sterngrößen [On an important psychophysical fundamental law and its relationship to the estimation of star sizes]. *Abhandlungen der Königl. Sächsische Gesellschaft der Wissenschaften*, 6(4), 455–532.
- Fechner, G. T. (1860). *Elements of psychophysics*, volume I (Helmut E. Adler, Trans.). Holt, Rinehart and Winston, 1966.
- Fechner, G. T. (1871). Zur experimentellen Ästhetik [On experimental aesthetics]. *Abhandlungen der Königl. Sächsische Gesellschaft der Wissenschaften, Mathematisch-physische Klasse*, 9(1), 555–635.
- Fechner, G. T. (1882). *Revision der Hauptpunkte der Psychophysik* [Revision of the Main Points of Psychophysics]. Breikopf und Härtel.
- Fechner, G. T. (1897). *Kollektivmasslehre* [Collective Mass Doctrine] (A. F. Lipps, Ed.). Engelmann.
- Feuerhahn, W. (2004). Die Wolff'sche Psychometrie [Wolffian Psychometrics]. In O.-P. Rudolph & J.-F. Goubet (Eds.), *Die Psychologie Christian Wolffs: Systematische und historische Untersuchungen* (pp. 227–236). Max Niemeyer Verlag. <https://doi.org/10.1515/9783110932317.227>
- Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Sage.
- Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. MacMillan.
- Galton, F. (1874). *English men of science: Their nature and nurture*. MacMillan.
- Galton, F. (1885). On the anthropometric laboratory at the late international health exhibition. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 14(1885), 205–221. <https://www.jstor.org/stable/2841978>
- Godin, B. (2006). On the origins of bibliometrics. *Scientometrics*, 68(1), 109–133. <https://doi.org/10.1007/s11192-006-0086-0>
- Godin, B. (2007). From eugenics to scientometrics: Galton, Cattell, and men of science. *Social Studies of Science*, 37(5), 691–728. <https://doi.org/10.1177/0306312706075338>
- Green, C. D. (1996). All that glitters: A review of psychological research on the aesthetics of the golden section. *Perception*, 24(8), 937–968. <https://doi.org/10.1068/p240937>
- Guilford, J. P. (1936). *Psychometric methods*. McGraw-Hill.
- Hall, N. S. (2007). R.A. Fisher and his advocacy of randomization. *Journal of the History of Biology*, 40(2), 295–325. <https://doi.org/10.1007/s10739-006-9119-z>
- Heiser, W. J. (2008). Psychometric roots of multidimensional data analysis in the Netherlands: From Gerard Heymans to John van de Geer. *Electronic Journal for History of Probability and Statistics*, 4(2), 1–25. <https://www.jehps.net/Decembre2008/Heiser.pdf>

- Heiser, W. J., & De Leeuw, J. (1981). Multidimensional mapping of preference data. *Mathématiques et Sciences Humaines*, 19(1981), 39–96. http://www.numdam.org/item/MSH_1981__73__39_0/
- Heymans, G., & Wiersma, E. (1906). Beiträge zur speziellen Psychologie auf Grund einer Massenuntersuchung [Contributions to differential psychology on the basis of a mass study]. *Zeitschrift für Psychologie*, 42(81–127), 258–301.
- Hock, H. S., & Schöner, G. (2010). Measuring perceptual hysteresis with the modified method of limits: Dynamics at the threshold. *Seeing and Perceiving*, 23(2), 173–195. <https://doi.org/10.1163/187847510X503597>
- Hornstein, G. A. (1988). Quantifying psychological phenomena: Debates, dilemmas, and implications. In J. G. Morawski (Ed.), *The rise of experimentation in American psychology* (pp. 1–34). Yale University Press.
- Hunter, I. (2000). Christian Thomasius and the desacralization of philosophy. *Journal of the History of Ideas*, 61(4), 595–616. <https://www.jstor.org/stable/3654071>
- Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 1–27). Elsevier. [https://doi.org/10.1016/S0169-7161\(06\)26001-2](https://doi.org/10.1016/S0169-7161(06)26001-2)
- Kelley, T. L. (1924). *Statistical method*. The Macmillan Company.
- Kouwer, B. J. (1963). *Het spel van de persoonlijkheid: Theorieën en systemen in de psychologie van de menselijke persoon* [The game of personality: Theories and systems in the psychology of the human person]. Erven J. Bijleveld.
- Leary, D. E. (1980). The historical foundation of Herbart's mathematization of psychology. *Journal of the History of the Behavioral Sciences*, 16(2), 150–163. [https://doi.org/10.1002/1520-6696\(198004\)16:2<150::AID-JHBS2300160206>3.0.CO;2-1](https://doi.org/10.1002/1520-6696(198004)16:2<150::AID-JHBS2300160206>3.0.CO;2-1)
- Link, S. W. (1994). Rediscovering the past: Gustav Fechner and signal detection theory. *Psychological Science*, 5(6), 335–340. <https://doi.org/10.1111/j.1467-9280.1994.tb00282.x>
- Lord, F. (1952). *A theory of test scores* (Psychometric Monograph, Number 7). Psychometric Corporation.
- Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, 66(2), 81–95. <https://doi.org/10.1037/h0043178>
- Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement*, 4(1), 1–7. <https://doi.org/10.1177/014662168000400101>
- Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, 66(2), 209–227. <https://doi.org/10.1007/BF02294836>
- McKeown, B., & Thomas, D. B. (2013). *Q-Methodology* (2nd ed.). Sage.
- McReynolds, P., & Ludwig, K. (1984). Christian Thomasius and the origin of psychological rating scales. *ISIS*, 75(3), 546–553. <https://doi.org/10.1086/353573>
- McReynolds, P., & Ludwig, K. (1987). On the history of rating scales. *Personality and Individual Differences*, 8(2), 281–283. [https://doi.org/10.1016/0191-8869\(87\)90188-7](https://doi.org/10.1016/0191-8869(87)90188-7)
- Mei, M. (2021). Wolff's idea of psychometria. In S. de Freitas Araujo, T. C. R. Rereira, & T. Sturm (Eds.), *The force of an idea, studies in history and philosophy of science* (Vol. 50, pp. 89–103). https://doi.org/10.1007/978-3-030-74435-9_6
- Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics*. Eleven International Publishing.
- Michell, J. (2006). Psychophysics, intensive magnitudes, and the psychometricians' fallacy. *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 17(3), 414–432. <https://doi.org/10.1016/j.shpsc.2006.06.011>
- Miller, G. A. (1964). *Mathematics and psychology*. Wiley.
- Mülberger, A. (2017). Mental association: Testing individual differences before Binet. *Journal of the History of the Behavioral Sciences*, 53(2), 176–198. <https://doi.org/10.1002/jhbs.21850>
- Müller, G. E. (1878). *Zur Grundlegung der Psychophysik, Kritische Beiträge* [On the foundation of psychophysics, critical contributions]. Grieben.

- Müller, G. E. (1879). Über die Maßbestimmungen des Ortsinnes der Haut mittels der Methode der richtigen und falschen Fälle [On measuring the spatial sense of the skin by means of the method of right and wrong cases]. *Archiv für die gesamte Physiologie des Menschen und der Tiere*, 19, 191–235. <https://doi.org/10.1007/BF01639850>
- Murray, D. J. (2021). *The creation of scientific psychology* (S.W. Link, Ed.) (1st ed.). Routledge. <https://doi.org/10.4324/9781315620985>
- Peirce, C. S., & Jastrow, J. (1885). On small differences in sensation. *Memoirs of the National Academy of Sciences for 1984*, Vol. III, 5th Memoir, 75–83.
- Pratt, C. C. (1928). Bisection of tonal intervals larger than the octave. *Journal of Experimental Psychology*, 11(2), 17–26. <https://doi.org/10.1037/h0075337>
- Ramul, K. (1960). The problem of measurement in the psychology of the eighteenth century. *American Psychologist*, 15(4), 256–265. <https://doi.org/10.1037/h0047753>
- Ramul, K. (1963). Some early measurements and ratings in psychology. *American Psychologist*, 18(10), 653–659. <https://doi.org/10.1037/h0040858>
- Rozeboom, W. W. (1962). The untenability of Luce's principle. *Psychological Review*, 69(6), 542–547. <https://doi.org/10.1037/h0041419>
- Saffir, M. (1937). A comparative study of scales constructed by three psychophysical methods. *Psychometrika*, 2(3), 179–198. <https://doi.org/10.1007/BF02288395>
- Sanford, E. C. (1898). *A course in experimental psychology: Part I, Sensation and perception*. Heath.
- Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, 136(3), 414. <https://doi.org/10.1037/0096-3445.136.3.414>
- Sheynin, O. (2004). Fechner as a statistician. *British Journal of Mathematical and Statistical Psychology*, 57(1), 53–72. <https://doi.org/10.1348/000711004849196>
- Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*, 22(6), 786–809. <https://doi.org/10.1177/0959354312454353>
- Sijtsma, K. (2016). Invited discussion of Cronbach (1951 Coefficient alpha and the internal structure of tests). *Psychometrika*, 81(4), 1205–1208. <https://doi.org/10.1007/s11336-016-9540-y>
- Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika*, 33(1), 75–102. <https://doi.org/10.2333/bhmk.33.75>
- Simonton, D. K. (2014). The personal characteristics of political leaders: Quantitative multiple-case assessments. In G. R. Goethals, S. T. Allison, R. M. Kramer, & D. M. Messick (Eds.), *Conceptions of Leadership* (Jepson Studies in Leadership) (pp. 53–69). Palgrave MacMillan. https://doi.org/10.1057/9781137472038_4
- Spearman, C. (1904). General Intelligence, objectively determined and measured. *American Journal of Psychology*, 15(2), 201–293. <https://www.jstor.org/stable/1412107>
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3), 271–295. <https://doi.org/10.1111/j.2044-8295.1910.tb00206.x>
- Stephenson, W. (1936). The foundations of psychometry: Four factor systems. *Psychometrika*, 1(3), 195–209. <https://doi.org/10.1007/BF02288366>
- Stephenson, W. (1953). *The Study of Behavior: Q-Technique and its Methodology*. University of Chicago Press.
- Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science*, 133(3446), 80–86. <https://www.jstor.org/stable/1706724>
- Stigler, S. M. (1978). Mathematical statistics in the early States. *The Annals of Statistics*, 6(2), 239–265. <https://www.jstor.org/stable/2958876>
- Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. The Belknap Press of Harvard University Press.
- Stigler, S. M. (1992). A historical view of statistical concepts in psychology and educational research. *American Journal of Education*, 101(1), 60–70. <https://doi.org/10.1086/444032>

- Stigler, S. M. (2010). Darwin, Galton and the statistical enlightenment. *Journal of the Royal Statistical Society, Series A*, 173(3), 469–482. <https://doi.org/10.1111/j.1467-985X.2010.00643.x>
- Sturm, T. (2006). Is there a problem with mathematical psychology in the eighteenth century? A fresh look at Kant's old argument. *Journal of the History of the Behavioral Sciences*, 42(4), 353–377. <https://doi.org/10.1002/jhbs.20191>
- Szabó, A. T. (1985). Alphonse de Candolle's early scientometrics (1883, 1885) with references to recent trends in the field (1978–1983). *Scientometrics*, 8(1–2), 13–33. <https://doi.org/10.1007/bf02025219>
- Takane, Y. (1989). Analysis of covariance structures and probabilistic binary choice data. In G. De Soete, H. Feger, & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 139–160). North-Holland. <http://takane.brinkster.net/Yoshio/p027.pdf>
- Thomasius, C. (1692a). *Die neue Erfindung einer wohlbegründeten und für das gemeine Wesen höchstmöthigen Wissenschaft das Verborgene des Herzens anderer Menschen auch wider ihren Willen aus der täglichen Conversation zu erkennen* [New Discovery of a Solid Science, Most Necessary for the Community for Discerning the Secrets of the Heart of Other Men from Daily Conversation, Even Against Their Will]. Christoph Salfeld.
- Thomasius, C. (1692b). *Weitere Erleuterung durch unterschiedene Exempel des ohnelängst gethanen Vorschlags wegen der neuen Wissenschaft anderer Menschen Gemüther erkennen zu lernen* [Further Elucidation by Different Examples of the Recent Proposal for a New Science for Discerning the Nature of Other Men's Minds]. Christoph Salfeld.
- Thomson, G. H. (1919). A direct deduction of the constant process used in the method of right and wrong cases. *Psychological Review*, 26(6), 454–464. <https://doi.org/10.1037/h0070741>
- Thorndike, E. L. (1910). Handwriting. Part I. The measurement of the quality of handwriting: Criticisms of the scale. *Teachers College Record*, 11(2), 8–46.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology*, 16(7), 433–451. <https://doi.org/10.1037/h0073357>
- Thurstone, L. L. (1927a). Psychophysical analysis. *The American Journal of Psychology*, 38(3), 368–389. <https://www.jstor.org/stable/1415006>
- Thurstone, L. L. (1927b). A law of comparative judgment. *Psychological Review*, 34(4), 273–286. <https://doi.org/10.1037/h0070288>
- Thurstone, L. L. (1929). Fechner's law and the method of equal appearing intervals. *Journal of Experimental Psychology*, 12(3), 221–238. <https://doi.org/10.1037/h0070968>
- Titchener, E. B. (1905). *Experimental psychology: A manual of laboratory practice, Vol. II, Instructor's Manual*. MacMillan.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley.
- Urban, F. M. (1907). On the method of just perceptible differences. *Psychological Review*, 14(4), 244–253. <https://doi.org/10.1037/h0073288>
- Urban, F. M. (1908). *The application of statistical methods to the problems of psychophysics*. Psychological Clinic Press.
- Urban, F. M. (1910). The method of constant stimuli and its generalizations. *Psychological Review*, 17(4), 229–259. <https://doi.org/10.1037/h0074515>
- Urban, F. M. (1933). The Weber-Fechner law and mental measurement. *Journal of Experimental Psychology*, 16(2), 221–238. <https://doi.org/10.1037/h0070805>
- Van der Heijden, P. G. M., & Sijtsma, K. (1996). Fifty years of measurements and scaling in the Dutch social sciences. *Statistica Neerlandica*, 50(1), 111–135. <https://doi.org/10.1111/j.1467-9574.1996.tb01483.x>
- Vidal, F. (2011). *The sciences of the soul: The early modern origins of psychology*. University of Chicago Press.
- Von Mises, R. (1912). Über die Grundbegriffe der Kollektivmasslehre [About the basic concepts of the collective mass doctrine]. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 21(1), 9–20. <http://eudml.org/doc/145318>
- Von Plato, J. (1994). *Creating modern probability: Its mathematics, physics and philosophy in historical perspective*. Cambridge University Press.

- Walker, H. M. (1929). *Studies in the history of statistical method, with special reference to certain educational problems*. The William & Wilkins Company.
- Webster, D. S. (1985). James McKeen Cattell and the invention of academic quality ratings, 1903–1910. *The Review of Higher Education*, 8(2), 107–121. <https://doi.org/10.1353/rhe.1985.0023>
- Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces—One element of social cognition. *Journal of Personality and Social Psychology*, 99(3), 530–548. <https://doi.org/10.1037/a0019972>
- Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, 4(1), 433. <https://doi.org/10.3389/fpsyg.2013.00433>
- Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Review: Monograph Supplements*, 3(6), i–62. <https://doi.org/10.1037/h0092995>
- Wolff, C. (1962a). Psychologia empirica, methodo scientifica pertractata [Empirical psychology, treated according to the scientific method]. In J. École (Ed.), *Christian Wolff: Vol. 5. Gesammelte Werke* (Psychologia empirica), Series II. Olms (Original work published 1732).
- Wolff, C. (1962b). Philosophia prima, sive Ontologia, methodo scientifica pertractata, qua omnis cognitionis humanae principia continentur [First philosophy, or Ontology, treated according to the scientific method, in which all the principles of human cognition are contained]. In J. École (Ed.), *Christian Wolff: Vol. 3. Gesammelte Werke* (Ontologia), Series II. Olms (Original work published 1736).
- Wolff, C. (1963). *Preliminary discourse on philosophy in general* (R. J. Blackwell, Trans.). Bobbs-Merrill (Original work published 1728).
- Wright, W. F. (2016). Early evolution of the thermometer and application to clinical medicine. *Journal of Thermal Biology*, 56(1), 18–30. <https://doi.org/10.1016/j.jtherbio.2015.12.003>
- Zhang, H. (1988). Psychological measurement in China. *International Journal of Psychology*, 23(1), 101–117. <https://doi.org/10.1080/00207598808247755>
- Zudini, V. (2011). The Euclidean model of measurement in Fechner's psychophysics. *Journal of the History of the Behavioral Sciences*, 47(1), 70–87. <https://doi.org/10.1002/jhbs.20472>

Chapter 2

The Janus Face of Psychometrics



Paul De Boeck and L. Robert Gore

Abstract Most psychometric data are behavioral data: responses to cognitive problems and to questionnaire items referring to behavior in a direct or indirect way. Therefore, measurement models are at the same time psychological models. The Janus face metaphor refers to these two sides of psychometrics. Measurement models can fail as psychological models. We discuss three examples, called vignettes in this chapter. The first refers to reflective measurement models not being in line with the psychology of what is measured. The second example concerns measurement invariance and the psychological meaningfulness of measurement invariance violations. The third example refers to the error variance (unexplained variance) in measurement models and models in general and how the error may be explained by individual-specific psychological phenomena.

Psychological measurement is the quantification of person variables of interest, such as cognition, skills, achievement levels, affect, and motivation, among many others. Psychological tests can quantify rather stable traits, variables subject to growth and change, and states depending on situations and occasions of measurement (Cronbach et al., 1972). Nearly all measurements quantify behavior of the person measured, including introspective self-reports (McFall & Townsend, 1998). Outside of physiology and group sociology, psychologists measure by observing participants and recording or rating their behaviors, using archival records or ratings of behavior, but in most cases, they ask participants to provide self-ratings or quantification of their own behaviors or experiences, and they present cognitive and other problems to work on in tests. Here we do not consider biological measures, such as

P. De Boeck (✉)

Department of Psychology, The Ohio State University, Columbus, OH, USA

e-mail: deboeck.2@osu.edu

L. R. Gore

Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, USA

e-mail: Bob.Gore@moffitt.org

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

L. A. van der Ark et al. (eds.), *Essays on Contemporary Psychometrics*,

Methodology of Educational Measurement and Assessment,

https://doi.org/10.1007/978-3-031-10370-4_2

cortisol, or neuroscience measures, such as fMRI, but focus instead on classically psychological variables such as attitudes, personality traits, moods, and intentions, as well as cognitive variables such as problem-solving, judgments, and response times.

It is noteworthy that often the measurement tool and the measurement object coincide. The tool consists of a person's behavior (e.g., test responses), and the object of measurement is a behavioral proclivity of the same person. This is not the case for the measurement of a person's weight or height. How a scale functions is independent of a person's weight, how a ruler functions is independent of a person's height, and how a thermometer functions is independent of the temperature to be measured. It is only in case that something is wrong with a measurement tool that the measurement tells us something about the tool. If all goes well, a thermometer tells us about the temperature of the room, of the body, etc., and a scale tells us about the weight of another object than itself. Indeed, psychology was born out of the difficulty human judges have in providing objective accounts of physical reality, such as the transit times of stars observed through telescopes (Traub, 1997), such that the observed times reflect something about the observer and not merely something about celestial mechanics and optics.

Because the tool and the objective of measurement are intrinsically inseparable, there always are two sides to psychological measurement, even when the researcher does not realize there are. The measurement model is at the same time a behavioral model, a model for how persons act while being measured. Classical test theory, factor models, and item response models (IRT) are at the same time measurement models and behavioral models. The researcher can focus on the first side and just consider the quantification of target behavioral proclivities or can focus on the implied behavioral model to understand people's test behavior, which in many cases is relevant as such, independent of the measurement outcome. Test items can require reports of knowledge, judgments, or decisions, so that the psychometric models are cognitive models, or test items can be self-descriptive and reflect a person's attitudes, feelings, and behavior, in which case the models are models of attitudes, feelings, and behaviors and models of how people describe themselves. A psychometric model is at the same time a measurement model (a model of the instrument) and a psychological model (a model of the person). The lack of separation and the two sides have inspired us to use the metaphor of a *Janus face*.

For example, in a cognitive test, information is collected to measure cognitive abilities, but the processes underlying the responses are of interest for substantive reasons, independent of the resulting measurement, and this has long been of interest to clinical psychologists (Lezak et al., 2012). The situation is different for the thermometer. We are not measuring the thermometer, and the temperature to be measured is not a feature of thermometer but of an object or space outside the thermometer. We do know how an analog thermometer works (the mechanism) to change heat into measured distance by causing expansion of the mercury in a linear tube, but we do not know with much certainty how a set of cognitive problems works on a person's mind to result in a response and

response time. We can learn from the cognitive test not only what the level of a person's cognitive ability is but, in principle, also to some extent how cognitive problem-solving works. Apart from the test responses themselves, measurement may require assumptions that are driven by domain knowledge, intuition, and potentially also self-reflection or introspection on the part of the researcher, given that the researcher is also an object of the same class (humans) as the object being measured. But some of these assumptions may be unjustified. For example, much response time research assumes that a set of response times is independently and identically distributed. However, this assumption was proven false decades ago (Luce, 1986).

Another aspect of two-sidedness of psychological measurement is that the tool can influence the measured object. Self-monitoring changes smoking behavior, for example (McFall, 1970). Taking a test can affect cognitive processes, such as when a person has a memorable insight that affects their future responses, perhaps years after taking the test (e.g., the Cognitive Reflection Test; Shane, 2005). Responding to questions on one's feelings can affect the feelings, a fact that has been exploited by political opinion pollsters (Gerstmann & Streb, 2004). Measuring a person's weight does not change the weight; measuring a person's temperature does not change their temperature appreciably.

A cognitive ability test yields a measure that is necessarily based on cognitive processes. The two sides to a cognitive test are the measurement and the underlying processes. A measurement model may not fit well with the resulting data, and that kind of failure is informative regarding the underlying focal processes. Ideally, the measurement model is at the same time a process model, which means that the two faces of psychometrics are consistent (McFall & Townsend, 1998). Outside of mathematical models in cognitive science (see, e.g., Ratcliff & McKoon, 2008), this condition is rarely met.

The Janus face of psychometrics implies that psychometric qualities are aspects of human behavior with relevance for psychology. We tend to isolate the measured quantity from its object (e.g., ignoring the response process) and to consider the measurement outcome as the objective output of an impartial instrument. A balance is a measurement instrument, but there is more to a psychological test than its role as a measurement instrument, of which the "psychometric" qualities are to be investigated and reported without psychology itself being at stake. Psychometrics is as much psychology as it is metrics.

In the following, we will discuss three possible cases, which we will call "vignettes," to illustrate the two faces. Vignette 1 concerns the internal consistency of test items, and vignette 2 concerns measurement invariance. Vignette 3 concerns error variance as unexplained variance and offers a clinical and idiographic perspective on that variance. They all three illustrate how a psychometric model is a psychological model and how psychometric qualities of an instrument can reflect important psychological principles.

2.1 Internal Consistency

Cronbach's alpha and alternative coefficients are popular quantifications of reliability. Cronbach's alpha is often interpreted as a measure of internal consistency. However, as Sijtsma (2009) explains, internal consistency is a vague notion. In this first vignette, we refer to internal consistency as an "average degree of 'interrelatedness'" between items (p. 114 in Sijtsma, 2009), which does contribute to coefficient alpha. Internal consistency arises from a reflective psychometric model with one dimension or multiple positively correlated dimensions, where each indicator reflects the latent variable in a direct way and with conditional independence. Not only are there psychologically meaningful other types of reflection than an independent direct reflection, but reflection is not the only way indicators can be linked to a latent variable. In *formative* latent variable modeling, for example, the link is from indicators to the latent variable and the link is cumulative. For both types of links (reflective, cumulative), at least three variants can exist: direct link, competitive link, and intermittent link.

2.1.1 Reflective Models

The three variants of reflective links are investigated by Tuerlinckx et al. (2002) for the Thematic Apperception Test (TAT). The TAT consists of a set of cards, and the respondent is requested to give a narrative interpretation of each card, a description which is believed to reflect underlying situationally specific psychological motivational tendencies.

The default type of reflection is *independent direct reflection* which means that the indicators do not affect one another (i.e., are independent conditional on the latent trait value) and that the reflection is not just intermittent (i.e., does not depend on the occasion). The common factor models and IRT models are in line with independent direct reflection. As a result, the common reliability coefficients for tau equivalent and congeneric models apply, such as the alpha coefficient and the omega coefficient, respectively.

A different type of reflection is *competitive reflection*, which means that reflection through one form of manifestation competes with reflection through other forms of manifestation. In psychometric models this would show through negative local dependencies and a reduction of internal consistency of indicators for the same trait. In the Tuerlinckx et al. (2002) study, the underlying principle is based on the Atkinson and Birch (1970) dynamics of action theory. The implication of the theory is that after an achievement motivation expression, the achievement action tendency is reduced, which shows as a negative effect of a response on the next response and thus as negative serial dependence. Competitive reflection is a more general phenomenon based on the dynamics of action theory. Any time there is restriction of resources related to the expression of a trait, competition follows. Time and

finances are examples of resources one needs when one follows interests related to leisure activities and social activities. One can have only so many interests, so many leisure activities, so many social activities, independent of the strength of one's needs, the breadth of one's interest, and the intensity of one's social motivation. Other principles at the basis of competition are habit formation and specialization. Habits may exclude other habits, as anxiety finds its expression in specialized fears, extreme political opinions fixate on certain topics and not on other topics associated with one's adversaries, and the set of fixations flowing from a particular ideological point is in flux. It is in theory possible that internal consistency of a test is very low and even zero or negative, although the indicators are all indicators of the same trait, albeit competing indicators.

Related to this but not discussed by Tuerlinckx et al. (2002), we might posit its opposite: accelerating reflection. This might occur when a behavior, once emitted, tends to raise the tendencies toward similar behavior. An example from social psychology is priming (Molden, 2014). A person who cooperates with an experimental confederate in one task may become more likely to cooperate on the next. On a measure of personality traits, as a person scans memory for examples of a particular trait (consider generosity for example), more such memories may come to mind, such that their proclivity to agree with similar trait descriptors increases as their progress through the test continues. On a multi-factorial test with shuffled items, this would suggest that the internal consistency of items grows from the first to the second half of the test.

The third type of reflection is *intermittent reflection*, which means that a trait is reflected only now and then but not on all occasions. Tuerlinckx et al. (2002) use the term "stochastic drop out" for this phenomenon. For example, a person can have a high need for achievement, but the need does not show at all possible achievement occasions. For the TAT, that would mean the need would not be reflected in the responses to all cards, which is a possibility suggested by Murray (1943, p. 15). When intermittent reflection is random, it is formally equivalent with an upper-asymptote model (as in the four-parameter model, but with a zero lower asymptote). What this means is that there always is a chance that the need for achievement is not expressed, which implies that the maximum probability (the upper asymptote) of an achievement response is smaller than 1.00. Dependent on the card, the upper asymptote is higher or lower (Tuerlinckx et al., 2002). When a response drops out of the normal response process, the response does not reflect the respondent's need for achievement, but instead some other need-induced phantasy is reflected in response to a TAT card. The assumption that intermittent reflection is random (conditional on the level of the upper asymptote) is an approximation for the fact that many different needs may take over to be expressed, conditionally independent of the achievement-related content of the card, induced by the varying strength of those other needs. Without the simplifying approximation with an upper asymptote, it would be too complex a model to be estimated, although it is possible to simulate the resulting intermittent reflection phenomena based on the dynamics of action theory (Atkinson & Birch, 1970). Intermittent reflection is the consequence of changing competitive strengths as postulated in the dynamics of action theory.

To give an example from another behavioral domain, suppose that a trait we are interested in is punctuality. Behaviors such as showing up on time for an appointment and making deadlines in time can be interpreted as reflections of punctuality. It is possible though (depending on the occasion) that another trait takes over to determine the behavior, which may lead to a violation of punctuality. For example, helpfulness may take over from punctuality if one needs more time than expected to solve someone else's problem, with consequences for the next appointment. Other traits taking over from a trait one wants to measure can explain intermittent reflection. Whereas competitive reflection refers to competition between indicators, intermittent reflection may refer to competition between traits for expression in a single behavior. Like competitive reflection, intermittent reflection also reduces the internal consistency. In physics, Brownian motion describes the process by which a dust particle is buffeted by random atomic collisions causing it to drift around in still air. Psychological tendencies may have a similar character, buffeting behavior in different directions depending partly on truly random factors. This cannot be explained by an error term when the whole response itself is captured by another tendency related to a trait one does not intend to measure, just as for an upper-asymptote model, the response cannot be captured by the common notion of an error term. An upper-asymptote IRT model is a mixture model for the pairs of persons and items, just as the three-parameter IRT model with a lower asymptote also is a mixture model.

Based on the empirical application in Tuerlinckx et al. (2002), the model with intermittent reflection was the best-fitting model for the TAT. The dropout probability in a constrained (but well-fitting) model with a common upper asymptote for all cards was 0.34, a close approximation of Murray's (1943) guess that 30% of responses are nondiagnostic responses.

2.1.2 *Cumulation Models*

As mentioned earlier, an alternative to reflection is *cumulation*, as in formative models. To explain the concept, let us use the example of happiness and assume that happiness has different sources (referring to different aspects of life). Let us further assume that the happiness from these different sources adds up (i.e., accumulates): relational happiness, happiness in one's job, and leisure time happiness. These sources do not need to be correlated, but they can as in the following examples. When people experience less happiness from one source, they may compensate by seeking and obtaining more happiness from another source, just as different sources of income add up and one source can compensate for another. Alternatively, when people reach a threshold of happiness, they may stop seeking happiness from untapped sources. In both these cases, the correlation would be negative. Independent of the relationships, if the sources of happiness add up, happiness is a cumulative trait, and internal consistency must not be expected. Control theory more generally describes a variety of phenomena where a person, motivated to

maintain homeostasis, experiences changes in appetite and behavior due to variation in goal satisfaction levels (Carver & Scheier, 1982). For accumulation, the same three types as for reflection can exist: independent direct accumulation, competition (and acceleration), and intermittence. The above examples of compensation and satisfaction are formally equivalent with competition as they lead to negative correlations. Possibly, the different sources of happiness do literally compete with one another or reinforce each other. For example, if happiness depends on the time invested in the sources of happiness (i.e., codetermines how much happiness is derived from the source), then investing in one's job may come at the cost of investing in relationships, which may lead to a negative correlation. It also is possible that happiness in one respect of life carries over to other respects of life. Intermittent cumulation would imply that the same source does not always contribute to one's happiness, depending on one's focus of the occasion. As a result, an inventory of pleasant activities a person has enjoyed (MacPhillamy & Lewinsohn, 1982), for example, does not necessarily lead to a high internal consistency.

The different kinds of reflection and accumulation illustrate how internal consistency is not just a measurement quality but a possible indicator of psychological processes. From a measurement point of view, a high internal consistency may seem desirable, while from a psychological point of view, a low internal consistency may be a meaningful result, even when it would lead to a low coefficient alpha value.

2.2 Measurement Invariance

2.2.1 *Relevance of Measurement Invariance and Its Violation*

Psychologists would like to quantify differences and changes to understand influences on human behavior. It is a well-known rule that measures cannot be compared if the condition of measurement invariance is not met (Millsap, 2011). A violation of the condition implies that using the same instrument results in measures of different variables, as if a scale does not always measure weight but sometimes quantifies volume or height instead. As a result, variations in the numeric output of instruments may be the quantification of dissimilar qualities, while the person doing the measuring believes they are comparing dissimilar individuals on the same quality.

Violations of measurement invariance are interpreted as an issue and may lead to adjustments of the measure. Rarely are the violations interpreted as interesting psychological phenomena, while a result that is undesirable from a *metric* perspective can be helpful from a *psychological* perspective. To illustrate, after a psychotherapeutic intervention, it can be expected that a trait is expressed in a different way and that the same behavior (the same response to an item) now has a different meaning. After a treatment for anxiety, perhaps not only the level of anxiety is reduced, but the threshold of some fearful behaviors has increased (a change

of the intercept, corresponding to a violation of scalar invariance), or previous behaviors driven by anxiety are now carried out for other reasons when they occur (a change of the loadings, corresponding to a violation of metric invariance). The lowering of a threshold for a fearful behavior, independent of an overall decrease of anxiety, is reflected in the intercept parameter of the behavior and is a violation of scalar invariance comparing pre- and post-intervention conditions. The lowering of a loading means that anxiety has less influence on the behavior and is a violation of metric invariance. These formal kinds of differences in thresholds and in the relationship with an underlying latent variable may exist between groups, across gender categories, between ethnic groups, between cohorts, between cultures, and between different points in time for the same set of persons, such as before an intervention and after an intervention. Fokkema et al. (2013) find evidence for such processes, called response shifts, with respect to depression.

The differences between groups and within groups across time that correspond to violations of measurement invariance will be called qualitative differences, and they can be of interest as psychological phenomena as such even though they interfere with the conditions of measurement invariance. One may have to give up on making inferences about quantitative differences such as differences between the means of a latent variable or between sum scores, but instead one may follow up on the specific violations and make inferences on qualitative differences instead.

For example, developmental psychologists hope to measure the growth in logical reasoning and vocabulary across childhood, and clinical psychologists hope to measure the reduction in anxiety, depression, or addictive cravings resulting from intervention. Social psychologists hope to measure geographical differences and cohort effects in implicit bias. To measure changes and differences requires that the measuring instrument preserve the conditional relationship between the behavioral proclivity as input and the output of the instrument, such as a sum score, across groups (such as geography) and time (in development or treatment outcome studies). This is considered a precondition for valid measurement, but its violations are interesting phenomena themselves, and violations may be a foreseeable result of the psychologist's theory of difference or change.

2.2.2 An Example

To give a clinical example, consider a group of spider-phobic undergraduates who have heretofore avoided spiders at all costs. Imagine that these individuals undergo a single-session arachnophobia treatment and that they make ratings of anxiety (0 = "not at all anxious" to 100 = "as anxious as you have ever been or could imagine being") before and after the session, in response to different items such as mentally imagining a spider, viewing a real spider, and touching a spider (which they may never have done), each time regarding a spider that sits still or crawls. In the sessions, the individuals approach and eventually touch spiders provided by the researcher. As a result of the treatment itself, anxiety ratings provoked by really

touching spiders may arise from different processes, perhaps with a clearly lower anxiety rating for touching spiders, while ratings of anxiety levels on seeing and thinking of spiders may have decreased less. It also is possible that the scale is recalibrated after the experience and not just by an additive constant for all items. In both cases there would be a violation of scalar measurement invariance, but the finding may be informative about the specific effects of the session, even though no inference can be made regarding an increased or decreased fear for spiders as measured by a latent variable or a sum score. The measurement invariance failure (viewed conventionally) invalidates the evidence base for the treatment.

In some cases, psychological theory predicts that a group effect or an intervention effect is different for a subset of items compared with other items. The Saltus model (Wilson, 1989) is a model for theory-based violations of measurement invariance in which a subset of items shares a common violation of scalar invariance, for example, a subset of Piagetian tasks becomes more difficult or easier with age.

2.2.3 Mathematical Models and Clinical Interpretations

Psychologists can formulate models simultaneously of the person being measured and the process of measuring the person. Such models could incorporate shifts in the judgments (such as in the arachnophobia example) and more complicated shifts in the meaning of measures. Although one way to jointly examine the state or trait being measured along with the response process would be with sophisticated, tailored mathematical models (McFall & Townsend, 1998), that would not be the only way. Psychologists who listen to the people who provide the measures, who give the measured the voice to speak about their experience of providing numbers or quantifiable behaviors, could gather a great deal of useful information in qualitative form (e.g., mixed methods research, Tashakkori & Teddie, 2003). Practicing clinical psychologists do this routinely and may find qualitative evidence that makes violations of measurement invariance interpretable. Clinical interpretations may also open the black box of error variance or unexplained variance and try to understand the particularity of individual human behavior and an individual person's life, as discussed in the following.

2.3 Error Variance and Unexplained Variance

Error variance is a common parameter in psychometric models, in classical test theory, in factor models, and in item response theories (if formulated in terms of latent responses). From a statistical and measurement point of view, error is specific yet unexplained variance: specific to the measurement indicator in question, unrelated to other measurement indicators (and unrelated to the latent variable or true score in CTT), and therefore unexplained.

2.3.1 *Two Views*

An interesting view on error variance is Kahneman's (2011) distinction between statistical thinking and causal thinking. Error variance is an example of what he calls statistical thinking, foregoing the meaningful effects of specific events and circumstances in a person's life. What Kahneman understands by causal thinking is thinking based on individual cases and individual events instead. Individual events and circumstance may affect a person's behavior and responses in a test, and such effects are globally summarized in error variance – (Kahneman et al. (2020) use the term “noise”) – while they may refer to psychologically meaningful phenomena that cannot be captured because the events and circumstances are person-specific and not part of the design (van Bork, 2019). Common sense causal thinking is often necessary in situations where causal inference is statistically underdetermined, and this kind of common sense has a place in psychological science.

When psychologists shift from academic to applied roles, in some cases the importance of peer-reviewed, published analyses of reliability (and validity) increases, while they may benefit from qualitative information on possible sources of the error variance that leads to a lower reliability. Whereas the statistical way of thinking is important, a more individualized approach in the line of “causal thinking” can be a useful complementary perspective.

2.3.2 *An Example*

Consider a parent whose fitness has been questioned in a contentious divorce proceeding. The parents in such a case may be court ordered to undergo an extensive psychological evaluation, which in some areas of the United States may include psychological testing (such as with the MMPI-2-RF), extensive reviews of background records including criminal background checks and children's medical records, interviews with people who know the parent and their children well, observation of the children's behavior with and without each parent present, and diagnostic interviewing of each parent (see vignette in Emery et al., 2005). Each of the resulting scores is an evaluation component, but these scores also contain error variance. However, it would be impractical and too ambitious to quantify fully the amount of this error in real-world, high-stakes cases, and adding individual information for a clinical judgment may be problematic. Clinical judgment research suggests that clinicians should be modest in their claims because complex constellations of additional individual information have been shown to be rife with judgment error (Garb, 2003; Dawes et al., 1989).

The interpretation of the joint collection of information components is subjective: different clinicians confronting the same collection of aggregated data could reach different conclusions (Garb, 1989), and the attorneys and judges in the case may select, block, amplify, and downplay different aspects of the record. The kind of

extensive evaluation performed in child custody cases may run to 100 pages, and each person who reads the record will no doubt face problems of how to consider the mass of information provided. Issues of selective attention and recall and individual bias will enter the process. In principle the content in the child custody record is individual information that may explain psychometric error. Unfortunately, relying on human judgment may not be a good way to interpret the information.

2.3.3 Two Issues

This third vignette highlights two issues. The complexity of error variance is not captured in an estimate of its size, whereas the multiple sources are psychologically meaningful and can be noticed in individual cases. This could help inoculate non-psychologists and psychologists alike against any tendencies to ascribe exclusive meaning to the global psychometric information while being blind to the qualitative information.

However, as a second point, there may be objective quantitative indicators of severe problems with reliability that ought to be highlighted for any users of the data. If, for example, an observed MMPI-2-RF score profile is to be used to comment on the future parenting abilities of the parties over the course of several years, and if the observed fluctuation in MMPI-2-RF scores across measurement occasions separated by a much shorter interval is such that the use of the test to forecast years into the future is in doubt, this fact is crucial (Faust, 2012). In a case such as this, the size of the unexplained variance is vitally important to the fair application of psychology in forensic settings, and speculations on a qualitative basis and causal thinking (in Kahneman's terms) may be largely misleading.

2.3.4 Clinical and Statistical

This vignette also highlights the potential value of training clinical psychologists to engage in nuanced analyses of their measurement procedures and the ways their findings are processed by end users. If applied psychologists were systematically trained to avoid focusing so narrowly on the justification of their measures with specific coefficients and instead were taught to think of their measures as intrinsically influenced by the contexts of measurement and the motivations and cognitions test takers have in relation to their performance, the quantitative indications of error variance would not be interpreted as the final word. What seems to be error – from a statistical point of view – may correspond to meaningful events in a person's life.

The challenge for decision-making may be to integrate across multiple sources of information. It is well-known from the judgment and decision-making literature that simple methods such as equal weighting of standardized scores (called improper linear models) could be useful, as illustrated, for example, by Dawes (1979). To

formulate an improper linear model, a set of judges rates a set of objects on a set of attributes. Ratings for each attribute are standardized across judges, and sums or means of standardized scores across attributes form the overall score. It is incumbent on psychology to educate end users of high stakes tests about the many sources of unexplained variance and imperfect validity. We also need to help test users find ways to reconcile a statistical approach for an optimization of prediction across a set of persons with an awareness that unforeseeable variation may invalidate predictions and decisions. Yet the complexity of this task is daunting.

Statistical reasoning may be the optimal way from a global perspective and across the set of individuals under consideration, but this does not guarantee it is the optimal way in individual cases where idiographic information is available about an individual's specific circumstances. From the perspective of causal reasoning in Kahneman's (2011) terms, which is more idiographic than statistical, the error and unexplained variance reflect meaningful information with consequences for how a test result should be interpreted. The problem with such an approach is that human judgment suffers from various shortcomings as amply described by Kahneman (2011) and Kahneman and Tversky (1996), which also explains why a statistical (actuarial) approach frequently works better than clinical judgment for predictive purposes (Dawes et al., 1989).

2.3.5 An Idiographic Alternative

A possible alternative for purely clinical judgment without giving up on individualized information is quantitative idiographic measurement, with different variables for each individual person, and within-person relationships of those variables across situations or stimuli. For example, people may be asked to rate their feelings toward important others in their life, while they each choose their own feeling terms as well as the important others. The data can then be analyzed in an objective way, for example, using cluster analysis or a dimensional analysis. Such approaches may be a way to counter the subjectivity and biases inherent to human judgment. Examples of such an approach can be found in Kelly's (1955) personal construct theory approaches based on the repertory grid and in Herman's self-confrontation method (Hermans, 1991; Lamiell, 1991; Lyddon et al., 2006). A method of Boolean factor analysis and cluster analysis for within-person data matrices that may help to understand the particularity of individual persons can be found in De Boeck and Rosenberg (1988) and Van Mechelen and De Boeck (1989). In addition, the application of idiographic data collection as for the repertory grid and for the self-confrontation method (the measurement tool) may have an effect, hopefully a beneficial effect, on the individuals being measured (the objects of measurement). Although more useful for understanding than for prediction, these methods may help to have a meaningful view on individualized factors that may contribute to unexplained variance in measurement models.

Most methods and the whole field of psychometrics are focused on an interindividual variance paradigm. The inherent complexity of psychological phenomena may require a somewhat different paradigm, with a stronger focus on intraindividual approaches. This may lead to a better understanding of what shows as measurement error in methods based on interindividual variance. While a qualitative way makes sense in the context of discovery, a more quantitative intraindividual approach can take care of the justification.

2.4 Discussion

It was not within the scope of this article to provide a compendium of statistical solutions, but rather to exemplify the psychology side of psychometrics. Our goal here was in the first place to shift perception. Janus faces were posted on gates, so that travelers coming and going saw different faces. As a result, Janus was thought to see both the past and the future. While entering a domain, one would see a particular face, and while leaving one would see a different face. At the start of an investigation, the researcher sees one of the Janus faces, and when the results are in, the researcher may see the other, one that could be disappointing from a measurement quality point of view but informative from a psychological point of view.

Psychology can continue its traditional attempt to separate the measurement tool from the human proclivities being measured, or it can turn around and regard Janus's other face: the face that might smile on us as we change course, perhaps even reverse course partly. Just as we have tended to regard test instruments as objective reflections of behavioral proclivities and to try to develop instruments that achieve this purpose, we have also tended to regard statistical procedures in the same way, and we have developed a reflex (and trained it into our students) according to which certain forms of reliability or measurement invariance have to be established before a measure can be considered worthy of use, and typically this demonstration relies on standard statistical methods such as confirmatory factor analysis, item response theory, or the computation of reliability coefficients such as alpha and test-retest reliability. But there is another path, which is to learn about psychology from so-called psychometric shortcomings and to use these indications and psychological theory to formulate models in line with measurement principles but also with psychological processes. Such models could incorporate shifts in the judgments (such as in the arachnophobia example) and more complicated shifts in the meaning of measures. Although one way to improve the meaningfulness of clinical methods would be with sophisticated, tailored mathematical models, that would not be the only way.

If perception shifted, and we began to regard the measurement as double-faced, we would be less apt to offer sweeping generalizations about human behavior that ultimately undermine our credibility when they are swept away by the facts in an individual person's life or in the next round of generalization in research. What

might result would be an approach to measurement that equally respects the two sides, that respects and provides a place for insights developed through looking into possible violations of measurement qualities and qualitative sources of error variance, to proceed more cautiously to conclusions. We suggest that this might also reduce some of the tendency psychologists have shown toward acrimonious debate and would provide legitimacy for researchers seeking to diversify the range of cultural contexts in which psychological research findings can be applied.

We do not want to replace quantitative approaches with qualitative ones. In the context of predicting specific outcome variables in an individual-differences paradigm, a statistical approach is clearly superior to a clinical approach, and adding qualitative information may not improve predictive accuracy, most likely because clinical judgment is vulnerable to distortions of various kinds. However, qualitative information may contain hints about prediction errors rooted in people's individual contexts. Hints are not proofs, but they help explain the omnipresence of errors and how such errors reflect the complex psychology of individual persons.

Reflecting on the Janus face of psychometrics may help us admit that our understanding of the world is only very partially captured by the current quantitative models we use and that deviations can refer to (1) meaningful but deviating models as discussed in the first vignette, (2) meaningful violations of measurement invariance as discussed in the second vignette, and (3) meaningful content of what is commonly called measurement error. The result might be a more investigative attitude, a stronger awareness of the two-sidedness of psychometric models, an openness to alternatives for the most prominent measurement models (CTT, confirmatory factor models, item response theory), and an awareness that replication and prediction failures do not necessarily stem from measurement shortcomings but are inherent to the meaningful complexity of the psychological reality (De Boeck & Jeon, 2018; De Boeck et al., 2019, 2021).

In his article on Cronbach's alpha, Sijtsma (2009) describes the unfortunate gap between psychology and psychometrics, which shows in misunderstandings and lack of interest from both sides. The gap has also led to the perception of psychometrics as an extraneous technical discipline with its own criteria and to the perception of psychometricians as gatekeepers and law enforcement agents. This view is not surprising, because psychometric models are usually not inspired by psychology (Borsboom, 2006). However, we believe that psychometric models are psychological models by implication, although primarily inspired by metric principles, and that psychometrics cannot be just a toolbox kind of discipline. The two faces of psychometrics cannot be separated.

References

- Atkinson, J. W., & Birch, D. (1970). *The dynamics of action*. Wiley.
- Boeck, P. D., & Rosenberg, S. (1988). Hierarchical classes: Model and data analysis. *Psychometrika*, 53(3), 361–381. <https://doi.org/10.1007/BF02294218>

- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425–440. <https://doi.org/10.1007/s11336-006-1447-6>
- Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical, and health psychology. *Psychological Bulletin*, 92(1), 111–135. <https://doi.org/10.1037/0033-2909.92.1.111>
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.
- Dawes, R. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582. <https://doi.org/10.1037/0003-066X.34.7.571>
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668–1674. <https://doi.org/10.1126/science.2648573>
- De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin*, 144(7), 757–777. <https://doi.org/10.1037/bul0000154>
- De Boeck, P., Jeon, M., & Gore, L. (2019). Beyond registration pre and post. *Computational Brain & Behavior*, 2(3), 183–186. <https://doi.org/10.1007/s42113-019-00063-w>
- De Boeck, P., DeKay, M. L., Gore, L. R., & Jeon, M. (2021). The trees and the forest: Investigating variability surrounding an aggregate result. *Theory and Psychology*, 31(3), 399–404. <https://doi.org/10.1177/09593543211016084>
- Emery, R. E., Otto, R. K., & O'Donohue, W. T. (2005). A critical assessment of child custody evaluations: Limited science and a flawed system. *Psychological Science in the Public Interest*, 6(1), 1–29. <https://doi.org/10.1111/j.1529-1006.2005.00020.x>
- Faust, D. (2012). *Coping with psychiatric and psychological testimony* (6th ed.). Oxford University Press.
- Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, 25(2), 520–531. <https://doi.org/10.1037/a0031669>
- Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, 105(3), 387–396. <https://doi.org/10.1037/0033-2909.105.3.387>
- Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment*, 15(4), 508–520. <https://doi.org/10.1037/1040-3590.15.4.508>
- Gerstmann, E., & Streb, M. J. (2004). Putting an end to push polling: Why it should be banned and why the first amendment lets congress ban it. *Election Law Journal: Rules, Politics, and Policy*, 3(1), 37–46. <https://doi.org/10.1089/153312904322739916>
- Hermans, H. J. M. (1991). The person as co-investigator in self-research: Valuation theory. *European Journal of Personality*, 5(3), 217–234. <https://doi.org/10.1002/per.2410050304>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Starus, and Giroux.
- Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103(3), 582–591. <https://doi.org/10.1037/0033-295X.103.3.582>
- Kahneman, D., Sibony, O., & Sunstein, C. R. (2020). *Noise: A flaw in human judgment*. Little.
- Kelly, G. (1955). *On the psychology of personal constructs*. Norton.
- Lamiell, J. T. (1991). Valuation theory, the self-confrontation method, and scientific personality psychology. *European Journal of Personality*, 5(3), 235–244. <https://doi.org/10.1002/per.2410050305>
- Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). Oxford University Press.
- Luce, R. D. (1986). *Response times: Their role in inferring elementary psychological organization*. Oxford University Press.
- Lyddon, W. J., Yowell, D. R., & Hermans, H. J. M. (2006). The self-confrontation method: Theory, research, and practical utility. *Counseling Psychology Quarterly*, 19(1), 27–43. <https://doi.org/10.1080/09515070600589719>
- MacPhillamy, D. J. & Lewinsohn, P. M. (1982). The pleasant events schedule: Studies on reliability, validity, and scale intercorrelation. *Journal of Consulting and Clinical Psychology*, 50(3), 363–380. <https://doi.org/10.1037/0022-006X.50.3.363>

- McFall, R. M. (1970). Effects of self-monitoring on normal smoking behavior. *Journal of Consulting and Clinical Psychology, 35*(2), 135–142. <https://doi.org/10.1037/h0030087>
- McFall, R. M., & Townsend, J. T. (1998). Foundations for psychological assessment: Implications for cognitive assessment in clinical science. *Psychological Assessment, 10*(4), 316–330. <https://doi.org/10.1037/1040-3590.10.4.316>
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Taylor and Francis. <https://doi.org/10.4324/9780203821961>
- Molden, D. C. (2014). Understanding priming effects in social psychology: An overview and integration. *Social Cognition, 32*(Supplement), 243–249. <https://doi.org/10.1521/soco.2014.32.supp.243>
- Murray, H. A. (1943). *Thematic apperception test manual*. Harvard University Press.
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922. <https://doi.org/10.1162/neco.2008.12-06-420>
- Shane, F. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Tashakkori, A., & Teddie, C. (2003). *Handbook of mixed methods in social & behavioral research*. Sage.
- Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8–14. <https://doi.org/10.1111/j.1745-3992.1997.tb00603.x>
- Tuerlinckx, F., De Boeck, P., & Lens, W. (2002). Measuring needs with the thematic apperception test: A psychometric study. *Journal of Personality and Social Psychology, 82*(3), 448–461. <https://doi.org/10.1037/0022-3514.82.3.448>
- Van Bork, R., (2019). *Interpreting psychometric models*. Unpublished doctoral dissertation. University of Amsterdam.
- Van Mechelen, I., & De Boeck, P. (1989). Implicit taxonomy in psychiatric diagnosis: A case study. *Journal of Social and Clinical Psychology, 8*(2), 276–287. <https://doi.org/10.1002/per.2410040207>
- Wilson, M. (1989). Saltus: A psychometric model or discontinuity in cognitive development. *Psychological Bulletin, 105*(2), 276–289. <https://doi.org/10.1037/0033-2909.105.2.276>

Chapter 3

Psychological and Educational Testing and Decision-Making: The Lack of Knowledge Dissemination in Textbooks and Test Guidelines



Rob R. Meijer, A. Susan M. Niessen, and Marvin Neumann

Abstract When it comes to decision-making based on psychological and educational assessments, there is compelling evidence that statistical judgment is superior to holistic judgment. Yet, implementing this finding in practice has proven to be difficult for both academic and professional psychologists. Knowledge transfer from research findings to practitioners and other stakeholders in psychological assessment is a necessary condition to close this gap. To obtain insight into how academic specialists in psychological testing disseminate knowledge about research findings in this area, we investigated how textbooks on testing and guidelines on test use report on, or do not to report on, decision-making in psychological and educational assessment. Second, we discuss some commonly encountered misunderstandings, and third we argue for a broader and more in-depth dissemination of research findings on this topic in textbooks and test standards; to this end we provide some suggestions.

3.1 Psychological and Educational Testing and Decision-Making: The Lack of Knowledge Dissemination in Textbooks and Test Guidelines

For decades, many Dutch psychology students' first acquaintance with psychometrics included studying the book by Drenth (1965, 1975) or the more recent editions by Drenth and Sijtsma (1990, 2006). Although, at some Dutch universities, this book

Authors Rob R. Meijer and A. Susan M. Niessen have equally contributed to this chapter.

R. R. Meijer (✉) · A. S. M. Niessen · M. Neumann
Department Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, The Netherlands
e-mail: r.r.meijer@rug.nl; a.s.m.niessen@rug.nl; m.neumann@rug.nl

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
L. A. van der Ark et al. (eds.), *Essays on Contemporary Psychometrics*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-031-10370-4_3

has since been replaced by more recent books, its influence on psychological testing in the Netherlands is significant. In our discussion with practitioners and academics, the book is still often mentioned as an authority textbook on test design and test use.

We still use the 2006 edition for our lectures to Dutch students, and one of the best features of the book is that it contains a chapter (Chap. 9) about “The contribution of a test in the decision-making process.” As we discuss and illustrate in this chapter, there are not many introductory textbooks on test theory or psychological and educational testing that devote much attention, let alone a whole chapter, to test use and decision-making. Most textbooks pay close attention to topics like reliability, validity, and types of tests, but test *use*, that is, the basic principles on how professionals should use tests when they make decisions, is often not discussed. Also, on conferences where psychometric research is presented, such as those of the National Council on Measurement in Education (NCME), the International Test Commission, or the Psychometric Society, presentations on test use and decision-making are almost nonexistent.

This is perhaps not that surprising because, as discussed in van der Linden (1991), although the practice of testing is firmly rooted in the field of decision-making (educational selection, selection for the military and companies), test theory or psychometrics has been mainly developed as a measurement theory. There are a few exceptions: the well-known work by Taylor and Russel (1939) and the book by Cronbach and Gleser (1965); this latter work provided a theoretical basis for test-based decision-making. Thus, in courses on psychometrics, students learn about measurement theories like the principles of classical test theory, item response theory, and factor analysis and in more advanced courses about the development of different psychometric models, parameter estimation procedures, fit statistics, and the application of these models to empirical data. But, in psychological testing or related courses, test *use* is not really instructed. While most textbooks on psychological testing discuss the decision-making perspective (e.g., Taylor-Russel tables) and some focus on utility models, there is a lack of focus on usage, that is, how to combine test scores with other information, as we discuss below.

This underrepresentation of knowledge and skill in test use in academic education is problematic. As future professionals, most of our students will mainly use psychological tests as a decision-making tool. In most applied settings, psychological tests are part of an assessment used to make judgments and predictions about behavior of individuals (Kuncel, 2008). For example, consider the following two scenarios.

A parole board consisting of different professionals, including two clinical psychologists, has to decide about temporary or permanent release of a prisoner before the expiry of the sentence, on the promise of good behavior. This decision has important consequences for the prisoner and for society, and many factors determine the prisoner’s future behavior. One of the standardized instruments that can be used to make this important decision is the Level of Service/Case Management Inventory (Andrews et al., 2004). This instrument assesses static and dynamic factors linked to recidivism risk based on 43 items, divided into 8 major categories. The total score provides information on the risk posed by the offender, and the

subcategories indicate individual characteristics that increase the risk of recidivism (i.e., criminogenic needs). The total score is used to determine the offender's initial risk level on a five-point ordinal scale ranging from very low risk to very high risk. Importantly, individual assessors can often override the initial risk level to create a final risk level when they see reasons to do so (see Guay & Parent, 2018 for more details). An important question is: Is it wise to override the initial risk level, for example, on the basis of professional expertise and experiences with a delinquent?

A hospital is searching for a consultant occupational physician. Requirements are "enthusiastic to continue the success of the team with innovative ideas, a careful decision maker, always putting the patients first, an excellent communicator, able to influence others positively and supportively, able to demonstrate leadership in a multi-professional environment" (these requirements were taken from an actual ad). A search team under the supervision of an I/O psychologist is advising the management which of 18 applicants is most suited for the job. They use an intelligence test, a situational judgment test, and an interview to decide which candidate is most suited for the job.

How should the information from the tests and the interview be combined to optimize the predictive validity of the decision? Should management review the scores on these three assessments and make a global judgment or should they compute a weighted average of the scores on these assessments and hire on the basis of this weighted average?

These two examples demonstrate test use by professional psychologists in (highly) consequential contexts. Other examples are deciding what diagnosis is the most suitable for a client, whether a client is eligible for a particular treatment, whether an athlete belongs to the 10% most capable athletes for a sports team, or whether a child needs extra training in particular subject matters in school.

Such decisions are rarely made using a single assessment tool. For example, in personnel selection, ability tests and interviews are used because these assessments are easy to administer and are expected to increase the criterion-related validity for later job performance, compared to only using one of these assessment tools (Schmidt & Hunter, 1998). Similarly, diagnoses and treatment recommendations in clinical psychology are often made based on a combination of tests, observations, biographical information, and clinical interviews. Therefore, it is not only important for professionals to know *what* information to use when making decisions (what are valid predictors and how can they best be measured) but also to know how to *combine* information from different sources to optimize prediction.

Many studies have been conducted to investigate how information can best be combined to optimize prediction. A major topic of investigation in this respect has been the distinction between holistic and statistical prediction. In holistic (or clinical, impressionistic, intuitive, informal) prediction, information is combined "in the head" of the decision-maker. Conversely, in statistical (or actuarial, mechanical) prediction, information is combined based on formal weighting procedures. In a classic review of 20 studies, Meehl (1954, inspired by Sarbin, 1943) showed that statistical prediction resulted in better predictions than holistic prediction. Many other studies confirmed these findings ever since (e.g., Grove et al., 2000).

Using statistical prediction is arguably one of the most effective ways to improve predictions and decisions in practice (Milkman et al., 2009). However, statistical prediction is not popular among professionals (e.g., Arkes, 2008; Highhouse, 2008; Meijer et al., 2020; Kuncel et al., 2013; Ryan & Sackett, 1987; Terpstra & Rozell, 1997; Vrieze & Grove, 2009).

There are several explanations for the underutilization of statistical prediction in practice, such as lack of perceived autonomy and fear of losing professional status (Highhouse, 2008; Nolan et al., 2020; Neumann et al., 2021b, 2021c). One important prerequisite, however, is knowledge. Without having knowledge about how to best combine information, psychologists will not use statistical decision-making (Neumann et al., 2021a). Therefore, in the present study, we first discuss a number of important characteristics of statistical prediction.

Second, we investigated how research findings on holistic and statistical prediction are disseminated. Textbooks are meant as summaries of academic research that synthesize findings and translate them into accessible information for students and professionals. Through studying how textbooks discuss holistic and statistical prediction, we learn about how research in this area is disseminated, which elements are unclear, and what misconceptions and controversies still exist. This knowledge is useful for two reasons (1): it may help improve the dissemination of research findings and (2) it provides input for research that is aimed at closing the science-practice gap (see Neumann et al., 2021b, for a research agenda).

Besides textbooks, test standards play an important role in disseminating information about evidence-based test use. Therefore, third, we describe if and how test standards disseminate knowledge on this topic. As we discuss below, test standards do not seem to be aimed at discussing or prescribing how test information can best be combined to optimize decision-making. We provide arguments for the importance of including research findings on information combination and decision-making to optimize test use in psychological practice. We want to emphasize that our aim is not to point fingers at authors of the textbooks and guidelines we reviewed, but to improve the dissemination of important research findings with respect to decision-making and prediction to strengthen psychology as an evidence-based, applied science.

3.1.1 Theory of Social Representation

To better understand how textbooks and test standards represent scientific theory of decision-making, we used the theory of social representation as discussed and used in Roulin and Bangerter (2012). They investigated the science-practice gap by studying how the use of structured interviews was diffused to practitioners in practitioner-oriented advice books. As they discussed “the theory of social representations (...) seeks to describe the social processes by which scientific knowledge is transformed into everyday knowledge used by laypersons” (p. 150). An interesting phenomenon is that laypersons often integrate new theories in

existing schemes or ideas. This is called anchoring. Second, this theory suggests focusing on the intermediary actors that translate scientific findings into social representations.

Authors of textbooks are the intermediary actors that delve into expert knowledge with the intention of diffusing it to students and professionals. They thus play a key role in the potential transformation of scientific findings, because (1) they may have different understandings of concepts than the experts they cite; and (2) they are designing their message to fit their audience's knowledge (Clark & Murphy, 1982). Compared to journalists and mass media, authors of textbooks are intermediary actors that stand much closer to the original research (Krathwohl, 1998, pp. 54–55) and are often specialists on the topic of their books.

3.2 Using Tests to Make Decisions

3.2.1 *Basic Distinctions: Data Collection and Data Combination*

For professionals that use assessment results for decision-making or prediction, which are almost all professionals in psychology and related disciplines, it is important to have knowledge about the way information can best be combined. Below we first provide descriptions of holistic and statistical prediction given by Meehl (1954, p. 3) and some later remarks given in Dawes et al. (1989) and Grove and Meehl (1996) because these articles are often cited in textbooks we discuss below. Meehl (1954, p. 3) discussed statistical prediction in the context of diagnosing persons for therapeutic sessions as follows:

“We may order the individual to a class or set of classes on the basis of objective facts concerning his life history, his scores on psychometric tests, behavior ratings or check lists, or subjective judgments gained from interviews”. The mechanical combination of information for classification purposes, and the resultant probability figure which is an empirically determined relative frequency, are the characteristics that define the actuarial or statistical type of prediction.

Three important elements of statistical prediction are (1) both “objective” and “subjective” (but quantified) impressions can be considered; (2) there is a mechanical combination rule; and (3) the rule is based on empirically established relations between the combined scores and observations and the behavior we want to predict. So, statistical prediction is not restricted to psychological test use; an assumption sometimes made in textbooks as we discuss below.

Holistic prediction is described as follows by Meehl (1954, pp. 3–4):

On the basis of interview impressions, other data from the history, and possibly also psychometric information of the same type as in the first sort of prediction, we formulate, as in a psychiatric staff conference, some psychological hypothesis regarding the structure and the dynamics of this particular individual. On the basis of this hypothesis and certain reasonable expectations as of the course of outer events, we arrive at a prediction of what

is going to happen. This type of procedure has been loosely called the clinical or case-study method of prediction.

Importantly, in holistic (clinical) decision-making, a prediction is made by “thinking about” the available information, not by using a pre-defined rule or on the basis of explicit empirically established relations. Relatedly, Dawes et al. (1989) described holistic and statistical predictions as

in the clinical method the decision-maker combines or processes information in his or her head. In the actuarial or statistical method the human judge is eliminated and conclusions rests solely on empirically established relations between data and the condition or event of interest.

Furthermore, Dawes et al. (1989) noted that

Virtually any type of data is amenable to actuarial interpretation. For example, interview observations can be coded quantitatively (patient appears withdrawn: [1] yes, [2] no). It is thereby possible to incorporate qualitative observations and quantitative data into the predictive mix. Actuarial output statements, or conclusions, can address virtually any type of diagnosis, description, or prediction of human interest.

Thus, in short, statistical prediction is about the way information is *combined*, not about *what* information is used to make decisions.

3.2.2 Statistical Prediction Is Superior to Holistic Prediction

As mentioned above, many empirical studies and meta-analyses convincingly showed that following structured decision rules results in better prediction than combining information “in the head” (Meehl, 1954; Kuncel et al., 2013; Grove et al., 2000; Karelaia & Hogarth, 2008; Ægisdóttir et al., 2006; Morris et al., 2015). More specifically: Dawes et al. (1989) cited almost 100 comparative studies and found that the statistical method performed better than the holistic method. Grove et al. (2000) analyzed 136 studies from medicine, education, and clinical psychology, where professionals predicted outcomes such as academic performance, job success, medical or psychiatric treatment success, criminal recidivism, and suicide. They concluded that “Even though outliers can be found, no systematic exceptions to the general superiority (or at least material equivalence) of mechanical prediction were identified.” Grove and Meehl (1996, p. 26) discussed that, from a theoretical perspective, this conclusion should be expected:

From a theoretical viewpoint the issue may be rather uninteresting, because it is trivial. Given an encodable set of data – including such first-order inferences as skilled clinicians’ ratings on single traits from a diagnostic interview – there exists an optimal formal procedure (actuarial table, regression equation, linear, nonlinear, configural, etc.) for inferring any prespecified predictand. This formula, fallible but best (for a specific clinical population), is known to Omniscient Jones but not to the statistician or clinician. However, the statistician is sure to approximate it better, if this is done properly. If the empirical comparisons had consistently favored informal judgment, we would have considerable explaining to do.

The argument that statisticians should do (and do) a better job at approximating the optimal way to combine information for prediction, and the sections in definitions of statistical prediction by Meehl (1954) and Dawes et al. (1989) that emphasize using statistical rules based on *empirically established* relations between information and the behavior we want to predict, reveal the most significant practical challenge for the application of statistical prediction in practice. They require the availability of data to design empirically based statistical prediction rules.

3.2.3 Robustness of Simple Rules

So, ideally, large datasets based on representative samples of the target population are collected to estimate optimal weights for each variable (e.g., in regression analysis), and the results are cross-validated. Clearly, this is often not possible in practice because such datasets are not available. Effective methods to tackle this steep hurdle are described by Dawes (1979). He discussed that, instead of using optimal weights derived from large, primary data, using the same weight for all variables (i.e., unit weighting) or even using randomly chosen but consistent weights in mechanical procedures still often results in better predictions than using holistic prediction.

However, under particular conditions, unit weighting can result in less valid predictions compared to using the single best predictor alone (Murphy, 2019; Sackett et al., 2017). A simple rule was discussed in Murphy (2019): avoid using predictors (i.e., give them a zero weight instead of a unit weight) that correlate more strongly with the other predictors than with the criterion. Moreover, this advice holds when decisions are made holistically as well, since adding such information could “dilute” the most predictive information (Dana et al., 2013).

3.2.4 People Are Bad at Identifying Exceptions to the Rule

When statistical rules are used in practice, they typically serve as decision aids that can be overruled when professionals believe that is appropriate (e.g., Guay & Parent, 2018). Importantly, research shows that overriding a statistical prediction because a certain specific case is believed to be an exception to the rule is a bad idea: people are not very good at correctly identifying these exceptions (Guay & Parent, 2018; Dietvorst et al., 2018; Dawes, 1979). This conclusion can be logically derived from the findings that statistical prediction outperforms holistic prediction; if people were good at identifying exceptions, holistic procedures would outperform mechanical procedures (see Dana et al., 2013 for a similar remark).

A question that arises from the above is whether psychologists can learn to match the predictive accuracy of statistical rules through experience. Kahneman and Klein (2009) discussed this question in depth and concluded that professionals

in psychology have a hard time to match the accuracy of their holistic predictions to the accuracy of decision rules, because (1) the environment in which psychologists act is difficult to predict and (2) feedback is absent or incomplete and delayed at best, which both seriously hinder learning. The biggest problem, however, is that these findings are in conflict with the *perceptions* of making accurate predictions that many professionals have when making decisions. As Kahneman discussed “If people can construct a simple and coherent story, they will feel confident regardless of how well grounded it is in reality” (Kahneman & Klein, 2010, p. 4).

3.2.5 Transparency

Another important characteristic of statistical prediction as defined above that we would like to mention is their transparency. By combining information in a pre-defined, transparent rule, we can replicate decisions, evaluate our policies, and adapt decision rules accordingly, *because we know exactly what we did*. In contrast, that is not the case when decisions are made holistically, because it cannot be directly observed how an assessor combines information “in the head.” This makes it harder to evaluate and improve our decisions.

3.3 What Textbooks Communicate About Test Use and Data Combination

We investigated the following research questions:

1. Do textbooks on psychological testing discuss statistical/holistic decision-making?
2. Which references to sources do they use as the basis of their treatment of this topic?
3. Are their conclusions in line with the literature on this topic? In particular, we investigated five criteria: (3a) Is the overall conclusion in line with the empirical literature: statistical prediction should be preferred over holistic prediction? (3b) Do textbooks make a distinction between data collection methods (e.g., tests, interviews, observation) and data combination methods (according to a rule or in the head)? (3c) Is there a discussion about the robustness of using non-optimal weights? (3d) Do textbooks mention exceptions to the rule, and do they correctly discuss how to handle them? (3e) Is there a discussion about transparency of decision making? Although we consider transparency a very important aspect of decision-making, it is not often discussed in the statistical/holistic literature and therefore we did not take this aspect into account when evaluating criterion 3a.

3.4 Method

3.4.1 Sample

We conducted a broad search of textbooks on psychological testing. We started with an electronic search using the library search engine *SmartCat* with the search term “books on psychological testing” with restriction that books should be written in English and published after 1995. This date was a bit arbitrary; we were interested in how statistical versus holistic prediction using tests is discussed in the more recent literature. This resulted in 3031 hits. The first author of this study then selected books using the following inclusion criterion: the books should be broad introductory books on psychological testing. Books on specific topics, such as books exclusively on intelligence testing or test use in minority groups were excluded. This strongly reduced the number of hits. The third author independently selected books using the same search engine and based on the same criteria discussed above as the first author, and he found one book that was not identified by the first author, which was added to the list. This resulted in a selection of 13 textbooks (Table 3.1).

3.4.2 Coding

In each book, we analyzed the content of the text to evaluate if and how statistical and holistic prediction were presented. Because textbooks contain a large amount of information (often several hundred pages), we first looked at the index and the references to identify potentially useful sections. Index terms we used were *clinical*, *holistic*, *actuarial*, *mechanical*, *statistical prediction*, and *decision making*. Authors we looked for in the references were *Meehl* and *Dawes*. When these references did not provide any results, we also checked *Highhouse* and *Kuncel* and *Grove*. However, this did not provide additional information as all textbooks referring to *Highhouse*, *Kuncel*, or *Grove* also referred to *Meehl* or *Dawes*.

Two independent raters (first and third author) searched the books and coded the texts on the basis of the five research questions mentioned above under 3(a)-3(e). The two raters checked the text passages on the basis of the five criteria discussed above. Each criterion was rated on a four-point scale: (0) no description at all; (1) description is wrong; (2) there is some description, but lacks important points; and (3) fair, accurate description.¹ The two raters first coded the textbooks independently and then discussed any score differences between them until consensus was reached.

¹ We agree with an anonymous reviewer of this chapter that, although technically the ratings are nominal, the coding scheme we used may suggest that they are ordinal. An ordinal interpretation would imply that a wrong description is “better” than no description, which is not the case. As this reviewer correctly remarked “One may argue that the reverse is true, which is reflected in the opening lines of the great must-see movie *The Big Short*: ‘It ain’t what you don’t know that gets you into trouble. It’s what you know for sure that just ain’t so.’ – also see <https://quoteinvestigator.com/2018/11/18/know-trouble/>”

Table 3.1 Scores that reflect the way textbooks discuss different criteria

	Conclusion in line with literature	Data collection/combination	Robustness weights	Exceptions to the rule	Transparency
Anastasi and Urbina (1997)	1	1	0	1	0
Aiken (2003)	2	2	0	0	0
Murphy and Davidshofer (2005)	2	2	3	2	0
Kline (2005)	0	0	0	0	0
Domino and Domino (2006)	2	2	0	0	0
Reynolds and Livingston (2012)	2	0	0	2	0
Kaplan and Saccuzzo (2013)	2	0	0	0	0
Gregory (2013)	2	0	0	0	0
Hogan (2015)	1	0	1	0	0
Miller et al. (2015)	2	1	0	0	0
Cohen and Swerdijk (2015)	1	2	0	0	0
Furr (2018)	0	0	0	0	0
Cooper (2019)	0	0	0	0	0

Note: 0 = not discussed, 1 = incorrect description; 2 = description lacks important points; 3 = fair description

3.4.3 Results

In Table 3.1 we provide an overview of the textbook literature. Note that Kline (2005), Furr (2018), and Cooper (2019) did not discuss mechanical versus holistic prediction. Below we summarize the most important findings.

1. Most textbooks on psychological testing discuss statistical versus holistic prediction using a limited number of pages (between 1 and 9 pages, mostly 1–3 pages). There was no textbook that wholeheartedly endorsed the main conclusion from the empirical literature that statistical prediction should be preferred over holistic prediction. Some textbooks only mentioned the empirical results found, without drawing any conclusions or mentioning implications. Almost all textbooks suggested a middle-of-the-road compromise, where they indicate that a rule can be used in some cases, but that there are situations in which that is not possible or desirable. Most reasoning is of the form: *Meehl (1954) or some other meta-analysis found that statistical prediction is superior to clinical prediction. We generally agree with this conclusion, but there are conditions where clinical prediction is preferred (because there are exceptions, because you cannot use tests in all cases, because it is difficult to formulate a rule).* For example, Murphy and Davidshofer (2005) provided an elaborate summary of the research on statistical versus holistic decision-making, but they also conclude:

However, in the long run, the automation of clinical prediction would limit the accuracy of clinical predictions, since it would preclude the use of behavioral observation data or the selection of appropriate tests to optimally assess the status of the individual patient. (p. 529)

There is, however, no reason why *quantified* behavioral observations could not be incorporated in statistical predictions. Furthermore, the “selection of appropriate tests to optimally assess the status of the individual patient” is still possible under mechanical decision-making.

In many passages, there was no explicit distinction between “the nature of information” and “how to combine information.” Textbooks rarely explicitly described this distinction. Many passages provide examples of holistic versus statistical prediction which incorrectly suggest that statistical decision-making is tied to using tests and holistic decision-making is tied to using other information (sometimes in addition to tests). For example, Miller et al. (2015, p. 419) discussed that: “For more than 50 years, researchers have debated the accuracy of making diagnoses using the unstructured interview (called the clinical method) compared with using structured psychological tests (called the statistical method). In 1954, Meehl published the results of his examination of 20 studies that compared clinical and statistical predictions (Meehl, 1954). His conclusion was that statistical methods were as accurate as, and often more accurate than clinical methods.”

2. Only optimal regression models are described as superior to holistic decision-making. The advantages of suboptimal rules such as unit weighting or expert weighting are not discussed. If authors mention specific examples, they often

come from the clinical context. An interesting example on the use of the MMPI is provided by Gregory (2013, pp. 487–493). Gregory (2013) discussed that “computerized narrative test reports should use existing actuarial formulas to determine the likelihood of various psychiatric diagnosis” (p. 491). However, Gregory (2013) also discussed that a drawback of statistical prediction is that when the rules are applied to a new client population, new rules should be determined because they will perform less well in a new population. Ideally, this would indeed be the case, at least when sufficiently large samples would be available. However, this remark ignores the empirical results that suboptimal weights generally do a better job than holistic combinations (Murphy et al., 2013; Yu & Kuncel, 2020).

Also, Hogan (2015, p. 177) noted that:

Can we replace clinicians with formulas? Sometimes yes, sometimes no. Development of formulas requires an adequate database. When we have an adequate database, we should rely on it. But we do not always have an adequate database. In that case, we must rely on clinical judgment to make the best of the situation.

This is an often-encountered misunderstanding that despite articles like those by Grove and Meehl (1996) and Dawes and Corrigan (1974) seems to be ineradicable. As we discussed above, research showed that picking a number of valid predictors and choosing reasonable weights based on empirical research (e.g., meta-analysis) will often result in more accurate decisions than holistic judgment. If textbooks keep communicating that adequate databases are a necessary condition to be able to use statistical prediction, it is no wonder that practitioners almost exclusively use holistic judgment, because adequate data are rarely available.

3. Some textbooks state that, sometimes, holistic methods should be preferred. These are perhaps the most interesting passages because most of the time, no references are provided to support those statements; they seem to rely on “common sense” or “authority” arguments. Most importantly, there is no evidence that holistic methods should be preferred over mechanical procedures in *any* situation.

Some authors seem to imply that we do not know which decision-making method is superior. For example, Kaplan and Saccuzzo (p. 554) noted “Further, the question remains as to whether computer interpretations can ever be as good as, let alone better than, those of the clinician.” Sometimes references are used, but then the content of these references is refuted by more recent articles, or the original articles are misinterpreted. For example, Aiken (p. 337) discussed that “under certain circumstances trained practitioners employing data from a variety of sources (case history, interview, test battery, and the like) are better than actuarial formulas (Goldberg, 1970; Holt, 1970; Wiggins & Kohen, 1971).” This is incorrect, because Goldberg (1970) showed the opposite, namely, that statistical rules created based on decisions made by the assessors were better than assessors themselves. Additionally, Holt (1970) is sometimes used as a reference in favor of holistic prediction, but Holt (1986, p. 378) himself conceded that statistical judgment is superior when he wrote:

Maybe there are still lots of clinicians who believe that they can predict anything better than a suitably programmed computer; if so, I agree that it is not only foolish but at times unethical of them to do so . . . If I ever accused him [Paul Meehl] or Ted Sarbin of “fomenting the controversy”, I am glad to withdraw any implication that either deliberately stirred up trouble, which I surely did not intend.

3.4.4 Conclusion on Decision-Making as Discussed in Textbooks

The way textbooks on testing discuss decision-making based on a combination of information is mostly not in agreement with the empirical literature. It seems as if authors of textbooks anchor mechanical decision-making to pre-existing schemes, as the theory of social representation would predict. These pre-existing schemes consist of ideas of how we make decisions in daily life: holistically. For example, Anastasi (p. 520):

A major contribution of the clinical method for example is that data are obtained in areas where satisfactory tests are unavailable through interviewing and observations of behavior. The clinical method is also better suited than the statistical method to the processing of rare and idiosyncratic events whose frequency is too low to permit development of statistical strategies.

This remark seems to be based on “common sense,” but not on results from the empirical literature which showed the opposite, namely, that people have a hard time in identifying valid idiosyncrasies. As a result, we speculate that many textbook authors (unintendedly) mix empirical findings in the literature with their own experiences. Furthermore, because the topic is more complex than many textbook authors perhaps realize, not enough space is devoted to carefully and accurately explaining the literature.

3.5 What Test Standards Communicate on Decision-Making with Tests

We investigated the following research questions:

1. Do test standards on psychological testing discuss statistical/holistic prediction?
2. Are their conclusions in line with the literature on this topic?²

There are different guidelines on test use. Internationally, the most important ones are the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014; in the remaining of this article referred to as the

² In contrast to the textbook research questions, we did not research which references were used because test standards include very few references.

Standards) and the International Test Commission Guidelines on Test Use (2013; in short, the ITC guidelines). The latter is available in many languages. Both guidelines fulfill an important role to transfer scientific assessment research to professional practice and contain important and very useful information.

3.5.1 *Standards for Psychological and Educational Testing*

To answer the first research question, it is important to first look at the mission of the Standards. On p. 1 it says

The purpose of the standards is to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test use. Although such evaluations should depend heavily on professional judgment, the standards provide a frame of reference to ensure that relevant issues are addressed.

Furthermore, on p. 2 it is noted that

Although the principles and concepts underlying the standards can be fruitfully applied to day-to-day decisions – such as when a business owner interviews a job applicant, a manager evaluates the performance of subordinates, a teacher develops a classroom assessment to monitor student progress to an educational goal, or a coach evaluates a prospective athlete – it would be overreaching to expect that the standards of the educational and psychological testing field would be followed by those making such decisions. In contrast, a structured interviewing system developed by a psychologist and accompanied by claims that the system has been found to be predictive of job performance in a variety of settings falls within the purview of the standards. Adhering to the Standards becomes more critical as the stakes for the test taker and the need to protect the public increases.

From these quotes it is clear that decisions made by persons not being a psychologist are considered beyond the scope of the Standards. It may also be inferred that the Standards are particularly concerned with the quality of individual assessment tools. However, decisions are seldom made based on one individual test or instrument. The Standards (p. 198) indeed discuss “In educational settings, a decision or characterization that will have major influences on a student should take into consideration not just scores from a single test, but other relevant information.” How this may be done is discussed on p. 170.

In some instances, test information is used in a mechanical, automated fashion. This is the case when scores on a test battery are combined by formula and candidates are selected in strict top-down rank order, or when candidates above specific cut scores are eligible to continue subsequent stages of a selection system. In other instances, information from a test is judgmentally integrated with information from other tests and with nontest information to form an overall assessment of the candidate.

Thus, the *Standards* discuss the difference between mechanical and judgmental (what we call holistic) decision-making, indicating that this is considered a topic of relevance for users of psychological tests. However, the *Standards* do not mention that mechanical judgment leads to more reliable and valid judgments

than holistic combinations of information. Second, the Standards incorrectly imply that mechanical decision-making can only be used when decisions are based exclusively on test scores and that taking information derived from other sources than standardized tests (such as interviews, biodata) into account requires holistic decision-making.

3.5.2 *International Test Guidelines*

The aim of the ITC test guidelines is described as follows (p. 7):

The Test Use guidelines relate to the competencies (knowledge, skills, abilities and other personal characteristics) needed by test users. These competencies are specified in terms of assessable performance criteria. These criteria provide the basis for developing specifications of the evidence of competence that would be expected from someone seeking qualification as a test user. Such competencies cover such issues as professional and ethical standards in testing, rights of the test taker and other parties involved in the testing process, choice and evaluation of alternative tests, test administration, scoring and interpretation, and report writing and feedback.

Furthermore, we encountered several statements that encourage using multiple sources of information and thus indicate that information will need to be combined (listed below, with original reference numbers). However, no explicit statement on *how* to combine information was found.

2.1.4 Seek other relevant collateral sources of information.

2.1.6 Ensure that full use is made of all available collateral sources of information.

4. Make clear that the test data represent just one source of information and should always be considered in conjunction with other information.

Thus, although the potential utility of testing in an assessment situation is discussed in the ITC guidelines, statistical versus holistic combination is not discussed. Furthermore, the statement in the ITC guidelines that “collateral information” is useful seems to imply that more information is better. However plausible this may sound, this is not true in general and can encourage problematic decision-making. For example, information from unstructured interviews when combined with valid grades can lower predictive validity compared to using grades alone, but at the same time increase the *feeling* of a valid decision (e.g., Dana et al., 2013).

3.5.3 *Conclusion on Decision-Making as Discussed in Test Guidelines*

Both guidelines pay little attention to obtaining reliable and valid judgments and decisions based on a combination of different sources of information (e.g., tests, interviews, questionnaires). In the vast majority of cases, psychological tests are

used with the main aim to aid decision-making about an individual, but the research literature on this issue is not discussed and its influence is minimal.

3.6 Concluding Remarks

The findings on how decisions can best be made based on a combination of information are exceptionally robust and should be highly consequential for psychological and educational practice, as well as other fields such as medicine and law (e.g., Arkes et al., 2008; Guay & Parent, 2018; Hanson & Morton-Bourgon, 2009; Schwab, 2008). Professionals and academic psychologists have a hard time accepting the superiority of statistical over holistic decision-making. Since Meehl (1954), a number of articles (e.g., Dawes, 1979; Grove & Meehl, 1996; Highhouse, 2008) addressed different types of objections with insightful explanations why these objections were unwarranted. As our results showed, 67 years after Meehl's publication, time has not resulted in a good understanding or appreciation of this topic in textbooks on psychological testing.

In some textbooks it is remarked that ethical guidelines of psychologists do not allow to completely rely on statistical decision-making. But as Murphy and Davidshofer (2005) discussed: "there are few excuses for not at least considering what a statistical model would say" (p. 530). Furthermore, using a statistical decision-making procedure does not imply that the psychologist is not responsible for the appropriateness of the procedure. As a reviewer remarked, The responsibility lies in selecting the relevant predictors, and setting up the rule to combine the information, but not so much second guess what the outcome is, every time the professional gets a 'hunch.' In fact, a psychologist should closely monitor the outcomes of statistical decision-making, use pilot studies, and intervene when things go wrong, for example, by excluding less valid predictors or adjusting the weights of a statistical rule. In fact, Dawes (2005) argued, and we agree, that it is unethical to not use a method that optimizes valid prediction.

If we take psychology as an applied science seriously, textbooks and test guidelines cannot stay behind in promoting an important finding in our field. Test guidelines form the link between scientific psychometrics and practice. It is thus the place where scientific findings can be disseminated to a wider audience. If we do not translate important empirical findings into guidelines for practice, our scientific findings will have very limited merit. When it comes to decision-making based on test scores, we think we can and should do a better job than we are doing at the moment.

When professionals do not adopt evidence-based procedures for test use, there are at least four possible reasons.

1. They do not have sufficient knowledge about the most appropriate procedures (Neumann et al., 2021a).
2. They do not believe in the evidence presented in scientific studies.

3. They know about and believe in the evidence presented in scientific studies, but do not act upon the evidence because of internal conflict (e.g., need for autonomy, Nolan & Highhouse, 2014).
4. They know about and believe in the evidence presented in scientific studies, but do not act upon them because of external pressures (e.g., stakeholder perceptions, being valued in their work, Nolan et al., 2020).

Including guidelines on test use and decision-making in test standards can help relieve all of these reasons. They can communicate the existing evidence to overcome reason 1, they can discuss common misconceptions and invalid counterarguments to overcome reason 2, and they serve to set a standard to resist both internal and external pressures that hinder using evidence-based prediction procedures. Therefore, we ask authors of textbooks and test guidelines to pay more attention to statistical decision-making.

As a final note, we started this chapter with observing that Drenth and Sijtsma (2006) devoted a whole chapter to the contribution of a test to the decision-making process. Statistical versus holistic prediction is part of this decision-making process, and the question the reader may have now is: How did they reflect on mechanical versus holistic prediction? Well, they did a good job. In response to the question, how should we combine the results of different tests? They noted that (p. 414; our translation):

First, this can be done via a statistical process of weighing test scores and possibly calculation of probabilities, and secondly via an intuitive, not statistical process of weighing and prediction. In this intuitive way it often concerns the different weighting across different cases; the process is less formalized, one does not follow a fixed strategy like in a statistical procedure.

Furthermore, they discussed that:

An evaluation of the many research findings in this context was in agreement with the original conclusions by Meehl that the statistical procedure is superior to the holistic method

And their explanation is (p. 414):

This result can be understood as follows. In a holistic combination of objective data, such as obtained in assessments with tests to predict an objective criterion, all kinds of biases, stereotypes, and unfounded assumptions play a role besides knowledge of the professional literature. One determines often on the basis of intuition the different weights, often in an inconsistent way. In this way some test scores are weighted too heavily, some are getting too few weights and per case and across different measurements there are fluctuations and inconsistencies.

Although they did not tick all the boxes in their chapter as suggested by us in Table 3.1, this phrasing of the main message of the statistical prediction literature was perhaps the most accurate description we found in the textbooks on psychological testing on statistical prediction.³

³ *References marked with an asterisk indicate works included in the review.

References

- Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341–382. <https://doi.org/10.1177/0011000005285875>
- Aiken, L. R. (2003). *Psychological testing and assessment*. Pearson Education Group.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- *Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.
- Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *The level of service/case management inventory*. Multi-Health Systems.
- Arkes, H. R. (2008). Being and advocate for linear models of judgment is not an easy life. In J. I. Krueger (Ed.), *Modern pioneers in psychological science: An APS-Psychology press series. Rationality and social responsibility: Essays in honor of Robyn Mason Dawes* (pp. 47–70). Psychology Press.
- Arkes, H. R., Shaffer, V. A., & Medow, M. A. (2008). The influence of a physician's use of a diagnostic decision aid on the malpractice verdicts of mock jurors. *Medical Decision Making, 28*(2), 201–208. <https://doi.org/10.1177/0272989X07313280>
- Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J.-F. Le Ny & W. Kintsch (Eds.), *Language and comprehension* (pp. 287–299). North-Holland Publishing Company.
- *Cohen, R. J., & Swerdijk, M. E. (2015). *Psychological testing and assessment: An introduction to tests and measurement*. McGraw-Hill Education.
- *Cooper, C. (2019). *Psychological testing, theory and practice*. Routledge.
- Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. University of Illinois Press.
- Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making, 8*(5), 512–520.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571–582. <https://doi.org/10.1037/0003-066X.34.7.571>
- Dawes, R. M. (2005). The ethical implications of Paul Meehl's work on comparing clinical versus actuarial prediction methods. *Journal of Clinical Psychology, 61*(10), 1245–1255. <https://doi.org/10.1002/jclp.2-180>
- Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95–106. <https://doi.org/10.1037/h0037613>
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. <https://doi.org/10.1126/science.2648573>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. <https://doi.org/10.1287/mnsc.2016.2643>
- *Domino, G., & Domino, M. L. (2006). *Psychological testing an introduction*. Cambridge University Press.
- Drenth, P. J. D. (1965). *De psychologische test* [The psychological test]. Van Loghum Slaterus.
- Drenth, P. J. D. (1975). *De psychologische test* [The psychological test] (2nd ed.). Van Loghum Slaterus.
- Drenth, P. J. D., & Sijtsma, K. (1990). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* [Test theory. Introduction to the theory of the psychological test and its applications]. Bohn Stafleu van Loghum.

- Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* [Test theory. Introduction to the theory of the psychological test and its applications]. Bohn Stafleu van Loghum.
- *Furr, R. M. (2018). *Psychometrics, an introduction*. SAGE.
- Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence for a method of improving on clinical inferences. *Psychological Bulletin*, 73(6), 422–432. <https://doi.org/10.1037/h0029230>
- *Gregory, R. J. (2013). *Psychological testing, history, principles, and applications* (7th ed.). Pearson Education.
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, 2(2), 293–323. <https://doi.org/10.1037/1076-8971.2.2.293>
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19–30. <https://doi.org/10.1037/1040-3590.12.1.19>
- Guay, J. P., & Parent, G. (2018). Broken legs, clinical overrides, and recidivism risk: An analysis of decisions to adjust risk levels with the LS/CMI. *Criminal Justice and Behavior*, 45(1), 82–100. <https://doi.org/10.1177/0093854817719482>
- Hanson, R. K. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment*, 21(1), 1–21. <https://doi.org/10.1037/a0014421>
- Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1(3), 333–342. <https://doi.org/10.1111/j.1754-943>
- *Hogan, T. P. (2015). *Statistical testing, a practical introduction*. Wiley.
- Holt, R. R. (1970). Yet another look at clinical and statistical prediction. *American Psychologist*, 25(4), 337–339. <https://doi.org/10.1037/h0029481>
- Holt. (1986). Clinical and statistical prediction. A retrospect and would be integrative perspective. *Journal of Personality Assessment*, 50(3), 376–386. https://doi.org/10.1207/s15327752jpa5003_7
- International Test Commission. (2013). *ITC guidelines on test use version 1.2*. Retrieved from: https://www.intestcom.org/files/guideline_test_use.pdf
- Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6), 515–526. <https://doi.org/10.1037/a0016755>
- Kahneman, D., & Klein, G. (2010). Strategic decisions: When can you trust your gut? *McKinsey Quarterly*, from <https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/strategic-decisions-when-can-you-trust-your-gut?cid=other-soc-lkn-mip-mck-oth-1912%2D%2D&sid=2972122698&linkId=79428658#>
- *Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological assessment and theory. Creating and using psychological tests. International Edition*. Central Learning.
- Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, 134(3), 404–426. <https://doi.org/10.1037/0033-2909.134.3.404>
- *Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. SAGE. <https://doi.org/10.4135/9781483385693>
- Krathwohl, D. R. (1998). *Methods of educational and social science research: An integrated approach*. Longman.
- Kuncel, N. R. (2008). Some new (and old) suggestions for improving personnel selection. *Industrial and Organizational Psychology*, 1(3), 343–346. <https://doi.org/10.1111/j.1754-9434.2008.00059.x>
- Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, 98(6), 1060–1072. <https://doi.org/10.1037/a0034156>

- Meehl, P. E. (1954). Empirical comparisons of clinical and actuarial prediction. In *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.
- Meijer, R. R., Neumann, M., Hemker, B. T., & Niessen, A. S. M. (2020). A tutorial on mechanical decision-making for personnel and educational selection. *Frontiers in Psychology, 10*, 3002. <https://doi.org/10.3389/fpsyg.2019.03002>
- Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science, 4*(4), 379–383. <https://doi.org/10.1111/j.1745-6924.2009.01142.x>
- *Miller, L. A., McIntire, S. A., & Lovler, R. L. (2015). *Foundation of psychological testing. A practical approach* (5th ed.). SAGE.
- Morris, S. B., Daisley, R. L., Wheeler, M., & Boyer, P. (2015). A meta-analysis of the relationship between individual assessments and job performance. *Journal of Applied Psychology, 100*(1), 5–20. <https://doi.org/10.1037/a0036938>
- Murphy, K. R. (2019). Understanding how and why adding valid predictors can decrease the validity of selection composites: A generalization of Sackett, Dahlke, Shewach, and Kuncel (2017). *International Journal of Selection and Assessment, 27*(3), 249–255. <https://doi.org/10.1111/ijsa.12253>
- *Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing, principles and applications* (6th ed.). Pearson Prentice Hall.
- Murphy, K. R., Deckert, P. J., Kinney, T. B., & Kung, M. C. (2013). Subject matter expert judgments regarding the relative importance of competencies are not useful for choosing the test batteries that best predict performance. *International Journal of Selection and Assessment, 21*(4), 419–429. <https://doi.org/10.1111/ijsa.12051>
- Neumann, M., Hengeveld, M., Niessen, A. S. M., Tendeiro, J. N., & Meijer, R. R. (2021a). Education increases decision-rule use: An investigation of education and incentives to improve decision making. *Journal of Experimental Psychology: Applied*. Advance online publication. <https://doi.org/10.1037/xap0000372>
- Neumann, M., Niessen, A. S. M., & Meijer, R. R. (2021b). Implementing evidence-based assessment and selection in organizations: A review and an agenda for future research. *Organizational Psychology Review, 11*(3), 205–239. <https://doi.org/10.1177/2041386620983419>
- Neumann, M., Niessen, A. S. M., Tendeiro, J. N., & Meijer, R. R. (2021c). The autonomy-validity dilemma in mechanical prediction procedures: The quest for a compromise. *Journal of Behavioral Decision Making*. Advance online publication. <https://doi.org/10.1002/bdm.2270>
- Nolan, K. P., & Highhouse, S. (2014). Need for autonomy and resistance to standardized employee selection practices. *Human Performance, 27*(4), 328–346. <https://doi.org/10.1080/08959285.2014.929691>
- Nolan, K. P., Dalal, D. K., & Carter, N. (2020). Threat of technological unemployment, use intentions, and the promotion of structured interviews in personnel selection. *Personnel Assessment and Decisions, 6*(2), 38–53. <https://doi.org/10.25035/pad.2020.02.006>
- *Reynolds, C. R., & Livingston, R. B. (2012). *Mastering modern psychological testing*. Pearson.
- Roulin, N., & Bangerter, A. (2012). Understanding the academic-practitioner gap for structured interviews: “Behavioral” interviews diffuse, “structured” interviews do not. *International Journal of Selection and Assessment, 20*(2), 149–158. <https://doi.org/10.1111/j.1468-2389.2012.00588.x>
- Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology, 40*(3), 455–488. <https://doi.org/10.1111/j.1744-6570.1987.tb00610.x>
- Sackett, P. R., Dahlke, J. A., Shewach, O. R., & Kuncel, N. R. (2017). Effects of predictor weighting methods on incremental validity. *Journal of Applied Psychology, 102*(10), 1421–1434. <https://doi.org/10.1037/apl0000235>
- Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology, 48*(5), 593–602. <https://doi.org/10.1086/219248>

- 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(2), 262–274. <https://doi.org/10.1037/0033-2909.124.2.262>
- Schwab, A. P. (2008). Putting cognitive psychology to work: Improving decision-making in the medical encounter. *Social Science & Medicine*, 67(11), 1861–1869. <https://doi.org/10.1016/j.socscimed.2008.09.005>
- Taylor, H. C., & Russel, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, 23(5), 565–578. <https://doi.org/10.1037/h0057079>
- Terpstra, D. E., & Rozell, E. J. (1997). Why some potentially effective staffing practices are seldom used. *Public Personnel Management*, 26(4), 483–495. <https://doi.org/10.1177/009102609702600405>
- Van der Linden, W. J. (1991). Applications of decision theory to test-based decision making. In R. K. Hambleton et al. (Eds.), *Advances of educational and psychological testing: Theory and applications*. Springer Science + Business Media.
- Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice*, 40(5), 525–531. <https://doi.org/10.1037/a0014693>
- Wiggins, N., & Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology*, 19(1), 100–106. <https://doi.org/10.1037/h0031147>
- Yu, M. C., & Kuncel, N. R. (2020). Pushing the limits for judgmental consistency: Comparing random weighting schemes with expert judgments. *Personnel Assessment and Decisions*, 6(2), 1–10. <https://doi.org/https://scholarworks.bgsu.edu/pad/vol6/iss2/2>

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 4

Trustworthy Artificial Intelligence in Psychometrics



Bernard P. Veldkamp 

Abstract The availability of sensors, eye-trackers, smartwatches, Wi-Fi trackers, or other digital devices facilitates the collection of new types of data that can be used for measurement. The question is how to analyze them. Several psychometric models are available, but even though they have been applied successfully in many testing programs, they do have their limits with respect to the kind of data they can be applied to. Artificial intelligence (AI) offers many methods for dealing with these new and more complex datasets. They do have some limitations when it comes to reliable and valid measurement thought. The question arises how to apply them in the field of psychometrics. To answer this question, the field of psychometrics is introduced first. Besides, the benefits and disadvantages of artificial intelligence are illustrated in three examples. A promising development, when it comes to the application of AI in the field of psychometrics, is referred to as trustworthy AI (TAI), with principles related to fairness, explainability, and accountability. Based on the examples of the use of AI in social and health science and the lessons learned from the approaches to integrate new data types in existing psychometric models, a framework is defined with nine steps for the use of AI in psychometrics. For each of these steps, it is evaluated how TAI can be applied for reliable and valid measurement. The chapter concludes with the observation that straightforward application of AI in the field of Psychometrics might still be a step too far, but that the developments related to TAI go fast and offer new and exciting opportunities for the application of AI to psychometrics.

Accurately measuring and confirming conclusions about human behavior, skills, knowledge, abilities, interests, values, and attitudes or the impact of interventions is considered to be of great importance in the social and health sciences. For this

B. P. Veldkamp (✉)

Department of Learning Data and Technology, Faculty of Behavioral Management and Social Sciences, University of Twente, Enschede, The Netherlands

e-mail: b.p.veldkamp@utwente.nl

purpose, standardized tests have been used in the past decades. Psychometric models are available for analyzing the response data, and reliable and valid inferences can be made. Recently though, different types of data became available. Text, video, audio, logfile, or sensor data related to human behavior have been generated in large volumes. Social media could be a source, or data could be found online in data repositories. Besides, the availability of sensors, eye-trackers, smartwatches, Wi-Fi trackers, or other digital devices even facilitated the collection of different types of data. The vast pool of available data could be a potential source of information to reveal new insights about human behavior, but the question is how to unlock and analyze it.

Artificial intelligence (AI) offers many methods for dealing with large datasets (Veldkamp, 2018). Natural language processing, image and speech recognition, or computer vision could be applied to derive meaningful information from the data. Machine learning, deep learning, supervised learning, unsupervised learning, or reinforcement learning (Panch et al., 2018) can be applied to derive inferences. This can be done with the help of specialized software like CRAN (2021), Python (Van Rossum & Drake, 2009), or MATLAB (Mathworks, 2021). AI can be seen as a methodology that combines data, algorithms, and computing power. Great successes have been achieved with these methods in areas like geography (Zhu et al., 2017), computer vision (Voulodimos et al., 2018), health care (Miotto et al., 2018), or text mining (Liang et al., 2017). Unfortunately, these AI methods are often very complex and hard to interpret for humans. For example, deep learning algorithms mainly function as black boxes where it is impossible to understand the functional relationship between the data (input) and the inferences (output). Besides, many examples can be found where the application of AI resulted in biased inferences (e.g., O'Neil, 2016). For the application of AI in social and health science, this poses some challenges.

A promising direction, therefore, is referred to as trustworthy AI (TAI; Floridi, 2019). Thiebes et al. (2021) mention the following foundational principles of TAI: (1) beneficence, (2) non-maleficence, (3) autonomy, (4) justice, and (5) explicability. Where beneficence, non-maleficence, and justice have many implications for the use and effect of the use of AI in society, the principles of autonomy and explicability have many implications for doing research with AI. Autonomy refers to the ability of humans to be in the lead. In a research context, this implies that researchers need to implement an oversight mechanism that enables them to control the algorithms during the entire process. This is also referred to as keeping the human in the loop when implementing AI methods. Explicability, on the other hand, is about both explainability and accountability. Explainable AI creates models that are interpretable while maintaining high levels of performance and accuracy (Floridi et al., 2018). Accountable AI creates models that are transparent and controllable. Thiebes et al. (2021) therefore conclude that AI can be trusted if its algorithms are fair, can be understood, and are capable to do what needs to be done.

This chapter focuses on the use of TAI for inferring conclusions about the abilities, aptitudes, or attitudes of individuals based on data. The research question that guided this study is:

How can trustworthy artificial intelligence be integrated into the field of psychometrics?

First, the field of psychometrics is introduced, and several models to analyze various data types within the existing framework of psychometrics are described. After that, three examples are presented that illustrate the issues involved with analyzing data with AI in the social and health sciences. Subsequently, a general framework for the use of AI in the field of psychometrics is presented. The chapter concludes with an outlook on the trustworthy application of AI in psychometrics.

4.1 Psychometrics

The science of inferring conclusions about abilities, aptitudes, or attitudes of individuals based on data is generally referred to as psychometrics. More specifically, psychometrics comprises the development, appraisal, and interpretation of psychological tests and other measures used to assess variation in behavior and to link such variation to psychological conditions (Committee on psychological testing, 2015). Psychometrics is about relating variability in data to variability in latent constructs. The concept of latent in latent construct refers to the fact that these constructs cannot be observed directly. We have to rely on data and make inferences about the underlying constructs that account for them. By carefully collecting data using tests of questionnaires, an attempt is made to minimize the measurement error and to come to reliable inferences. Psychometric models have been developed to guide this process. Based on dichotomously or polytomously scored items, they can be applied to make inferences about the underlying constructs.

The classical test theory (CTT) model is among the earliest models available. It is built on the assumption that the observed score (X) of a respondent consists of a true score (T) and an error component (E):

$$X = T + E$$

By adding the assumption that the errors are normally distributed around zero and that the errors are uncorrelated with the true score, it can be shown that the expected observed score equals the true score of the respondent. See also Lord and Novick (1968) for an in-depth introduction to classical test theory. With the help of CTT, it was not only possible to draw inferences about the true underlying constructs or abilities of the respondents; with the help of reliability and validity indices (Lord & Novick, 1968; Sijtsma, 2009; Sijtsma & van der Ark, 2015), statements could be made about the quality of these inferences as well.

There was some criticism about CTT though. The score of the respondent can only be interpreted in the context of the specific measurement instrument, the standard error of measurement is assumed to be the same for all respondents, most generally applied reliability indices only provide lower bounds for the reliability,

and inferences can only be made at the test level and not at the individual item level. As an alternative, item response theory (IRT) models were proposed (Hambleton & Swaminathan, 1985; Lord, 1980). Based on the assumptions of unidimensionality and local independence, a series of models have been developed for both dichotomous and polytomous data. The Rasch model for dichotomous items states that the probability of a correct response ($X = 1$) to an item given the ability of the respondent (θ) can be modeled using a logistic function:

$$P(X = 1|\theta) = \frac{e^{(\theta-b)}}{1 + e^{(\theta-b)}},$$

where b denotes the difficulty of the item. More complex IRT models were developed with multiple item and person parameters, both for dichotomous and polytomous responses (see Embretson & Reise, 2013). Besides models that assume a parametric relationship between the ability, the item, and the probability of a correct response, non-parametric IRT (NIRT) models were developed (Sijtsma & Molenaar, 2002) that do not make assumptions about the specific functional relationship, but they typically specify order restrictions like the monotonicity assumption:

$$P(\theta_a) \leq P(\theta_b),$$

whenever $\theta_a < \theta_b$. Typical for these NIRT models is that they were developed to relax the IRT assumptions as much as possible while maintaining essential measurement properties (Sijtsma & Meijer, 2006).

Just to mention a few advantages of IRT, the possibility to compare test scores over different tests and different test administrations added to the popularity of these models. Besides, because item parameters are claimed to be sample independent, i.e., independent of the particular sample of items and/or examinees chosen (Hambleton & Swaminathan, 1985), IRT facilitated item banking and computerized adaptive testing. Nowadays, most large-scale operational testing programs, therefore, rely on IRT to relate response behavior to the underlying constructs.

Both CTT and IRT models have been developed to handle response data. Different models are available to work with other data types like response times and self-narratives as well.

4.1.1 Response Times

Both CTT and IRT models only rely on response data to make inferences about the underlying latent constructs. In addition, other sources of information are available that could be of use. These sources of information are often referred to as process data. Process data is about the series of actions a respondent performed but also

about the time they spent on these actions. Response times might reveal useful information about how well respondents master a given task. This notion was, for example, used to develop speed tests in which respondents are challenged to complete as many tasks as possible within a given time frame. All tasks are relatively easy, and the objective is to demonstrate processing speed while minimizing errors.

Also in educational measurement, test speededness plays a role. Unlike traditional speed tests, the tasks are not easy and there is a spread in the difficulty of the items. However, the time to finish a test is often still limited. Response times might therefore reveal information about the respondents and the items. Van der Linden (2007) proposed a general hierarchical framework for the concurrent modeling of response data and response times. In this framework, several assumptions were made: (1) a respondent is working at a constant speed; (2) for a fixed respondent, both response and response times are assumed to be random variables; (3) separate item and person parameters for both the distributions of responses and response times; (4) conditional independence between responses and response times given the levels of ability and speed; and (5) separate models for responses and response times can be combined and estimated within a hierarchical framework using a Bayesian approach.

The link between response and response time parameters within this framework allows the combination of information from both sources. Because of this, it is possible to improve inferences that are made on response data only, when an IRT model is applied. For example, response time information can be used to improve the estimation of ability and item parameters. Besides, response times can be used to deal with issues of speededness in testing or detection of aberrant response behavior.

By extending existing IRT models with log-normal response models, process data could be incorporated, and the performance of candidates could be modeled more accurately compared to models that model the performance based on the final responses only.

4.1.2 Self-Narratives

Questionnaires or other item-based measurement instruments are generally applied to measure different kinds of psychological constructs because they are easy to administer, don't take much time, and don't need the presence of a psychologist. These instruments have been carefully developed to provide the necessary information with a minimum number of items. However, respondents might feel limited in expressing themselves. An alternative would be to ask respondents to share their stories in a self-narrative. It is challenging to analyze the large body of open text, but such a story could provide useful information as well (He et al., 2012), even though the information might not be perfect and self-narratives might contain noninformation as well.

To combine textual data and information from questionnaires, He et al. (2019) applied a Bayesian approach. Within Bayesian statistics, inferences are made based

on a posterior, which combines prior beliefs about the construct of interest with a likelihood of the observed responses. He et al. (2019) used self-narratives for defining the prior beliefs and combined them with the likelihood of the questionnaire data. In this way, both data sources could be combined in one posterior distribution.

4.1.3 Lessons Learned

The use of response times and self-narratives demonstrates that response data can be combined with different data sources by extending existing psychometrics models. The general hierarchical framework of van der Linden (2007) combines the power of response time modeling and multilevel Bayesian statistics. He et al. (2019) also built upon the strengths of Bayesian statistics by eliciting informative priors using new data sources. By carefully modeling these new data types, they could be added to existing psychometrics, and the measurement process could be enriched.

To benefit from the opportunity to build and compute the complex Bayesian models that were needed, a high level of specialized knowledge in computational Bayesian psychometrics is required. Besides, given the variety of the new data, such an approach might not be applicable. In the next section, three examples will be presented that apply AI to come to inferences about underlying constructs based on a variety of data sources.

4.2 Examples of the Use of AI

Three examples are discussed that show various applications of AI in the social, behavioral, and health sciences. They illustrate the issues involved in analyzing data with AI. First, a deep learning algorithm will be presented that can be applied to analyze continuous physiological data streams to predict workload. Second, a face-scanning algorithm is described that was applied to predict personality. Finally, patient reports will be analyzed to predict post-intensive care syndrome, that is, cognitive, psychiatric, and/or physical disability of survivors or relatives after treatment in the intensive care unit (ICU).

4.2.1 Perceived Mental Workload

In the process of optimizing team or individual performance, attention is often paid to perceived mental workload. How does the task load related to the capacity of the persons involved? To perform optimally, the ratio between available and required cognitive resources is assumed to be more or less equal to 1 (Csikszentmihalyi, 1997).

Different physiological signals have been related to workload. Heart rate, stress, and brain activity increase when workload increases. Attention to the task relates to workload as well. To measure these variables, photoplethysmography (PPG; heart rate), galvanic skin response (GSR; stress), functional near-infrared spectroscopy (fNIRS; brain activity), and eye-tracking (ET; attention to the task) can be applied. To find out which signals would provide the most useful information, Dolmans et al. (2020) collected PPG, GSR, fNIRS, and ET data of participants who carried out tasks of different difficulty levels.

To handle such multimodal data, different approaches can be used. Continuous data streams can be merged using early, intermediate, or late fusion. When data fusion is applied, various data streams are merged or integrated to produce more consistent, reliable, and useful information. Each fusion strategy comes with its own (dis)advantages. Early fusion asks for data harmonization before the data is fed into the network. When sampling rates vary for the different data collection devices, this might be a problem. Intermediate fusion does not suffer from these problems, but it asks for a complicated multilevel architecture of the deep neural net. Late fusion analyzes the data streams separately and generates an output based on a kind of majority vote. Dolmans et al. (2020) opted for intermediate fusion. For each modality a separate neural net consisting of several layers was designed; the outcomes of these individual nets were the combined input of a second neural net consisting of another series of layers. Figure 4.1 illustrates the architecture of the network. First the four data streams are analyzed separately, then the outputs of the separate nets are combined in a concat layer, and finally they are analyzed concurrently in the second part of the network.

Even though the workload, as measured with the NASA Task Load Index (Hart & Staveland, 1988), was predicted rather successfully with the combined data streams coming from various devices, issues related to synchronizing the devices turned out to be quite difficult. Besides, training the deep neural nets and optimizing the settings were rather time-consuming and sensitive to small changes in settings of the different layers. This application is just one example of the capability of AI to analyze multimodal datasets.

4.2.2 Face Recognition

Personality testing with the help of AI recently received considerable attention (Liem et al., 2018) in job selection, because of its efficiency and scalability. Machine learning algorithms have been developed to predict personality based on facial features. Deep learning algorithms were trained to analyze video material. Röber (2021) implemented a deep neural net in OpenCV (2020) that was capable of detecting facial landmarks in faces at different orientations and scales and under substantial occlusion.

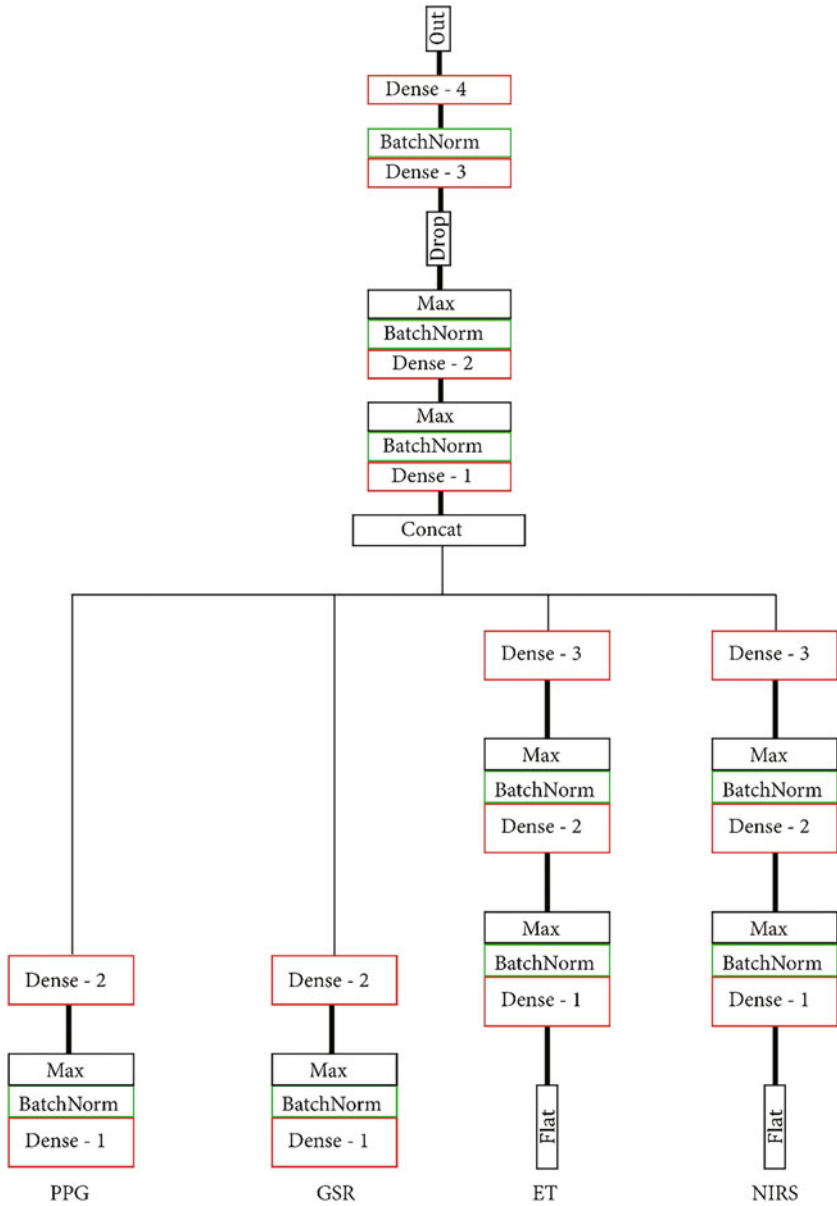


Fig. 4.1 The architecture of the deep neural net for analyzing multimodal data

Facial landmarks (Fig. 4.2) are salient features of faces such as eyes and eyebrows, nose, mouth, or jawline. Besides, pupil locations were detected. Based on these data, features or variables could be generated like distances between landmarks and variation in pupil location. When respondents were classified as

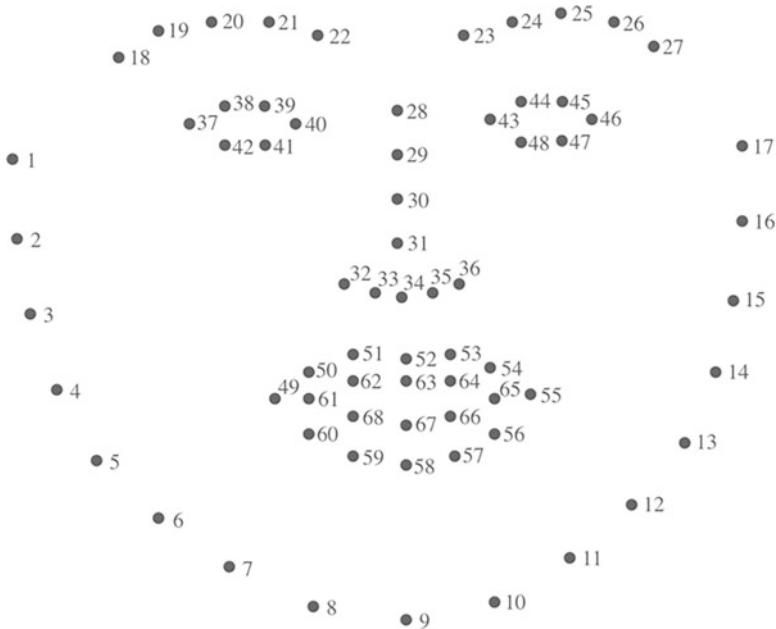


Fig. 4.2 68 Facial landmarks that were detected

either scoring high or low for the different personality traits, the accuracy, that is, the percentage of correct classifications of the algorithm ranged from 0.84 for openness to 0.94 for conscientiousness (Röber, 2021). Even though the outcome (high or low) is a rather rough dichotomous measure, these results seem to verify the use of face recognition as the first screening for candidates in a job selection context from a technical point of view.

Ethical and legal issues have to be considered though. How about privacy of participants? Do they provide consent for the use of their image or video? How long will the data be stored? The use of face recognition also received considerable criticism, both in the public and in the scientific domain. The Netflix documentary *Coded Bias* illustrated flaws in facial recognition technology, and it has been shown that results can be contradictory. More specifically, Kachur et al. (2020) reported that personality traits can be better predicted for female faces, whereas Hu et al. (2017) reported that personality could be better predicted from male faces. Escalante et al. (2018) described that ethnicity affects the performance of face recognition technology. Finally, Abdurrahim et al. (2018) mentioned that besides gender and ethnicity, age-related bias might occur when face recognition is applied.

Even though Keszler (2021) did not find any significant effects of gender, ethnicity, or age when the models of Röber (2021) were applied, it can be concluded from various studies in this field that the sample for which the face recognition models are trained is very important and that various sources of bias might disturb its performance and when they are applied. So, even though AI could be applied for

face recognition, this example also illustrates that ethical considerations might have to be taken into account when AI is applied.

4.2.3 Post-Intensive Care Syndrome (PICS)

PICS refers to cognitive, psychiatric, and physical problems after treatment at an intensive care unit (ICU). Cognitive impairment was reported to occur on average in 25% of ICU survivors, psychiatric illnesses were reported by up to 62%, and ICU-acquired neuromuscular weakness was reported by 25% of the survivors (Rawal et al., 2017). To detect survivors that are at risk of developing PICS in an early stage, the University of Twente and Medisch Spectrum Twente hospital started a study where patient open format self-reports, in which they describe how they are doing, were screened for risk of PICS. An initial data collection resulted in responses of 261 ICU survivors and relatives. Both survivors and relatives provided a (short) description of what they experienced and how they felt. Out of this sample of survivors and relatives, 33 were identified as having PICS-related symptoms. Data were pre-processed by stemming algorithms that among others remove conjugations and by removing stop words.

Analysis of the self-reports using sentiment analysis revealed that both the PICS and non-PICS groups hardly differed in the kind of words they used (see Fig. 4.3), when it comes to different sentiments or emotions. To summarize the different emotions expressed in the texts, Plutchik’s wheel of emotions (Abbasi & Beltikov, 2019) was applied. Survivors and relatives in the PICS group only showed higher word counts for sentiments “anger” and “surprise” and lower word counts for “joy.”

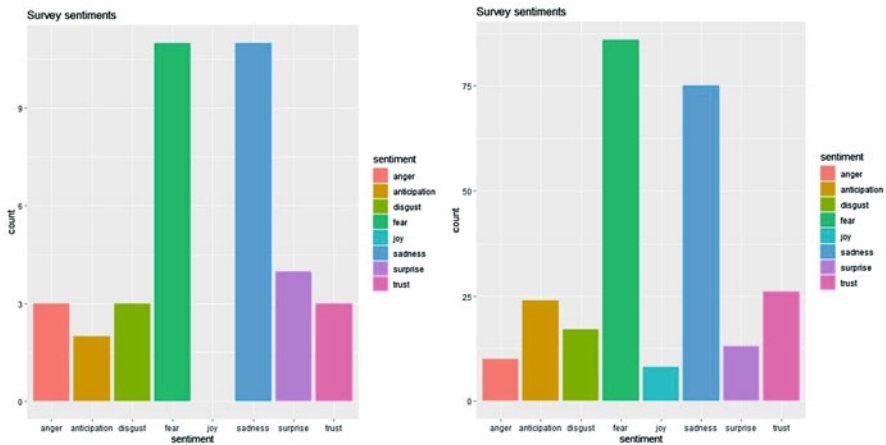


Fig. 4.3 Scores for Plutchik’s wheel of emotions for the PICS group (left) and the non-PICS group (right)



Fig. 4.4 Unigrams for relatives with (a) and without PICS (b) and survivors with (c) and without PICS (d)

Analysis of the unigrams (frequencies of individual words) revealed that the most common words in both groups were “Het gaat best wel goed” (“It goes rather well”). Only when these words were removed from the text corpus, differences in unigrams occurred for both the survivors and relatives with and without PICS (Fig. 4.4).

Unfortunately, the remaining word frequencies (after removing the frequently used words “it goes rather well”) were very low, and linguistic features were not suitable yet for building reliable prediction models. One of the explanations was that it is in the character of the people living in the Twente region not to complain. Taking this information into account, it makes sense that the words “it,” “goes,” “rather,” and “well” are among the most frequently used words. A self-report on how people from the Twente region are doing therefore does not provide useful information for predicting PICS. On top of this regional effect, it is quite common in the Netherlands to respond to the question how they are doing in quite general wordings. This might have had an effect as well.

Even though AI was applied successfully to unlock large bodies of open text, the resulting models did not have any predictive power for the construct of interest.

4.3 Framework for the Use of AI in Psychometrics

The field of psychometrics has been confronted with new voluminous, veracious, and variable data sources. AI methods have been developed to deal with these data types, but before they can be used in the field of psychometrics, it has to be guaranteed that they meet quality standards comparable to those of existing psychometric models. Research on response times modeling and the use of informative priors are successful examples of extensions of psychometrics with other data types. A more general approach is needed to fully integrate new data types and AI.

A promising development is the focus on trustworthy AI with principles related to fairness, explainability, and accountability. Based on the three examples of the use of AI in social and health science and the lessons learned from the approaches to integrate new data types in psychometric models, an attempt was made to define a framework with useful steps for the use of AI in psychometrics. It is important to notice that many AI applications have been developed with an emphasis on predictive validity, whereas in the area of psychometrics, construct validity is of great importance as well. The framework is broken down into nine steps that each emphasize a different aspect of the use of AI in psychometrics.

1. *Defining the Clear Measurement Goal*

AI can be used to handle large datasets. Both the PICS and the workload examples illustrate that a large number of variables, where some of the variables only carry a limited amount of information by themselves (individual words are not very informative) and are very different (e.g., heart rate and eye movement), can be combined into one model. One of the challenges involved in handling such large and diverse datasets is to prevent that the application of AI turns into a phishing expedition, where a model is being built by repeatedly adding variables until a relation is found.

One way to prevent this is by formulating the goal of research or the goal of measurement in advance. This is also very much in line with the empirical cycle of research as formulated by De Groot (2019). Based on observations or previous research, hypotheses are being formulated that can be tested using empirical information. So before a start is made to explore the data, a hypothesis should be formulated that specifies the constructs of interest and guides the process of AI. For research purposes, one could consider to preregister the study, which is quite common in, for example, random clinical trials.

2. *Making an Inventory of Available Data*

The next step is to take inventory of available data and to collect the information that is available about these datasets. Such an inventory can result in a large and diverse set of data sources. Veldkamp et al. (2021), for example, identified ten different types of data sources when they searched within schools for information that could be informative about the performance of individual pupils. The sources vary from large-scale international surveys like PISA (Schleicher, 2019) to notes

taken during teacher meetings. An important notion is that the General Data Protection Regulation (GDPR; Voigt & von dem Bussche, 2017) requires that appropriate consent needs to be given before the data can be used.

When the data has been identified, accessibility and usefulness have to be explored. The granularity of the data might vary, ownership of the data might vary, and permission to work with the data is not always granted. The data might be incomplete, and it might be unknown how it was collected, so there might be a lot of uncertainty involved. Given the goal of measurement, it has to be considered which data will be included in the analyses. It might even be necessary to collect new data if the available data does not suffice. One of the most important outcomes of this step is that the choices that have been made are described and accounted for.

3. *Preprocessing the Data*

Different data types come in different formats. For example, audio files, heart rate signals, or text documents have to be processed to make this raw data accessible for analysis. For many data formats, specialized software is available, but for others, this software is still under development. New data sources come with new challenges. In the workload example, we struggled with the problem of handling data from various sources. Physiological data were recorded by devices and the suppliers did their best to support us, but a lot of time had to be invested in harmonizing the various signals. Specialized knowledge is often needed, not only to handle the data but also to distinguish data from noise. Blázquez-García et al. (2021) wrote a review on methods that are available for outlier detection, and they distinguish three different kinds of outliers, individual observations, several consecutive observations that are unusual, or entire signals that are not useful for analysis. Notably, however, removing data is not without consequences. Bakker and Wicherts (2014) illustrate the dangers of outlier removal and the consequences of different analysis techniques. Several strategies have been developed to define, to detect, and to handle different kinds of outliers (e.g., Aguinis et al., 2013). It remains a topic of research how to apply these strategies to the development of AI models.

4. *Identification and Selection of Variables*

Once raw data is accessible, variables can be derived. The face recognition example illustrates this process. First, facial landmarks have to be detected in video data. Based on these landmarks, variables can be defined related to, for example, variation in pupil location. The number of variables that can be defined is often very large. With textual data, variables can be defined related to the frequency of individual words, frequency of combination of words, total word count, sentence length, but also linguistic to linguistic features like the number of verbs and complexity, or even interpretations of sentences or entire texts. Given the goals of measurement, a selection has to be made about which variables to generate. Suggestions for variable selection can often be found in the literature, or the variables might be suggested by the preprocessing software.

Once the variables have been defined, they can be applied in the modeling process. Here too it holds that selection of variables is both theory-based and data-driven. A theory might provide guidance about which variables are assumed to be relevant. On the other hand, one of the strengths of AI is that new variables of interest can be found by exploring the data realm. An initial selection of promising variables could be made based on theory and by using straightforward measures like the correlation between input and output variable in case of a continuous output or by applying chi square-based measures in case of a categorical output (e.g., He et al., 2012). Which and how many of the variables are selected for the final model depends on the performance of the models.

It should be mentioned though that the variable selection also depends on the purpose of the model. There is a difference between models that focus on prediction and models that focus on measurement. If the model is designed to measure a specific construct, a theoretical underpinning of the variable selection process is very relevant because of validity issues. On the other hand, when the purpose is prediction, the focus is on the quality of the outcomes, and it is considered less important which variables are identified and selected to come to an optimal prediction.

5. Separating Training from Test Data

In psychometrics, model fit indices are available that provide information about how well the model represents the data. These indices can be used to interpret the parameters. Person fit indices (e.g., Meijer & Sijtsma, 2001) provide information about how well the model can analyze an individual's response pattern and can provide information about possible aberrant behavior like cheating. These model fit indices provide information about how well the inferences of the model can be trusted.

Such model fit indices are not available in AI, unfortunately. Therefore a different approach is needed. Various performance measures have been proposed (e.g., Dinga et al., 2019). An important notion is that these performance measures should not be calculated over the same data set as the one that was used to develop the model, because of the risk of over-fitting. The set of available data is therefore split into a training set and a test set. The test set is set aside and the training set is used to develop the model. This could be a 50/50 percent split, a 70/30 percent split, or even an 80/20 percent split. The choice of split proportion should be made keeping in mind that the training set has to be large enough to train a stable model for the set of variables at hand, preferably using resampling methods to reduce the error rate and to increase model robustness (James et al., 2021). After training the model, the performance can be tested by applying the model to the hold-out test set. In such a way, reliable information about the performance is obtained.

It should be mentioned that both the training and the test set come from the same original data set. To prevent the resulting common method bias, it would even be stronger if a second independently collected dataset would exist, to test the performance of the model.

6. *Choice of Data Mining Method and Hyperparameter Optimization*

Many different methods for unsupervised and supervised learning, either for classification or prediction, have been proposed (Hastie et al., 2008; James et al., 2021). It is beyond the scope of this framework to mention all of them. One thing these methods have in common is the use of hyperparameters, parameters that are used to control the learning process. They could refer to, for example, the number of subsets in cross-validation or the way the data was split in a training and a test set. To optimize the learning process, optimal hyperparameters have to be chosen. A grid search, where all combinations of hyperparameters are systematically evaluated, could be used for this purpose but is very time-consuming. Other approaches for hyperparameter optimization of hyperparameter tuning are available (e.g., Feurer & Hutter, 2019).

It was already mentioned that the explainability of AI models is an important issue when it comes to gaining trust in the outcomes and applications in practice. To measure human behavior, either in the context of psychological, educational, or health measurement, the choice of data mining methods is therefore limited to those methods that are transparent and interpretable. This is different from an AI model that was developed to optimize prediction. Also in Step 6, it holds that the choices being made have to be registered and accounted for to gain trust in the use of the model.

7. *Interpreting the Results*

Fairness, explainability, and accountability are seen as core values in the use of trustworthy AI. Because of the interpretability of TAI algorithms, we can connect the outcomes of the model to the goal formulated in Step 1. Visual aids can be applied to facilitate the interpretation for a more general audience. Besides, the choices that were made during the process are documented, which guarantees that the whole process can be accounted for.

Within the context of psychometrics, error of measurement and reliability are typically reported to provide information about the quality of the model. For AI, such indices are missing. As an alternative, resampling methods like bootstrapping or cross-validation can be applied when training the model to provide information about the uncertainties in the parameters and about to what extent the resulting scores can be relied upon.

8. *Validating the Results*

Besides the reliability of the results, validity is a core concept in psychometrics. Validity refers to whether the resulting score provides trustworthy information about the constructs of interest that were mentioned when the goal of measurement was specified. In the case of AI, it implies that the outcomes or predictions can be applied with enough fidelity for measurement purposes. AI typically relies on hold-out samples to demonstrate whether the results hold for different populations. However, within a measurement context, validity comprises more than the performance of the model on a hold-out subset of the initial sample.

The argument-based approach of Kane (2006, 2013) is widely applied in validation theory and can be of help. Kane mentions that validity is related to the intended use of the outcomes. To substantiate whether the outcomes of a model are used appropriately, all steps to come from observations to the use of the outcomes in practice have to be substantiated. Therefore, the whole train of thought, to come from raw data to inferences about the scores, has to be analyzed and accounted for. Finally, the validity is based on a combination of these arguments. An independent check of both the results and the validation process will increase trust.

9. Implementation in Practice

Once the reliability and validity of the AI model have been substantiated, the outcomes can be implemented in practice. Developing a user-friendly robust implementation that can be applied by practitioners is a challenge on its own, which goes beyond this chapter. It is important nonetheless to keep in mind that the original goal of measurement should still be leading here. A final step is to provide the intended user with proper documentation of how the AI-based measurement process was conducted. To facilitate the communication of this process, data visualization tools should be applied that translate the data mining process into charts, graphs, or other visuals. This kind of tools can also be of much help in communicating the measurement results with the persons who have been measured.

4.4 Conclusion

Psychometrics is a field that is specialized in measuring human behavior, skills, knowledge, abilities, interests, values, and attitudes, or the impact of interventions. Classical test theory and (non-)parametric item response theory have proven their value in this field. Recently, an abundance of new data sources has become available that might be useful for deriving inferences about these constructs as well. Most of them cannot be analyzed using existing models. Besides, the volume, velocity, and variety of these data sources ask for a different approach. To handle these data sources, a set of tools and algorithms referred to as AI can be applied.

Unfortunately, black-box models, like deep neural nets, are so complicated that they cannot be interpreted by humans, which causes all kinds of validity issues. Recently, however, the topic of trustworthy AI has received considerable attention. These AI models meet the principles of interpretability and accountability. Neumann et al. (2021) recently demonstrated that people prefer algorithms in which they retain some influence or know how they can affect the predictions or outcomes. Emphasizing interpretability and accountability of AI will strengthen the acceptance of AI in the area of psychometrics. In this chapter, an attempt was made to formulate a pipeline for the trustworthy application of AI in psychometrics. The next step would be to operationalize reliability for various AI algorithms and to provide a framework for building validity arguments for AI models applied for measuring human behavior, skills, or abilities.

One of the questions that remain is whether AI can be used to its full potential when resulting models have to be explainable and accountable. Deep neural nets that do not meet these requirements have been applied successfully for many problems. It seems to be a pity that these powerful methods are deemed inapplicable. One should bear in mind however that these powerful methods focus on prediction and that there is a difference between prediction and measurement. With prediction, the quality of the outcome is decisive. With measurement, the process has to be accountable as well. So, for prediction, these models might still be a valid option. For application in psychometrics, it might still be a step too far. In the meantime, many attempts are being made to expand trustworthy AI with, for example, explainable deep neural nets (xDNN; e.g., Angelov & Soares, 2020). These developments go fast and might offer new opportunities for the application of AI to psychometrics as well.

References

- Abbasi, M. M., & Beltiukov, A. P. (2019). Summarizing emotions from text using Plutchik's wheel of emotions. In N. Yusupova, G. Shakhmametova, K. Mironov, & L. Galimova (Eds.), *Proceedings of the 7th scientific conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019): Vol. 166. Advances in intelligent systems research* (pp. 291–294). Atlantis Press. <https://doi.org/10.2991/itids-19.2019.52>
- Abdurrahim, S. H., Samad, S. A., & Huddin, A. B. (2018). Review on the effects of age, gender, and race demographics on automatic face recognition. *The Visual Computer*, 34(11), 1617–1630. <https://doi.org/10.1007/s00371-017-1428-z>
- Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2), 270–301. <https://doi.org/10.1177/1094428112470848>
- Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Networks*, 130(1), 185–194. <https://doi.org/10.1016/j.neunet.2020.07.010>
- Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t-tests: The power of alternatives and recommendations. *Psychological Methods*, 19(3), 409. <https://doi.org/10.1037/met0000014>
- Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3), 1–33. <https://doi.org/10.1145/3444690>
- CRAN, R. (2021). The R project for statistical computing. <http://www.r-project.org>
- Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. Basic Books. <https://psycnet.apa.org/record/1997-08434-000>
- De Groot, A. D. (2019). *Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen* [Methodology: Foundations of research and thinking in the behavioral sciences]. De Gruyter Mouton. <https://doi.org/10.1515/9783110875621>
- Dinga, R., Penninx, B. W., Veltman, D. J., Schmaal, L., & Marquand, A. F. (2019). *Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines*. bioRxiv, (p. 743138). <https://doi.org/10.1101/743138>
- Dolmans, T. C., Poel, M., Van't Klooster, J. W. J., & Veldkamp, B. P. (2020). Perceived mental workload classification using intermediate fusion multimodal deep learning. *Frontiers in Human Neuroscience*, 14(1), 609066. <https://doi.org/10.3389/fnhum.2020.609096>
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

- Escalante, H. J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güllü, U., . . . & van Lier, R. (Eds.). (2018). *Explainable and interpretable models in computer vision and machine learning*. Cham: Springer International Publishing.
- Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated machine learning: Methods, systems, challenges* (pp. 3–33). Springer. https://doi.org/10.1007/978-3-030-05318-5_1
- Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1(6), 261–262. <https://philpapers.org/archive/FLOETR.pdf>
- Floridi, L., Cows, J., Beltracchi, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology* (Vol. 52, pp. 139–183). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning*. Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- 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. <https://doi.org/10.1016/j.psychres.2012.01.032>
- He, Q., Veldkamp, B. P., Glas, C. A., & Van Den Berg, S. M. (2019). Combining text mining of long constructed responses and item-based measures: A hybrid test design to screen for posttraumatic stress disorder (PTSD). *Frontiers in Psychology*, 10(1), 2358. <https://doi.org/10.3389/fpsyg.2019.02358>
- Hu, S., Xiong, J., Fu, P., Qiao, L., Tan, J., Jin, L., & Tang, K. (2017). Signatures of personality on dense 3D facial images. *Scientific Reports*, 7(1), 1–10. <https://doi.org/10.1038/s41598-017-00071-5>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer. <https://doi.org/10.1007/978-1-4614-7138-7>
- Kachur, A., Osin, E., Davydov, D., Shutilov, K., & Novokshonov, A. (2020). Assessing the Big Five personality traits using real-life static facial images. *Scientific Reports*, 10(1), 1–11. <https://doi.org/10.1038/s41598-020-65358-6>
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–73). American Council on Education/Praeger Publishers.
- 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>
- Keszler, N. S. (2021). *Automatic personality prediction based on facial features: Race, gender, and age bias* [Unpublished bachelor thesis, University of Twente]. http://essay.utwente.nl/86496/1/Keszler_BA_BMS.pdf
- Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: A review. *EURASIP Journal on Wireless Communications and Networking*, 2017(1), 1–12. <https://doi.org/10.1186/s13638-017-0993-1>
- Liem, C. C. S., Langer, M., Demetriou, A., Hiemstra, A. M. F., Sukma Wicaksana, A., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 197–253). Springer. https://doi.org/10.1007/978-3-319-98131-4_9
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- MATLAB. (2021). *MATLAB* (Version R2021a) [Computer Software]. The MathWorks Inc.

- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107–135. <https://doi.org/10.1177/01466210122031957>
- Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities, and challenges. *Briefings in Bioinformatics*, 19(6), 1236–1246. <https://doi.org/10.1093/bib/bbx044>
- Neumann, M., Niessen, A. S. M., Tendeiro, J. N., & Meijer, R. R. (2021). The autonomy-validity dilemma in mechanical prediction procedures: The quest for a compromise. *Journal of Behavioral Decision Making* (Advance online publication). <https://doi.org/10.1002/bdm.2270>
- O’Neil, C. (2016). *Weapons of math destruction*. Crown Books.
- OpenCV. (2020). *Open source computer vision library*. <https://github.com/opencv/opencv>
- Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning, and health systems. *Journal of Global Health*, 8(2), 1–8. <https://doi.org/10.7189/jogh.08.020303>
- Rawal, G., Yadav, S., & Kumar, R. (2017). Post-intensive care syndrome: An overview. *Journal of Translational Internal Medicine*, 5(2), 90–92. <https://sciendo.com/pdf/10.1515/jtim-2016-0016>
- Röber, T. E. (2021). *Automated personality prediction based on facial features* [Unpublished master thesis, University of Utrecht].
- Schleicher, A. (2019). *PISA 2018: Insights and interpretations*. OECD Publishing. <https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach’s alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Sijtsma, K., & Meijer, R. R. (2006). Nonparametric item response theory and special topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 719–746). Elsevier. [https://doi.org/10.1016/S0169-7161\(06\)26022-X](https://doi.org/10.1016/S0169-7161(06)26022-X)
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. SAGE.
- Sijtsma, K., & van der Ark, L. A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research*, 64(2), 128–136. <https://doi.org/10.1097/NNR.0000000000000077>
- Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, 31(2), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287–308. <https://doi.org/10.1007/s11336-006-1478-z>
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace. <https://www.python.org>
- Veldkamp, B. P. (2018). *Mastering the data mass* [Inaugural address]. University of Twente. https://research.utwente.nl/files/28106874/oratie_Bernard_Veldkamp.pdf
- Veldkamp, B., Schildkamp, K., Keijsers, M., Visscher, A., & de Jong, T. (2021). *Big Data Analytics in Education: Big Challenges and Big Opportunities*. International Perspectives on School Settings, Education Policy and Digital Strategies: A Transatlantic Discourse in Education Research, 266.
- Voigt, P., & Von dem Bussche, A. (2017). *The EU general data protection regulation (GDPR): A practical guide*. Springer. <https://doi.org/10.1007/978-3-319-57959-7>
- Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018, e7068349. <https://doi.org/10.1155/2018/7068349>
- Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8–36. <https://doi.org/10.1109/MGRS.2017.2762307>

Chapter 5

Psychological Constructs as Organizing Principles



Denny Borsboom

Abstract Klaas Sijtsma has suggested that psychological constructs, such as those invoked in the study of intelligence, personality, and psychopathology, should be understood as *organizing principles* with respect to elements of behavior, including item response behavior. In a discussion in the journal *Psychometrika*, Sijtsma (*Psychometrika*, 71(3), 451–455 (2006)) contrasted this position with the *common cause* interpretation of Item Response Theory (IRT) models and the associated theory of validity that I had articulated some years earlier (Borsboom, *Psychological Review*, 111(4), 1061–1071 (2004)), arguing that this theory of validity was far too strong given the immature status of psychological constructs. In the present chapter, I present an alternative understanding of IRT models in terms of psychometric networks, which is inspired by Sijtsma’s idea of constructs as organizing principles. From the weak premise that psychological constructs organize behaviors, in the sense of identifying behavioral elements that structurally hang together, in the present chapter, I show how one can build up a psychometric approach that can motivate and guide the use of tests in psychology in the absence of strong common cause interpretations.

5.1 Introduction

Psychometrics is an intrinsically multidisciplinary project, and like all multidisciplinary projects, it tends to disintegrate into unconnected monodisciplinary components if left to its own devices. Klaas Sijtsma is one among a small group of psychometricians who have spent their careers trying to protect the brittle but essential connections between substance, mathematics, and philosophy. In this respect, Klaas and I are kindred spirits, because both of us have tried to find a balance between the messy reality of psychometric practice, the idealized structures

D. Borsboom (✉)

Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands
e-mail: D.Borsboom@uva.nl

of psychometric modeling, and the conceptual questions of what psychological measurement is and how it should be optimized.

Despite these shared commitments, in the past, Sijtsma and I also have defended different positions on the core question of how psychometrics should relate to its neighbors. Sijtsma has argued that psychometrics should operate as an auxiliary discipline to psychology, i.e., it should seek a partnership in which it plays the role of helper (Sijtsma 2006). I have tended to take a more directive position, primarily because I doubt that psychologists are sufficiently interested in measurement to tackle the problems involved (Borsboom 2006; Borsboom et al. 2004).

Unfortunately, the theoretical basis required for the research program I championed (Borsboom et al. 2004) is often unattainable in psychology, as Sijtsma (2006) astutely observed, because standard measurement models in psychometrics are unrealistic given the substance matter of psychology. In recent work, however, alternatives to standard measurement models have been developed that seem to align much more naturally to the way that psychologists think; in these models, constructs are not seen as common causes of manifest variable, but as network structures that connect such variables (Borsboom et al. 2021). It turns out that these models are actually finely tuned to a comment that Sijtsma (2006) made in discussion we had in *Psychometrika*, in which he presented the viewpoint that psychological constructs should operate as “organizing principles” that specify which psychometric items “hang together.”

In this chapter, I aim to bring this idea of Sijtsma (2006) in contact with the field of network psychometrics, which has been recently developed on the basis of the network perspective on psychometric constructs (Marsman et al. 2018; Borsboom et al. 2021; Van Borkulo et al. 2014; Cramer et al. 2010) to arrive at an alternative conceptualization of psychometrics in the context of network models. I first review the standard interpretation of latent variables as common causes, after which I discuss an alternative interpretation in terms of structurally connected variables. Finally, I examine the important psychological concepts of unidimensionality, reliability, and validity from this viewpoint.

5.2 Item Response Theory and Common Cause Structures

Item Response Theory (IRT) models the response of a person i to an item j as a function of a set of item and person parameters through an Item Response Function (IRF) that maps each combination of the parameters to a probability distribution over the item responses. In the case that there is only one person parameter θ_i , we have a unidimensional model. A commonly used example of such a model is the well-known Rasch (1960) model, in which the IRF is logistic and each item has one

parameter, β_j , which controls the location of the IRF:

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}} \quad (5.1)$$

Because of its ease of application, mathematical tractability, and favorable measurement properties, the Rasch model is popular among psychometricians. It is heavily used in fields like educational testing, intelligence and personality research, and the study of psychopathology. The model will therefore serve well as a leading example in the current chapter.

Looking at the Rasch model, it is evident that the item response probabilities are the result of a trade-off function between the item and person parameters, which are often called “difficulty” and “ability,” reflecting the origin of the model in educational measurement. This trade-off is possible because θ and β are on the same scale, which means that “difficulty” and “ability” are, in an important sense, exchangeable: “having a higher level of ability” is equivalent to “making an easier set of items,” not just in a figurative mode of speech, but exactly. The fact that all of the IRFs that describe a set of items are controlled by a single person parameter then means that each of the item difficulties trades off against the same ability. This, in turn, suggests that θ functions as a *common cause* of the item responses (Reichenbach 1956; Pearl 2009; Haig 2005a,b).

It is useful to briefly consider the notion of a common cause, as introduced by Reichenbach (1956), to establish this parallel. Reichenbach (1956) dealt with the situation in which a binary common cause, C , has two binary events A and B as its effects. In this case, a common cause is required to satisfy three conditions: (1) $P(A|C) > P(A|\neg C)$ and $P(B|C) > P(B|\neg C)$, (2) $P(A \cap B) > P(A)P(B)$, and (3) $P(A \cap B|C) = P(A|C)P(B|C)$. A classic example considers the relation between yellow-stained fingers (A) and lung cancer (B) as a function of smoking (C): the probability of both yellow-stained fingers and lung cancer is increased, given smoking (condition 1) yellow-stained fingers and lung cancer are positively associated (condition 2), and smoking “screens off” the association between yellow-stained fingers and lung cancer, rendering them conditionally independent (condition 3).

Translating this to a situation with m dichotomous effect variables X_j , $j = 1, \dots, m$ and a continuous common cause θ , as would match most IRT models, Reichenbach’s conditions become:

1. $P(X_j = 1|\theta)$ is increasing in θ .
2. $P(X_j = 1, X_k = 1) > P(X_j = 1)P(X_k = 1)$ for all j, k .
3. $P(x_1, \dots, x_j, \dots, x_m|\theta) = \prod_{j=1}^m P(x_j|\theta) = \prod_{j=1}^m P(X_j = 1|\theta)^{x_j} P(X_j = 0|\theta)^{1-x_j}$.

Condition 1 is satisfied in the Rasch model, as the logistic function (1) is strictly increasing in θ . Condition 2, positive association, is a well-known consequence of every unidimensional monotone latent variable model (Holland & Rosenbaum

1986) including that of Rasch. Condition 3 is local independence, a common property of IRT models, including that of Rasch. Thus, the Rasch model conforms to a common cause structure.

In fact, conditions 1–3 are satisfied in all unidimensional models for dichotomous item responses that have increasing IRFs, like the popular model of Birnbaum (1968). In less restrictive models, like the Mokken (1971) nonparametric model and its generalization, the monotone latent variable model (Holland & Rosenbaum 1986; Junker & Sijtsma 2001), a weaker form of monotonicity (i.e., that $P(X_j = 1|\theta)$ is non-decreasing in θ) exists that does not strictly conform to these conditions; however, in such models, the latent variable can be conceived of as the common cause of subsets of item responses, in those regions of θ where the corresponding items' IRFs are all increasing. Thus, Reichenbach's (1956) common cause structure applies to the relation between θ and the item responses in a broad class of IRT models.

This appears to be more than a statistical coincidence, because several other psychometric concepts have strong parallels with the causal modeling literature as well. For instance, in a measurement context, it is sensible to require that θ mediates the effects of a set of external factors $\{V\}$ on the set of items $\{X\}$. That is, if $\{X\}$ measures θ , then changes in the item response probabilities induced by conditioning on group variables (e.g., sex) or interventions (e.g., therapy) should affect the item responses only indirectly, that is, through θ . In causal terms, this means that θ should “block” all causal paths from variables in $\{V\}$ to variables in $\{X\}$. Via the criterion of *d-separation* (Pearl 2009), this implies the following conditional independence relation for all variables in $\{X, V\}$:

$$F(x|\theta) = F(x|\theta, v), \quad (5.2)$$

for all (θ, v) , where $F(x|\theta, v)$ denotes the value of the conditional distribution function of X evaluated at the point (θ, v) . In the psychometric literature, (2) is well known as the requirement of *measurement invariance* (Mellenbergh 1989; Meredith 1993; Millsap 2007). Interpreted causally, measurement invariance thus requires that no variables except for θ exert a direct causal effect on the item responses.

The idea that θ acts as a common cause of the item responses also matches the way many substantive researchers think about latent variables. Spearman (1904) set up the common factor model to analyze cognitive tests in accordance with this notion, as he interpreted general intelligence, or *g*, as a source of individual differences present in a wide range of cognitive tests (see also Jensen (1999), for a similar view); the condition of *vanishing tetrads* that Spearman introduced as a model test is currently seen as one of the hallmark conditions of the common cause model (Bollen & Ting 1993). In personality research, putative latent variables such as those in the Five-Factor Model are likewise seen as causes of behaviors; for instance, McCrae and Costa Jr. (2008) argue such things as “E[xtroversion] causes party-going” (p. 288). Finally, in clinical psychology, Reise and Waller (2009) note that “to model item responses to a clinical instrument [with IRT], a researcher must

first assume that the item covariation is caused by a continuous latent variable” (p. 26).

Thus, not just the “letter” (i.e., the formal correspondence given above) but also the “spirit” of latent variable modeling is driven by the idea that our item responses are the effects of a common attribute that underlies the observations, represented in the model structure by the symbol θ . As Reise and Waller (2009) note, this “sets limits on the type of constructs that can be appropriately modeled by IRT” (p. 26); namely, the type of constructs for which this is sensible is the type for which, minimally, it can be expected that the items will behave as if they are a function of a common cause.

5.3 The Causal Account of Test Validity

The common cause understanding of latent variable models is strong but clear. In 2004, I developed a straightforward consequence of the causal interpretation of measurement models for the concept of validity (Borsboom et al. 2004). My reasoning was that, *if* psychological constructs like depression or intelligence signify common cause of test scores, *and* validity refers to the question of whether these test scores measure what they should measure, *then* the core of any validity argument must lie in specifying the psychological processes by which the relevant psychological attributes play their causal role. This idea applies naturally for certain test types; an example may involve items as used in working memory capacity tests. In these tests, participants are instructed to recall different sequences of letters or numbers, while they are simultaneously executing another task (e.g., counting back from 100 to 0). Plausibly, one’s success in recalling the sequence 2, 6, 4, 7, 2 and the sequence 4, 6, 3, 8, 9, 4, 3, 4, 5 depend on the same resource, namely, working memory capacity. Clearly, then, working memory capacity acts as a common cause with respect to the individual differences in item responses.

This type of causal argument says how individual differences in a psychological attribute, which affects all of the item responses, are translated into individual differences in test scores. In my view, this forms the core of the validity concept. If one thinks about it, such specifications are not hard to come by in cases where questions of validity actually have a definite answer. Such examples, in my view, are too scarcely considered in validity theory. In fact, the idea that validity questions are unanswerable is taken for granted in certain lines of thinking about validity (one received view is that “validity is a never-ending process”). However, there are actually measurement problems that have been solved and validity questions that have been answered. And typically, the answer to a question like “why does instrument X measure attribute Y ?” hinges on a specification of *how the instrument works* (i.e., specifies a causal process where the measured attribute is the starting point and the meter readings are the endpoint). Why do mercury thermometers measure temperature? Because higher temperatures cause the mercury to expand and hence the meter rises. Why does the composition of air trapped in the Arctic

ice measure historical global carbon dioxide emissions? Because higher emissions cause more carbon dioxide in the relevant air pockets and higher concentrations of carbon dioxide cause higher readings in spectral analysis of the air contained in these pockets. Why does the item “what is your age?” measure age? Because people know how old they are and, if willing, will be able to supply that information.

If available, the causal answer to validity questions is the most forceful answer there is. It is explicit, testable, and suggestive of changes that might improve the measurement device. However, it is also a very taxing answer. It requires a convincing account of how the measured attributes exert their causal effects, and theories that can motivate such accounts are scarce in psychology (although they do exist, as some of the above examples show).

In 2004, I believed that this type of analyses could be made to work in psychology at large and should be investigated vigorously. Our task as psychometricians, in my view, was to come up with good analyses of response behavior in which the measured attribute played a causal role. It looked like that kind of analysis was there for the taking with the combination of advanced modeling techniques, cognitive diagnostic models, and good psychological theory. However, some colleagues were skeptical. Klaas Sijtsma was one of them (Sijtsma 2006). In response to a paper in which I pushed the causal psychometric account to its extreme (Borsboom 2006), he articulated doubts with respect to the research program I was advocating:

Borsboom’s assumption about the ontology and causality of psychological attributes seems to lead to a very restrictive conception of the process of construct validation: Elegant in its rigor but impractical for psychology (and many others areas). It seems to me that we still know so little about the functioning of the human brain in general and cognitive processes including those underlying personality traits and attitudes in particular, that it is difficult even to say what an ‘attribute’ is. In the absence of such knowledge, I prefer to consider psychological attributes as organizational principles with respect to behavior. Thus, my point of view is that psychological attributes define which behaviors hang together well and are useful to the degree in which tests sampling these behaviors play a role in predicting interesting psychological phenomena.

With some reluctance, I have to admit defeat to this charge when it comes to the more abstract entities in the psychometric pantheon—that is, the big psychometric players like general intelligence, neuroticism, attitudes, and psychopathological conditions. In the years that followed the conceptual articulation of the causal validity program, I attempted to come up with good measurement theories for such constructs but ultimately failed to provide a believable analysis in causal terms. Although this research line of mine is undocumented and impossible to replicate—a failure to construct conceptual analyses leads to the theoretical equivalent of a file-drawer problem; one can hardly publish failures to come up with a new theory—I did try hard. Apart from a few isolated successes (most notably the analysis of IRT model results in terms of drift diffusion parameters as developed by my colleague Han van der Maas (Van der Maas et al. 2011)), it just didn’t work.¹

¹ Naturally, that I could not come up with good theories of test validity does not mean that nobody else could. Perhaps I didn’t use the right framework; perhaps I just approached the problem from

In fact, that is an understatement. If one attempts to specify how general intelligence causes responses to the item “Who wrote the Iliad?”, how depression leads to sad mood, and how attitudes influence the answer to questions like “do you think Trump is a good leader?”, one arrives at theories that are far too strong and far too simplistic. In fact, the very idea that traits like intelligence, extraversion, and psychopathological syndromes are causes of human behavior, including the behavior that involves ticking boxes on questionnaires, appears to be rather far-fetched, more akin to Moliere’s *virtus dormitiva* than to any serious appreciation of the psychological complexity of the constructs in question.² B.F. Skinner (1987) once stated that “as soon as you have formed the noun *ability* from the adjective *able*, you are in trouble,” and indeed that seems to be accurate for many of the abilities and traits invoked in psychometric theory.

5.4 Structural Connections

The general failure to come up with adequate measurement theories forms an interesting contrast with the relative ease with which one can concoct psychometric models. Taking desirable measurement properties as axiomatic for measurement models, it is possible to deduce the general form and structure that psychometric models *should* have and work out the distributions of data they imply. This is what, in my view, psychometricians have been most successful at over the course of the past century. One can easily imagine the tests and theories employed in psychology today to become a laughing stock for future generations, but the intricate building of interrelated statistical measurement models of IRT, which Klaas Sijtsma and others erected in the past decades, will remain an important entry in the scientific record.

Because such models have more to do with philosophical ideas on what good measurements should look like, than with psychological ideas about whatever it is we are measuring, psychometric models are in my view best seen as applied philosophy of science. The models one can deduce from general philosophical measurement desiderata range from very strong to extremely weak. The Rasch model in Eq. 5.1 is an example of a strong model. Rasch (1960) started from some desirable measurement axioms (e.g., things that would be nice to have, like separate identifiability of person and item characteristics) and then deduced the

the wrong angle; may others come and do it better. However, as they say, insanity is trying the same thing over and over again and expecting different results, so it seemed more sensible to reconceptualize my problems than to keep trying.

² As an aside, if test score use and interpretation would actually require theories of this kind, then the whole scientific project of psychometrics would be in serious trouble, perhaps even trouble of the end-of-story kind. Realizing this, in hindsight, it is unsurprising that the reception of my validity theory was mixed. One influential validity theorist stated informally that what I said might all be good and true, but that my definition of validity would never be accepted because theories that specify how psychological constructs cause item scores “would not hold up in court.”

model formula in Eq. 5.1 as a consequence. One can also proceed from much weaker requirements and deduce weaker models as a result (Holland & Rosenbaum 1986; Ellis & Junker 1997); this more realistic approach is the cornerstone of nonparametric IRT, to which Junker and Sijtsma (2001) and Sijtsma and Molenaar (2002) provide excellent introductions.

The focus on desirable measurement properties leads to simple models. The Rasch model in Eq. 5.1 is one example, but basically all models in the IRT family (Mellenbergh 1994) are variants of the general structure. Usually, that structure specifies how people's position on a relatively simple latent variable (e.g., a point on a continuous line, membership of a latent category) is coordinated with a specific probability distribution over the item responses. Because nearly all models specify a form of conditional independence, in which the observed variables are independent given the latent variable, they can typically be understood along the lines of Reichenbach (1956) as explained in the previous paragraph. Thus, nearly all models can be understood as specifying a (possibly somewhat convoluted) common cause model.

However, if we think for a moment about, say, relations between symptoms of depression, attitude items, or cognitive processes, it is hard to see how causal interpretations of such simple models could possibly be on target. After all, it would be a small miracle if human behavior, embedded in a nexus of complex interactions between factors at genetic, physiological, psychological, and social levels, were literally governed by a model structure as simple as Eq. (5.1) and its relatives.

This realization, however, presents us with a paradox. This is because the latent variable modeling approach in topics, like intelligence, personality, and psychopathology, has *not* fared as badly as one should expect, given the complexity of human behavior. Although measurement models rarely fit adequately, they do generally provide a reasonable description of the data; for instance, the fact that the general factor of intelligence is now in the company of general factors of personality and psychopathology is not accidental. In recent years, I have investigated the hypothesis that the reason for this is that the tests used in such domains depend on distinct attributes and processes that do *not* depend on a common cause, but *are* structurally connected through relations that can reasonably be approximated by pairwise interactions; these pairwise interactions, in turn, generate probability distributions that tend to fit latent variable models reasonably well.

What does it mean for variables to be structurally connected? To preempt some obvious misinterpretations, let me first say what I do not mean. First, I do *not* mean to say that structurally connected variables merely correlate. Ice cream consumption and murder rates are famously correlated across the months of the year, but not structurally connected. Second, to be structurally connected does not necessarily mean that variables stand in directed causal relations. Sad mood and suicidal ideation, for instance, are probably to some extent involved in some reciprocal reinforcement process, but it is unlikely that this relation is of the smoking-causes-lung-cancer kind that modern theories of causality (e.g., Pearl (2009)) present as axiomatic. In addition, I intentionally cover cases where different items are in part related through semantic or logical pathways. For example, some items in

personality questionnaires contain very similar wordings, which leads responses to be structurally connected, but the queried attributes are unlikely to stand in directed causal relations.

As a working definition, I propose variables to be structurally connected if they (or their probability distributions) cannot vary independently. This definition is extremely broad and covers a wide variety of cases where relations between variables are systematic (i.e., they are not merely correlated) but not necessarily causally directed. For variables to be structurally connected thus means that these variables represent elements of behavior that, in the words of Sijtsma (2006), “hang together.”

Here are some examples. Responses to the item “do you think Trump is a good leader?” are structurally connected with responses to the item “do you like Trump?” because people strive to keep their attitude elements consistent. Responses to the item “do you like parties?” are structurally connected with responses to the item “did you like the last party you went to?” because the latter assesses a memory trace that a respondent will also use in answering the former. Responses to the item “have you felt fatigued over the past 2 weeks?” are structurally connected with responses to the item “have you slept more than usual over the past 2 weeks?” because people who are tired will tend to sleep more. Responses to the item “have you felt fatigued over the past 2 weeks?” are *also* structurally connected with responses to the item “have you slept less than usual over the past 2 weeks?” because people who don’t sleep well tend to get tired. In each of these cases, the relevant variables cannot vary independently, because they share meaning, are causally related, share resources, or are intertwined in development.³ In contrast, responses to the item “do you like parties?” are not structurally connected with responses to the item “who wrote the Iliad?”, because these variables can vary independently. For the same reason, responses to the item “have you felt fatigued over the past 2 weeks?” are not structurally connected with responses to the item “do you think Trump is a good leader?”.

5.5 Network Representations of Psychological Constructs

Shifting attention from a common cause principle to the idea of structural connections between variables invites a different way of setting up our basic psychometric apparatus. I propose to denote the structural connection between two variables X_j and X_k with a tilde:

$$\text{Structural connection} \equiv X_j \sim X_k \quad (5.3)$$

³ In a very weak interpretation of causality, one could say structural connections are a type of causal relations, but I think this stretches the meaning of the term beyond the limits of usefulness.

$X_j \sim X_k$ means that the variables in question cannot vary independently. One way of making this idea more precise might be taken by saying that intervening on X_j will affect X_k and vice versa, i.e., implying a bidirectional causal relation between the variables in question that can be expressed using the concept of a Do-operator (Pearl 2009). The Do-operator is used in the causality literature to represent interventions on a system in order to provide a semantics for causal relations. In particular, a causal effect of X_j on X_k would be expressed as $P(X_k|\text{Do}(X_j = x_j)) \neq P(X_k)$, i.e., a causal effect means that the probability distribution of X_k is not the same under manipulations that force X_j to take different values x_j . In the present case, one could imagine that a structural connection may be taken to imply bidirectional causal dependence:

$$X_j \sim X_k \Rightarrow P(X_j|\text{Do}(X_k = x_k)) \neq P(X_j) \wedge P(X_k|\text{Do}(X_j = x_j)) \neq P(X_k) \quad (5.4)$$

This type of characterization in causal terms may be useful to flesh out specific formalizations of structural dependence.⁴ For instance, given the causality calculus, the causal formulation implies the statistical consequence that two variables cannot be rendered statistically independent, given any other variable at our disposal. Thus, given a set of variables $\{X\}$ that characterize a system under study, if a structural connection exists between X_j and X_k , this implies that when conditioning on the complement set $\{X_c\}$ (all variables in $\{X\}$ excluding X_j and X_k):

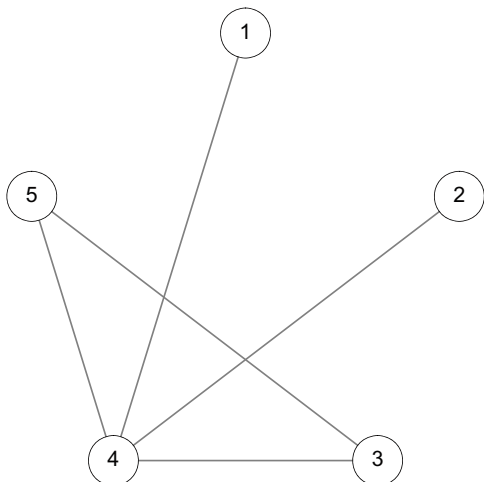
$$X_j \not\perp\!\!\!\perp X_k | X_c \quad (5.5)$$

In other words, the variables are not statistically independent given everything else we can measure on the system. Ordinarily, the set $\{X\}$ will be a pragmatically chosen collection of variables, and the question of whether any two variables are structurally connected is studied relative to this collection. It would be interesting to investigate what other choices would be sensible to define the set $\{X\}$ or what it means for $\{X\}$ to characterize the system under study, but I will not pursue these questions here and will simply assume $\{X\}$ to be composed of whatever a researcher chooses to include in the data. Also, for convenience, I will assume the bidirectional causal relation to be symmetric (i.e., equally strong in both directions) although I don't think much hinges on that.

For a given set of variables, the above definitions imply a network structure to which standard representations of network psychometrics apply. In particular,

⁴ I hasten to add that the above characterization should be seen as one of the various ways to make the idea of a structural connection concrete and not as definitional. Also I do not intend the notion of structural connection to require such things as decomposability (de Boer et al. 2021) and similar kinds of atomistic conception of the world of variables, which seems to have become common in the language of causality. Thus, the implication is not biconditional as the causal analysis does not exhaust the possibilities and may depend on auxiliary assumptions that are not satisfied in psychometrics.

Fig. 5.1 A network of five variables. Edges between variables indicate that the relevant variables are structurally connected



Eq. 5.5 defines a Pairwise Markov Random Field (PMRF), which has the attractive graphical representation as a network in which variables that are not directly connected are conditionally independent given the other variables. For binary variables, the PMRF can be estimated in various ways, for instance, through the R-package *IsingFit* (Van Borkulo et al. 2014).

Figure 5.1 provides an example network. Variables are represented as nodes, and structural connections as edges. The set of nodes that is connected to node j is known as the *neighborhood* of j and denoted N_j . We assume that the probability distribution of the variables has the Markov property, i.e., that it factorizes according to the graph structure. This implies that the joint probability distribution can be represented as log-linear model that includes main effects for all variables and pairwise interactions for any two variables that are connected in the graph. However, for my current purposes, it is more convenient to think of the model in terms of a set of logistic regressions, where each node is regressed only on the variables in its neighborhood:

$$\text{logit}(X_j) = \alpha_j + \sum_{k \in N_j} \beta_{jk} X_k \quad (5.6)$$

This formulation is the model used in the *IsingFit* representation (Van Borkulo et al. 2014). Now let us consider the relation between Eq. 5.6 and the typical IRT representation as in Eq. 5.1. Willem Heiser (personal communication) has observed that this representation connects the network approach to an older tradition in psychometrics, namely, that of image factor analysis (Guttman 1953). Image factor analysis is an approach to factor analysis that explicitly aims to avoid the use of latent variables. In image factor analysis, the regression of a variable on all other variables in the data creates the variable’s image (the weighted sumscores formed by the regression), while the residual of that regression defines its anti-image. The

IsingFit representation of the network model could be seen as an extension of the image factor analysis model to the dichotomous case; in this interpretation, the regression model defines the image of X_j (Guttman 1953).

5.6 Reinterpreting Psychometric Concepts

In the IRT model, the items are related to a single person parameter, and the regression parameter for that function depends only on the item considered. In the logistic regression formulation of the network model, the items are related to a set of independent variables, and the regression parameters are different for each of them. In the IRT formulation, the predictor is latent. In the logistic regression formulation, it is observed.

However, there are also similarities. In both cases, we see a generalized regression with a parameter that depends on the item (the intercept in the logistic regression, the difficulty parameter in the IRT model) and a regression parameter that controls the slope of the regression of the item on the predictor term. That predictor, in the IRT model, is the latent variable. In the logistic regression, it is a set of scores on the neighboring items. These scores are weighted by regression weights. We can imagine collecting the combined effects of all predictors in a weighted sumscore of the variables in the item j 's neighborhood, which for person i we may denote as

$$N_{ij}^+ = \sum_{k \in N_j} \beta_{jk} X_{ik} \quad (5.7)$$

Now things start to look quite analogous if we express person i 's expected score as a function of the latent variable model,

$$P(X_{ij} = 1 | \theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}, \quad (5.8)$$

and as a function of the regression model. We can make this similarity most apparent by putting the regression in the same form as the IRT model through suitable transformations of parameters, representing the model in terms of a trade-off between the internal field (the effects of the other nodes in the network) and the external field (e.g., in case the regression coefficients equal unity, this would directly correspond to the intercept parameter in Eq. 5.6 transformed to $\alpha_j^* = -\alpha_j$):

$$P(X_{ij} = 1 | N_{ij}^+, \alpha_j^*) = \frac{e^{N_{ij}^+ - \alpha_j^*}}{1 + e^{N_{ij}^+ - \alpha_j^*}} \quad (5.9)$$

Using this representation, we see that the neighborhood score N_{ij}^+ plays a role that is analogous to that of the latent variable θ in the IRT model, while the intercept of the regression α_j^* is the analogue of item difficulty in IRT. Via the concept of structural connection, one can think of any specific item as standing under the influence of the variables in its neighborhood, not in the sense that its value is directly caused by these, as in a billiard ball causation picture, but in the sense that the item's probability distribution cannot change independent of that of its neighbors. In a nontrivial sense, therefore, the item *measures* the influence of its neighbors: *ceteris paribus*, the more neighbors of item j are positive (take value $X = 1$), the more j will tend to be positive as well.⁵

The relation between the latent variable in IRT and the neighborhood score in network analysis in the dichotomous case mirrors the relation between latent variables in factor analysis and components in image factor analysis for the continuous case (Guttman 1953); also, the centrality measure of predictability that has been proposed in the network literature (Haslbeck & Waldorp 2018) is highly similar to the index of determination discussed in Guttman (1953). Finally, note that the dimensionality of the neighborhood scores is the same as that of the data (i.e., there are as many neighborhood scores as variables); a reduction of these neighborhood scores could be achieved through, for instance, a principal component analysis, which would compress the neighborhood scores into a smaller dimensionality. In the case where the network is fully connected, one would then expect the neighborhood scores to approximate unidimensionality, while a sparsely connected network would not.

Although the alignment between IRT and network models that I have constructed here is not as mathematically elegant as those used in the direct equivalence proofs between multidimensional IRT and Ising models that are now in the literature (Marsman et al. 2018; Epskamp et al. 2018), the logistic regression of an item on a neighborhood score has intuitive appeal and facilitates reinterpretation of psychometric concepts. This is because we can keep in mind the analogy between the latent variable and the neighborhood score. Substituting the concept of a neighborhood score in a network of structural connections for the concept of a common cause of item responses leads to several straightforward consequences for psychometric practice. In the following, I review some of the most important psychometric concepts from this point of view.

⁵ The relation between the item and the targeted latent variable is typically represented in an Item Characteristic Curve (ICC). Of course, we can do the same in the network model, if we put N_{ij}^+ on the x-axis and the probability of a positive item response on the y-axis; we may call this curve a Network Response Function (NRF). The items will have different neighborhoods, which means the NRFs have different domains, but the general concept clearly is similar.

5.6.1 *Unidimensionality*

The notion of unidimensionality plays a very important role in psychometrics. It encodes the idea that the correlations between item responses can be represented as a function of a single dimension. In the parlance of IRT, unidimensionality means that the logits of the expected scores of the items (the true item scores) are perfectly correlated, which means that if one knows a person's true score on one item, one cannot learn anything new about the ordering of the persons on the latent variable by consulting the other items. In terms of the causal interpretation of measurement models, this represents the hypothesis that the different items trade off against precisely the same ability.

Networks in general are unlikely to satisfy such requirements, but they can approximate them (and often do). This works as follows. If one looks at Fig. 5.1, it is clear that the variables have very different neighborhoods. Node 1 only has one neighbor (node 4), while node 4 has four (nodes 1, 2, 3, and 5). Clearly, in this case, the covariance matrix will depart from unidimensionality significantly. However, if one imagines an ever more densely connected network, one can see that the neighborhoods of different nodes will overlap more and more. Thus, the neighborhood scores of different items will get more and more correlated. In a perfectly connected network, the neighborhoods of any two nodes will differ by only one term (the scores on the evaluated nodes themselves, which are not part of their own neighborhood). Thus, the closer the network approaches perfect connectivity, the closer it will get to unidimensionality. In the network literature, this means that the network can be approximated by the so-called mean field approximation, which essentially substitutes a single number for all of the node neighborhoods (Finnemann et al. 2021). In a nontrivial sense, the latent variable in a unidimensional psychometric model corresponds to the mean field in a network model, which in turn is strongly related to the first factor of an image factor analysis (Guttman 1953).

One can also see that, as the network gets larger, the neighborhoods get ever more close. I conjecture that this, in effect, realizes the same process that Ellis and Junker (1997) describe through the concept of a tail measure. A tail measure is the equivalent of a sumscore on an infinite item domain, which Ellis and Junker (1997) showed is an adequate interpretation of a latent trait. Similarly, I suggest that an infinitely large network will produce equivalent tail measures on items' neighborhood scores, as in the limit all neighborhoods will coincide in terms of their ordering of persons. Thus, from a network perspective, unidimensionality can be interpreted as a measure of network homogeneity. Interestingly, a perfectly connected network with invariant edge weights (a so-called Curie-Weiss model) turns out to be statistically equivalent to the Rasch model (Marsman et al. 2018).

5.6.2 Reliability

It would not be much of an overstatement to say that Classical Test Theory (CTT) was invented to furnish a basis for the notion of reliability: the degree to which true scores are linearly predictable from observed scores. The most important estimator of reliability, Cronbach's α , is controversial in psychometrics, both because of misinterpretations of the concept and because it is statistically inferior to other estimators (Sijtsma 2009). However, it is probably also the most important quantity psychometrics has delivered, as it regulates the composition and size of item sets used in practical test applications.

Reliability is commonly seen as a property of a test.⁶ That is, it is a measurement concept, which indicates to what extent the total test score contains "measurement error." However, it is a well-kept secret among psychometricians that the noise in our test scores is rarely identifiable as measurement error independent of the psychometric model. Typically, what we call measurement error is simply variance that simply cannot be explained from the latent variable model (for whatever reason). Why this unexplained variance should be interpreted as measurement error is rarely explicated.

Interestingly, in the network representation, the psychometric representation of sumscore reliability is not (only) a measurement concept. Even if all items are measured without error, the network may still leave variance unexplained, for instance, because the items do not hang together perfectly (i.e., there is wiggle room for individual items given the other items) or because the network is not fully connected. This may very well be a property of a *construct* rather than of the *measurement instrument*. Indeed, Dalege and van der Mass (2020) hold that implicit measures of attitudes are necessarily unreliable because in the situation where people do not attend to the attitude, the attitude network operates in a high entropy regime (i.e., the network is weakly connected).

What *does* reliability imply, from a network perspective? In my view, high reliability means that the state of the individual items is highly predictable from the neighborhood scores. That is, the network has a low entropy (Dalege & van der Maas 2020), because the structural connections between items are strong so that items tend to align. Interestingly, low entropy implies that more extreme sumscores will become more prevalent, which will lead to higher variance of the sumscore. Thus, from a network perspective, the ratio of the sum of the item variances to the total test score variance—a standard operationalization of reliability—is actually a measure of how strongly connected the network is.

⁶ This is fundamentally mistaken because, even on its own terms, CTT represents reliability as a test \times population interaction (Mellenbergh 1996), but I will ignore this here and assume the population given.

5.6.3 *Validity*

As I noted earlier, in the past, I have articulated and defended the idea that validity is a causal concept, which hinges on the degree to which the measured attribute (represented as a latent variable) influences the item scores. Clearly, in the network representation, there is no latent variable (except as a mathematical representation of the joint probability distribution of the network; Epskamp et al. (2018)). Hence, the causal interpretation of validity is not on offer for the network as a whole. However, that conception can still be operational for the individual items in a network, for instance, if one asks whether the depression item “have you slept less than usual over the past 2 weeks?” actually measures insomnia (Cramer et al. 2010). In addition, if different items depend on a variable that is not represented in the network (i.e., a latent variable), then a latent variable model can be used to analyze that part of the network (e.g., in a latent network model; see Epskamp et al. (2017)), and in this case, the latent variable can be conceptualized as a common cause, which renders the causal account of validity applicable.

But what can one say about the validity of a test if the items in that test in fact measure properties that are structurally connected, rather than a single latent attribute? If the network model is true, then the construct label (e.g., “depression,” “intelligence,” “neuroticism”) does not refer to such a latent attribute but to the network as a whole. Thus, when we ask “does this depression questionnaire actually measure depression?”, the question should be understood as “do the variables assessed through the items included in this questionnaire actually correspond to the nodes in the depression network?”. This, in turn, leads to the question “which nodes are part of the depression network?”. And it is here, I submit, that the psychometric construct fulfills its function as an *organizing principle*. A construct label such as “depression” does not designate a latent attribute targeted in the measurement procedure, but instead indicates a family of variables that are structurally connected to produce the coordinated behavior of the network as a whole that we phenomenologically recognize as the overall state of individuals we are interested in.

Thus, the organizing principle of psychological constructs involves a simple but important task: to identify which nodes should be part of the network. In the special case that items are questions (rather than observations of behavior or other modes of investigating the human system, such as brain states or genetic profiles), this means that psychological constructs fulfill their main function in the area traditionally referred to as *content validity*. This is ironic, because in the literature on validity theory, content validity is typically seen as an outdated concept, if not an inferior one (Guion 1980). If the combination of Sijtsma’s “hanging together” and network psychometrics is in the right ball park, content validity may thus well see a revival.

5.7 Discussion

In the present chapter, I have offered a reinterpretation of standard psychometric concepts in terms of a network perspective, in which item responses are viewed as structurally connected components in a network. This perspective aligns remarkably well with the idea that item responses merely “hang together” (Sijtsma 2006). In the presented scheme, the role of the psychological construct is radically shifted: the construct label does not designate a latent variable that acts as a common cause with respect to the item responses, but a set of relevant properties that are structurally connected. The primary task of the construct theory, so understood, is to indicate *which* of the many potentially relevant properties actually is part of the psychometric network, i.e., is part of the set of structurally connected variables.

This is quite a different way of thinking about the function of construct theories, but it seems to fit psychological practice quite well. Whenever I proposed to substantive psychologists that their theories should provide information with respect to the question of how a latent attribute determined the responses to questionnaire items, they looked at me as if they witnessed water burning. However, most of these same psychologists will have little problems in identifying why certain items should be included in a test. Usually, their answers either implicitly or explicitly explain how the items tap attributes that hang together systematically. The reasons behind these connections can vary wildly from area to area, so they cannot be uniformly fleshed out. However, in many cases, the connections in question suggest that variables bear a connection that is stronger than mere association and weaker than directional causation. I have tried to capture this notion in the term “structural connection.”

My exploration of the mathematical conceptualization and the theoretical consequences of this idea has been preliminary. Especially the connection to the work by Ellis and Junker (1997) seems to harbor some interesting secrets that I have not developed here. In particular, because in a positive manifold that is consistent with a unidimensional factor model, pairwise conditional associations are always weaker than unconditional ones (Van Bork et al. 2018), it seems that in such cases the size of networks is limited by the strength of the structural connections that they consist of. That is, mathematically, a set of items that realizes an item domain can grow without bound (in fact, this is required for the proofs in Ellis and Junker (1997)). But a fully connected network like the Curie-Weiss model discussed in Marsman et al. (2018) cannot grow without bound, unless the conditional associations in the network get ever smaller in the process. It seems to me that this will not always be attainable. In other words, sets of items that cover a fully connected network may have a limited size. This would induce the notion of a construct that features a finite item domain which, to my knowledge, has not yet been developed in psychometrics.

As noted in the introduction to this chapter, psychometricians have many alliances. Their models, while cast in the language of mathematics, have an important connection to substantive realms (e.g., psychology, education, etc.) as well as to conceptual ideas about the nature of measurement. These alliances often clash. What

is desirable from the point of view of measurement theory (e.g., additivity of the model, separability of parameters, simplicity, and parsimony) is often substantively speaking unrealistic. On the other hand, processes that are relevant from a substantive point of view (e.g., in terms of cognitive processes involved in psychometric tests) often lead to theoretical models that are mathematically intractable and that do not respect the strictures that the occupants of measurement theory's ivory tower proscribe as normative. The challenge is therefore to find conceptions of psychometric constructs that have a natural representation as mathematical structures, so that they can play the essential role of connecting psychological theory to empirical observation—the cardinal purpose of measurement. Latent variables are one such conception and network structures another. However, it would be idle to think that the possibilities are exhausted by these representations, and I hope that future psychometricians will come up with many others, so that our discipline will remain a vibrant and developing one that honors the *psycho* in psychometrics.

Acknowledgments I would like to thank Willem Heiser and Riet van Bork for valuable comments on an earlier draft of this chapter. This work was supported by NWO Vici grant no. 181.029.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.
- Bollen, K. A. & Ting, K. (1993). Confirmatory tetrad analysis. *Sociological Methodology*, 23, 147–175.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425–440.
- Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, 1(1). <http://dx.doi.org/10.1038/s43586-021-00055-w>.
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>.
- Cramer, A. O., Waldorp, L. J., Van Der Maas, H. L., & Borsboom, D. (2010). *Comorbidity: A network perspective*. <https://doi.org/10.1017/S0140525X09991567>.
- Dalege, J. & van der Maas, H. L. J. (2020). Accurate by being noisy: A formal network model of implicit measures of attitudes. *Social Cognition*, 38(Supplement), s26–s41. <https://doi.org/10.1521/soco.2020.38.suppl.s26>.
- de Boer, N. S., de Bruin, L. C., Geurts, J. J. G., & Glas, G. (2021). The network theory of psychiatric disorders: A critical assessment of the inclusion of environmental factors. *Frontiers in Psychology*, 12. <https://www.frontiersin.org/article/10.3389/fpsyg.2021.623970>.
- Ellis, J. L. & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, 62, 495–523.
- Epskamp, S., Rhemtulla, M. T., & Borsboom, D. (2017). *Generalized network psychometrics: combining network and latent variable models*. *Psychometrika*. <https://doi.org/10.1007/s11336-017-9557>

- Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network psychometrics. In P. Irwing, Hughes, D., & T. Booth (Eds.), *The wiley handbook of psychometric testing*. New York: Wiley.
- Finnemann, A., Borsboom, D., Epskamp, S., & Maas, H. L. J. van der. (2021). The theoretical and statistical ising model: A practical guide in R. *Psych*, 3(4), 594–618. <https://www.mdpi.com/2624-8611/3/4/39>.
- Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11, 385–398.
- Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, 18(4), 277–296. <https://doi.org/10.1007/BF02289264>.
- Haig, B. D. (2005a). An abductive theory of scientific method. *Psychological Methods*, 10(4), 371–388. <https://doi.org/10.1037/1082-989X.10.4.371>.
- Haig, B. D. (2005b). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, 40(3), 303–329. https://doi.org/10.1207/s15327906mbr4003_2.
- Haslbeck, J. M. B. & Waldorp, L. J. (2018). How well do network models predict observations? On the importance of predictability in network models. *Behavior Research Methods*, 50(2), 853–861. <https://doi.org/10.3758/s13428-017-0910-x>.
- Holland, P. W. & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics*, 14, 1523–1543.
- Jensen, A. R. (1999). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Junker, B. W. & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272. <https://doi.org/10.1177/01466210122032064>.
- Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., Bork, R. van, Waldorp, L. J., et al. (2018). An introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research*. <https://doi.org/10.1080/00273171.2017.1379379>.
- McCrae, R. R. & Costa Jr., T. J. C. P. (2008). Empirical and theoretical status of the Five-Factor Model of personality traits. In G. M. Boyle & D. Saklofske (Eds.), *G* (pp. 273–294). Los Angeles: Sage.
- Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13, 127–143.
- Mellenbergh, G. J. (1994). Generalized Linear Item Response Theory. *Psychological Bulletin*, 115, 300–307.
- Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, 1, 293.
- Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525–543.
- Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, 72, 461.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.
- Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.
- Reichenbach, H. (1956). *The direction of time*. Los Angeles: The University of California Press. <https://doi.org/ppe>.
- Reise, S. P. & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5(1), 27–48. <https://doi.org/10.1146/annurev.clinpsy.032408.153553>.
- Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika*, 71(3), 451–455. <https://doi.org/10.1007/s11336-006-1497-9>.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>.
- Sijtsma, K. & Molenaar, I. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks: SAGE Publications Ltd. <https://doi.org/https://methods.sagepub.com/book/introduction-to-nonparametric-item-response-theory>
- Skinner, B. F. (1987). Whatever happened to psychology as the science of behavior? *American Psychologist*, 42(8), 780–786. <https://doi.org/10.1037/0003-066X.42.8.780>.

- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.
- Van Bork, R., Grasman, R. P. P. P., & Waldorp, L. J. (2018). Unidimensional factor models imply weaker partial correlations than zero-order correlations. *Psychometrika*, *83*(2), 443–452. <https://doi.org/10.1007/s11336-018-9607-z>.
- Van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., et al. (2014). A new method for constructing networks from binary data. *Scientific Reports*, *4*(1), 5918. <https://doi.org/10.1038/srep05918>.
- Van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339–356. <https://doi.org/10.1037/a0022749>.

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 licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence 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
Factor Analysis and Classical Test Theory

Chapter 6

A New Expression and Interpretation of Coefficient Omega Under the Congeneric One-Factor Model



David J. Hessen

Abstract A new expression for the communality of the total score under the one-factor model is presented. In general, the communality of the total score is a lower bound to the reliability of the total score. Under the one-factor model, the communality of the total score also assesses the validity of the total score as a measure of the common factor. Conditions are given under which the new expression equals coefficient alpha. Furthermore, new expressions for the communality of an arbitrary item score and the proportion of total variance explained are derived under the one-factor model. For all new communality expressions, closed-form distribution-free estimates are provided. In an example, the closed-form estimates are calculated for a classic data set.

6.1 Introduction

Coefficient alpha (Guttman 1945; Cronbach 1951) is a very popular lower bound to the reliability of the total score (the unweighted sum of the item scores). Sijsma (2009) criticized the use of coefficient alpha for assessing the reliability of the total score and recommended the use of greater lower bounds, such as the greatest lower bound (Woodhouse and Jackson 1977; ten Berge et al. 1981) and coefficient lambda-2 (Guttman 1945). Despite the existence of greater lower bounds to the reliability of the total score, coefficient alpha continues to be used in practice. For an overview of many other lower bounds to the reliability of the total score, see Revelle and Zinbarg (2009).

It has been shown that coefficient alpha equals the communality of the total score if the item scores follow the essentially tau-equivalent model (Bentler 2009). The essentially tau-equivalent model is the unrealistic special case of the one-factor model in which all item scores have the same factor loading (Lord and

D. J. Hessen (✉)
Utrecht University, Utrecht, The Netherlands
e-mail: d.j.hessen@uu.nl

Novick 1968). Under the essentially tau-equivalent model, coefficient alpha is only equal to the reliability of the total score if all unique factors only contain random measurement error. However, under the essentially tau-equivalent model, the communality of the total score equals the proportion of variance of the total score explained by the common factor. This means that under the essentially tau-equivalent model, coefficient alpha assesses the validity of the total score as a measure of the common factor.

A more realistic model for the measurement of a single factor by a set of items than the essentially tau-equivalent model is the one-factor model (Spearman 1950). Under the one-factor model, factor loadings are not restricted to be equal. Since, in practice, the items of a subtest are usually constructed to measure one and the same latent factor, the item scores of a subtest are often assumed to follow the one-factor model. Under the one-factor model, the communality of the total score is given by coefficient omega (Heise and Bohrnstedt 1970; McDonald 1978) and equals the proportion of variance of the total score explained by the common factor. So under the one-factor model, coefficient omega assesses the validity of the total score as a measure of the common factor.

In this chapter, a new expression for the communality of the total score under the one-factor model is presented. Whereas coefficient omega expresses the communality of the total score in terms of factor model parameters, this new expression is in terms of the variances of the item scores, the covariances between the item scores, and the number of item scores. It is shown that the new expression equals coefficient alpha if the item scores follow the essentially tau-equivalent model. Furthermore, new expressions of the communality of an arbitrary item score and the proportion of total variance explained are derived under the one-factor model. Since all new expressions are functions of the population variances of the item scores and the population covariances between the item scores, distribution-free closed-form estimates are obtained by replacing the population parameters with sample analogues.

First, however, the one-factor model is briefly outlined in the next section. Subsequently, the new communality expressions and their closed-form estimates are presented. Finally, the closed-form estimates of the new communality expressions are calculated for a classic example data set.

6.2 The One-Factor Model

Let the random variables X_1, X_2, \dots, X_J be J item scores for a randomly selected individual from a population. The means of X_1, X_2, \dots, X_J are denoted by b_1, b_2, \dots, b_J ; the variances are denoted by $\sigma_1^2, \sigma_2^2, \dots, \sigma_J^2$; and the covariance between two arbitrary item scores X_j and X_k is denoted by σ_{jk} , for all j and $k \neq j$. In the one-factor model, it is assumed that

$$X_j = b_j + a_j\xi + U_j, \quad \text{for all } j, \quad (6.1)$$

where a_j is a constant factor loading, for all j , ξ is the common factor, and U_j is a unique factor, for all j . Note that $U_j = S_j + E_j$, where S_j is an item-specific factor (only varying between persons) and E_j is random measurement error (varying between persons and within persons), so that $T_j = b_j + a_j\xi + S_j$ is the item true score. The common factor ξ is assumed to be independent of all unique factors U_1, U_2, \dots, U_J . The unique factors are assumed to be mutually independent. To identify the model, the variance of ξ is set to one. Let $var(U_j) = \delta_j$, for all j . Then, it follows that $\sigma_j^2 = a_j^2 + \delta_j$, for all j , and $\sigma_{jk} = a_j a_k$, for all j and $k \neq j$. Usually, the items are constructed such that it can be assumed that $a_j > 0$, for all j . If the items are not constructed this way, a transformation can be applied to some of the item scores such that it can be assumed that $a_j > 0$, for all j . Note that if $a_j > 0$, for all j , then $\sigma_{jk} > 0$, for all j and $k \neq j$.

6.3 Communality/Validity

Let $C_j = b_j + a_j\xi$, for all i . Then, the total score is given by $X = \sum_j X_j = C + U$, where $C = \sum_j C_j = \sum_j b_j + \sum_j a_j\xi$ and $U = \sum_j U_j$. The communality of X is given by coefficient omega (Heise and Bohrnstedt 1970; McDonald, 1978) and is the squared correlation between X and C , that is,

$$\omega = \rho_{XC}^2 = \frac{var(C)}{\sigma_X^2} = \frac{\left(\sum_j a_j\right)^2}{\sum_j \sigma_j^2 + \sum_j \sum_{k \neq j} \sigma_{jk}} = \frac{\sum_j a_j^2 + \sum_j \sum_{k \neq j} a_j a_k}{\sum_j a_j^2 + \sum_j \delta_j + \sum_j \sum_{k \neq j} a_j a_k}, \quad (6.2)$$

where $\sigma_X^2 = var(X)$. Note that under the one-factor model, ρ_{XC}^2 is equal to the squared correlation between X and ξ given by

$$\rho_{X\xi}^2 = \frac{\{cov(X, \xi)\}^2}{\sigma_X^2} = \frac{[E\{(\sum_j a_j\xi + \sum_j U_j)\xi\}]^2}{\sigma_X^2} = \frac{(\sum_j a_j)^2}{\sigma_X^2}. \quad (6.3)$$

From this, it can be concluded that under the one-factor model, the communality coefficient ρ_{XC}^2 also assesses the validity of X as a measure of ξ . Now, since $\sigma_{jk} = a_j a_k$, for all $j \neq k$, it follows that

$$\frac{\sigma_{jk}\sigma_{jl}}{\sigma_{kl}} = \frac{a_j a_k a_j a_l}{a_k a_l} = a_j^2, \quad \text{for all } j, k \neq j, \text{ and } l \neq j, k.$$

Taking the average over all $k \neq j$ and $l \neq j, k$ gives

$$\frac{1}{(J-1)(J-2)} \sum_{k \neq j} \sum_{l \neq j, k} \frac{\sigma_{jk}\sigma_{jl}}{\sigma_{kl}} = a_j^2, \quad \text{for all } j. \quad (6.4)$$

Substitution from Eq. 6.4 and $\sigma_{jk} = a_j a_k$ into

$$\rho_{XC}^2 = \frac{\sum_j a_j^2 + \sum_j \sum_{k \neq j} a_j a_k}{\sigma_X^2} \quad (6.5)$$

yields the new expression of the communality of the total score X given by

$$\rho_{XC}^2 = \left\{ \frac{1}{(J-1)(J-2)} \sum_j \sum_{k \neq j} \sum_{l \neq j, k} \frac{\sigma_{jk} \sigma_{jl}}{\sigma_{kl}} + \sum_j \sum_{k \neq j} \sigma_{jk} \right\} / \sigma_X^2. \quad (6.6)$$

Note that substitution from Eq. 6.4 into $\sigma_j^2 = a_j^2 + \delta_j$ and solving for δ_j yields

$$\delta_j = \sigma_j^2 - \frac{1}{(J-1)(J-2)} \sum_{k \neq j} \sum_{l \neq j, k} \frac{\sigma_{jk} \sigma_{jl}}{\sigma_{kl}}, \quad \text{for all } j. \quad (6.7)$$

Also note that if $a_j > 0$, for all i , then it follows from Eq. 6.4 that

$$a_j = \sqrt{\frac{1}{(J-1)(J-2)} \sum_{k \neq j} \sum_{l \neq j, k} \frac{\sigma_{jk} \sigma_{jl}}{\sigma_{kl}}}, \quad \text{for all } j. \quad (6.8)$$

Under the essentially tau-equivalent model, $\sigma_{jk} = a^2$, for all j and $k \neq j$, so that

$$\sum_{l \neq j, k} \frac{\sigma_{jk} \sigma_{jl}}{\sigma_{kl}} = (J-2) \sigma_{jk}, \quad \text{for all } j \text{ and } k \neq j. \quad (6.9)$$

Substitution from Eq. 6.9 into Eq. 6.6 and factoring $\sum_j \sum_{k \neq j} \sigma_{jk}$ yields coefficient alpha given by

$$\alpha = \frac{J}{J-1} \sum_j \sum_{k \neq j} \sigma_{jk} / \sigma_X^2. \quad (6.10)$$

In addition to the communality of the total score, the communalities of the individual item scores might be of interest in practice. The communality of item score X_j is defined as the squared correlation between X_j and C_j . Under the one-factor model, the communality of item score X_j is given by

$$h_j^2 = \rho_{X_j C_j}^2 = \frac{\text{cov}(X_j, C_j)^2}{\sigma_j^2 \text{var}(C_j)} = \frac{\text{var}(C_j)}{\sigma_j^2} = \frac{a_j^2}{a_j^2 + \delta_j}, \quad \text{for all } j. \quad (6.11)$$

Note that under the one-factor model, $\rho_{X_j C_j}^2$ is equal to $\rho_{X_j \xi}^2$. So under the one-factor model, the communality of item score X_j also assesses the validity of X_j as

a measure of ξ . Now, dividing the left-hand side of Eq. 6.4 by σ_j^2 yields

$$h_j^2 = \frac{1}{(J-1)(J-2)\sigma_j^2} \sum_{k \neq j} \sum_{l \neq j, k} \frac{\sigma_{jk}\sigma_{jl}}{\sigma_{kl}}, \quad \text{for all } j. \quad (6.12)$$

The total variance is defined as $\sum_j \sigma_j^2$. Under the one-factor model, the proportion of total variance explained by the common factor is given by

$$\pi = \frac{\sum_j a_j^2}{\sum_j a_j^2 + \sum_j \delta_j} = \frac{\sum_j a_j^2}{\sum_j \sigma_j^2} \quad (6.13)$$

and assesses the extent to which the items measure the common factor relative to the unique factors. Substitution from Eq. 6.4 into Eq. 6.13 yields

$$\pi = \frac{1}{(J-1)(J-2)} \sum_j \sum_{k \neq j} \sum_{l \neq j, k} \frac{\sigma_{jk}\sigma_{jl}}{\sigma_{kl}} / \sum_j \sigma_j^2. \quad (6.14)$$

6.3.1 Estimates

Let x_{ij} be the observed score of individual $i = 1, 2, \dots, N$ on item $j = 1, 2, \dots, J$. The sample mean score on item j is given by $\bar{x}_j = \sum_{i=1}^N x_{ij}/N$, for all j . The observed total score of individual i is then given by $x_i = \sum_{j=1}^J x_{ij}$, for all i , and the sample mean total score is then given by $\bar{x} = \sum_{i=1}^N x_i/N$. A closed-form estimate of ρ_{XC}^2 is now given by

$$\hat{\rho}_{XC}^2 = \left\{ \frac{1}{(J-1)(J-2)} \sum_j \sum_{k \neq j} \sum_{l \neq j, k} \frac{s_{jk}s_{jl}}{s_{kl}} + \sum_j \sum_{k \neq j} s_{jk} \right\} / s^2, \quad (6.15)$$

where $s_{jk} = \sum_{i=1}^N (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)/(N-1)$ is the estimate of σ_{jk} , for all j and $k \neq j$, and $s^2 = \sum_{i=1}^N (x_i - \bar{x})^2/(N-1)$ is the estimate of σ_X^2 . Note that $s^2 = \sum_j s_j^2 + \sum_j \sum_{k \neq j} s_{jk}$, where $s_j^2 = \sum_{i=1}^N (x_{ij} - \bar{x}_j)^2/(N-1)$ is the estimate of σ_j^2 , for all j . A closed-form estimate of δ_j is given by

$$\hat{\delta}_j = s_j^2 - \frac{1}{(J-1)(J-2)} \sum_{k \neq j} \sum_{l \neq j, k} \frac{s_{jk}s_{jl}}{s_{kl}}, \quad \text{for all } j. \quad (6.16)$$

If $s_{jk} > 0$, for all $j \neq k$, then a closed-form estimate of factor loading a_j is given by

$$\hat{a}_j = \sqrt{\frac{1}{(J-1)(J-2)} \sum_{k \neq j} \sum_{l \neq j, k} \frac{s_{jk}s_{jl}}{s_{kl}}}, \quad \text{for all } j. \quad (6.17)$$

A closed-form estimate of the item communality $\rho_{X_j C_j}^2 = h_j^2$ is given by

$$\hat{h}_j^2 = \frac{1}{(J-1)(J-2)s_j^2} \sum_{k \neq j} \sum_{l \neq j,k} \frac{s_{jk}s_{jl}}{s_{kl}}, \text{ for all } j. \quad (6.18)$$

Finally, a closed-form estimate of the proportion of total variance explained by the common factor is given by

$$\hat{\pi} = \frac{1}{(J-1)(J-2)} \sum_j \sum_{k \neq j} \sum_{l \neq j,k} \frac{s_{jk}s_{jl}}{s_{kl}} / \sum_j s_j^2. \quad (6.19)$$

6.4 An Example

The data in this example are taken from Lord and Novick (1968, p. 91) and are the entries of the sample covariance matrix of four measures of English as a foreign language. The sample covariance matrix is based upon a sample size of 1416 and is given by

$$\begin{bmatrix} s_1^2 & & & \\ s_{21} & s_2^2 & & \\ s_{31} & s_{32} & s_3^2 & \\ s_{41} & s_{42} & s_{43} & s_4^2 \end{bmatrix} = \begin{bmatrix} 94.7 & & & \\ 87.3 & 212.0 & & \\ 63.9 & 138.7 & 160.5 & \\ 58.4 & 128.2 & 109.8 & 115.4 \end{bmatrix}.$$

Estimates of the parameters of the one-factor model are often obtained by maximum likelihood estimation under the assumption of multivariate normality of the item scores in the population. For comparison, both the maximum likelihood estimates and the closed-form estimates of a_j , δ_j , and h_j^2 , for all j , and ω and π are calculated. The maximum likelihood estimates are denoted by \tilde{a}_j and $\tilde{\delta}_j \tilde{h}_j^2$, for all j , and $\tilde{\omega}$ and $\tilde{\pi}$. The estimates for all item parameters and coefficients are given in Table 6.1. Note that the item order given by the closed-form estimates $\hat{h}_1^2 < \hat{h}_3^2 < \hat{h}_4^2 < \hat{h}_2^2$ is different from the item order given by the maximum likelihood estimates $\tilde{h}_1^2 < \tilde{h}_3^2 < \tilde{h}_2^2 < \tilde{h}_4^2$. The estimate of coefficient α is $\hat{\alpha} = .891$. The maximum likelihood estimate of coefficient ω is $\tilde{\omega} = .909$, and its closed-form estimate is $\hat{\omega} = .912$. So, about 91% of the sample variance of the total score is explained by the common factor. The maximum likelihood estimate of the total variance π is $\tilde{\pi} = .725$, and its closed-form estimate is $\hat{\pi} = .735$. So, about 73% of the total sample variance is explained by the common factor.

Table 6.1 Estimates of all item parameters and coefficients under the one-factor model, for the Lord and Novick (1968) example data

j	Estimate					
	\hat{a}_j	\tilde{a}_j	$\hat{\delta}_j$	$\tilde{\delta}_j$	\hat{h}_j^2	\tilde{h}_j^2
1	6.164	6.133	56.708	57.020	.401	.397
2	13.455	12.939	30.975	44.432	.854	.790
3	10.653	10.882	47.015	41.963	.707	.738
4	9.791	9.937	19.534	16.581	.831	.856

6.5 Conclusion

Coefficient alpha has traditionally been used to assess the reliability of the total score. Since coefficient alpha equals the communality of the total score under the essentially tau-equivalent model, coefficient alpha is a lower bound to the reliability of the total score. The communality of the total score under the more realistic one-factor model is also a lower bound to the reliability of the total score. Under the one-factor model, however, the communality of the total score equals the proportion of variance of the total score explained by the common factor and therefore assesses the extent to which the common factor is measured by the total score. If items are constructed to measure one and the same latent factor, then the one-factor model can be used to study whether the items actually measure a single common factor. Once it has been concluded that the items measure a single common factor, it is of interest to assess how well the single common factor is measured by the item scores or the total score. To assess how well the single common factor is measured by the total score, coefficient omega and its new expression can be used. Under the one-factor model, coefficient omega and its new expression give the proportion of variance of the total score explained by the common factor. In practice, the maximum likelihood estimate of coefficient omega is often used as the estimate of the communality of the total score under the one-factor model. The closed-form estimate of omega now provides a distribution-free alternative.

References

- Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, 74(1), 137–143. <https://doi.org/10.1007/s11336-008-9100-1>.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–282. <https://doi.org/10.1007/BF02288892>.
- Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. In E. F. Borgatta (Ed.), *Sociological methodology* (pp. 104–129). Jossey-Bass.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- McDonald, R. P. (1978). Generalizability in factorable domains: “domain validity and generalizability”: 1. *Educational and Psychological Measurement*, 38(1), 75–79. <https://doi.org/10.1177/001316447803800111>.

- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. <https://doi.org/10.1007/s11336-008-9102-z>.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>.
- Spearman, C. (1950). *Human ability*. Macmillan.
- ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to reliability and constrained minimum trace factor analysis. *Psychometrika*, *46*(2), 201–213. <https://doi.org/10.1007/BF02293900>.
- Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika*, *42*(4), 579–591. <https://doi.org/10.1007/BF02295980>.

Chapter 7

A Factor Analysis Approach to Item Level Change Score Reliability



Dylan Molenaar

Abstract Reliability of change scores from a pretest-posttest design is important to establish the usefulness of change scores in drawing inferences about pretest-posttest differences. Besides the traditional sum score-based classical test theory approach, an item level classical test theory approach has been proposed to assess change score reliability. This approach was demonstrated to be superior to the traditional sum score-based approach. However, both the item level and the sum score-based approaches are biased in the case of multidimensionality and correlated errors. Therefore, in this chapter two factor analysis approaches to the item level classical test theory approach are presented. These approaches treat the item level data explicitly as ordinal and allow various psychometric aspects of the data to be investigated including multidimensionality, carry-over effects, and response shifts. As a result, using the factor analysis approaches, it can be assessed whether the results from the classical test theory approaches can be trusted. The classical test theory approaches and factor analysis approaches are studied in a simulation and applied to a real dataset pertaining to life satisfaction.

The pretest-posttest design is an important scientific research tool to study change. For instance, the effectiveness of an intervention (a newly developed psychotherapy, social skills training, teaching method, etc.) can be studied by comparing the differences in pre-intervention measurements and post-intervention measurements to those obtained from a placebo sample in which no intervention, a bogus intervention, and/or an existing intervention is applied. Besides scientific applications, the pretest-posttest design has its uses in, for instance, education where children complete an achievement test at the start and at the end of the school year to see how they progress. In addition, in large-scale online environments (e.g., websites, monitoring systems), a pretest-posttest design is used to test for the effectiveness of

D. Molenaar (✉)

Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands
e-mail: D.Molenaar@uva.nl

small differences in the online environment (e.g., the font of the text, or the ordering of the items) on some measure of interest (e.g., “time spent in the application”).

Of key importance in making inferences about the (size of the) effect of the intervention is the reliability of the difference between the pretest and posttest scores. That is, in unreliable difference scores, it is unlikely to find any statistically significant effects, and, more importantly, unreliable difference scores are unlikely to be of practical significance as they hardly tell something about the underlying change. The use of difference scores in general has been criticized by a number of authors (e.g., Cronbach & Furby, 1970; Linn & Slinde, 1977; Lord, 1963; Williams & Kaufmann, 2012). One of the arguments is that change score reliability cannot be large if the intervention is effective. That is, if the intervention is effective, a large pretest-posttest correlation arises due to most subjects improving on the posttest measurement which in turn results in a small reliability of the difference scores (Linn & Slinde, 1977). Gu et al. (2018) pointed out, however, that this critique, and four other common critiques on change scores, can be refuted. Among others, they argue that criticism on change score reliability is often based on problematic assumptions (e.g., that the test score variance is equal for the pretest and posttest) and on a confusion of measurement precision with test score reliability (see also Mellenbergh, 1996). One of the main conclusions by Gu et al. is that change score reliability depends on the data characteristics of a given application and is therefore an empirical question.

Addressing the empirical question of change score reliability has commonly been done using approaches from classical test theory (Lord & Novick, 1968). That is, by decomposing both the pretest scores and the posttest scores into a true score component and an error score component, an expression for the reliability of the difference between these scores can be derived (e.g., Cronbach & Furby, 1970). Such an approach, however, focusses on the summed item scores of the pretest and posttest, thereby aggregating over individual items. To this end, Gu et al. (2021) presented a related classical test theory approach based on the item level pretest and posttest data. From a simulation study, it appeared that the item level approach outperforms the traditional sum score approach in terms of the bias of the change score reliability estimates. That is, in the case of correlated errors and/or multidimensionality in the data, the item level approach was less biased than the sum score approach. For the item level approach itself, reliability estimates were unbiased in the case of unidimensional data and in the absence of correlated errors. For datasets with stronger multidimensional structures and data with stronger correlations among the errors, bias increased, but not as much as for the sum score approach.

In the present chapter, a factor analysis approach to the item level classical test theory approach by Gu et al. (2021) is proposed. Advantage of such an approach is that it naturally adds model fit assessment tools to the methodology of Gu et al. These tools can be used to verify unidimensionality of the difference scores. In addition, an extension is proposed in which correlated errors can be identified and in which the presence of a so-called response shift can be assessed (Howard & Dailey, 1979). If a response shift occurs, the pretest-posttest difference is (partly) due to

a change in the interpretation of the construct by the subjects and can thus not be interpreted in terms of change on the underlying construct (Sprangers & Schwartz, 1999; see also Oort et al., 2009). In practice, if a response shift occurs, this results in violations of measurement invariance (Meredith, 1993) across the pretest and posttest which can be detected in a factor analysis framework.

The outline is as follows: first the sum score and item level classical test theory approaches are discussed. Then, two factor analysis approaches are derived, one closely resembling the approach by Gu et al. (2021) and the other being the extended model. Next, the new approaches are studied in a small simulation study and applied to a real dataset pertaining to life satisfaction to illustrate their use in practice. This chapter is concluded with a general discussion.

7.1 Classical Test Theory Approaches to Difference Score Reliability

7.1.1 The Sum Score Approach

To introduce the sum score approach, let $X^{(\text{pre})} = \sum_{j=1}^J X_j^{(\text{pre})}$ and $X^{(\text{post})} = \sum_{j=1}^J X_j^{(\text{post})}$ denote the summed item scores over, respectively, the J pretest items, $X_j^{(\text{pre})}$, and the J posttest items, $X_j^{(\text{post})}$. Then, the difference scores, $D = X^{(\text{post})} - X^{(\text{pre})}$, can be submitted to a classical test theory decomposition, that is,

$$D = T_D + E_D \quad (7.1)$$

where T_D is the true difference score and E_D is the measurement error. Note that all classical test theory definitions apply, that is, for a given subject v ,

$$T_{D,v} = E(D_v) \quad (7.2)$$

and

$$E_{D,v} = D_v - T_{D,v}. \quad (7.3)$$

Therefore, reliability of the difference score can be expressed as

$$\rho_{DD'} = \frac{\sigma_{T_D}^2}{\sigma_D^2} \quad (7.4)$$

where $\sigma_{T_D}^2$ is the true difference score variance and σ_D^2 is the observed difference score variance. In practice, the true difference score is not observed, precluding

calculation of the difference score reliability using the above. However, under the assumption that the measurement errors of the pretest and posttest are independent, it can be derived that (see, e.g., Cronbach & Furby, 1970)

$$\rho_{DD'} = \frac{\rho_{pre,pre'}\sigma_{pre}^2 + \rho_{post,pre'}\sigma_{post}^2 - 2\sigma_{pre,post}}{\sigma_{pre}^2 + \sigma_{post}^2 - 2\sigma_{pre,post}}, \quad (7.5)$$

where $\rho_{pre,pre'}$ and $\rho_{post,pre'}$ are the reliability of the pretest scores $X^{(pre)}$ and the posttest scores $X^{(post)}$ and σ_{pre}^2 and σ_{post}^2 are the corresponding variances with covariance $\sigma_{pre,post}$. Although, again, this equation contains $\rho_{pre,pre'}$ and $\rho_{post,pre'}$ which cannot be directly estimated, in practice $\rho_{pre,pre'}$ and $\rho_{post,pre'}$ can be replaced by lower bound estimates of the reliability like Cronbach's alpha (Cronbach, 1951). Therefore, Eq. 7.5 above, with a lower bound estimate plugged in for the reliabilities of the pretest and posttest, provides a lower bound to the difference score reliability.

7.1.2 The Item Level Approach

In the above, the basis for estimating the difference score reliability is the summed item scores. As discussed by Gu et al. (2021), using the summed pretest and posttest scores will make the difference score reliability estimate sensitive to the presence of correlated errors and to multidimensionality of the pretest and posttest scores. That is, the items from the pretest and the posttest may be subject to correlated errors, for instance, due to carry-over effects which are common in the pretest-posttest design (McConnel et al., 1998). In the traditional sum score-based approach above, these correlations will be absorbed in the covariance between the pretest scores and the posttest scores, $\sigma_{pre,post}$, which biases the difference score reliability $\rho_{DD'}$. In addition, if the pretest item scores and the posttest item scores are multidimensional, in practice – where lower bound estimates are used – this will also bias $\rho_{DD'}$ via the effect that multidimensionality has on classical test theory reliability estimates $\rho_{pre,pre'}$ and $\rho_{post,post'}$ (Dunn et al., 2014; Sijtsma & Pfadt, 2021).

To address the above two issues, Gu et al. (2021) advocated the use of item difference scores as these are less sensitive to correlated errors and multidimensionality as compared to the summed pretest and posttest scores. Thus, Gu et al. proposed to directly estimate reliability $\rho_{DD'}$ from the item level difference scores, that is,

$$X_j^{(pre)} = T_j^{(pre)} + E_j^{(pre)} \quad (7.6)$$

and

$$X_j^{(post)} = T_j^{(post)} + E_j^{(post)}, \quad (7.7)$$

so that

$$D_j = X_j^{(\text{post})} - X_j^{(\text{pre})} = \left(T_j^{(\text{post})} - T_j^{(\text{pre})} \right) + \left(E_j^{(\text{post})} - E_j^{(\text{pre})} \right) = T_{D_j} + E_{D_j}. \quad (7.8)$$

Thus, the item level difference scores, D_j , are treated as a single test of which the reliability needs to be determined. By doing so, possible correlations between the errors $E_j^{(\text{pre})}$ and $E_j^{(\text{post})}$ are absorbed in $\sigma_{E_{D_j}}^2$ making the procedure less vulnerable to these correlations as compared to the traditional sum score-based classical test theory approach (at least in the situations considered by Gu et al.).¹ In addition, the item level approach above does not depend on the dimensionality of $T_j^{(\text{pre})}$ and $T_j^{(\text{post})}$ but only on the dimensionality of T_{D_j} which Gu et al. argue to be closer to unidimensionality in practice.

To apply the item level method above, a lower bound reliability estimate is needed to be applied to D_j . Gu et al. considered Cronbach's alpha and λ_2 and λ_4 (Guttman, 1945). The obtained estimates are directly interpretable in terms of lower bounds to the difference score reliability, e.g., for λ_2 ,

$$\lambda_2 = \frac{\sum \sum_{j \neq k} \sigma_{D_j D_k} + \sqrt{\frac{J}{J-1} \sum \sum_{j \neq k} \sigma_{D_j D_k}^2}}{\sigma_D^2} \leq \rho_{DD'} \quad (7.9)$$

where $\sigma_{D_j D_k}$ is the covariance between the difference score of item j and k and σ_D^2 is the variance of the summed item difference scores which are identical to D in the sum score approach above. This approach is taken as the point of departure for a factor analysis account of difference score reliability.

7.2 A Factor Analysis Approach to Difference Score Reliability

In this section, two approaches are presented. The first is a direct translation of the classical test theory item level approach by Gu et al. (2021) to a factor model for item differences which can be used to test for the presence of multidimensionality of the difference scores. The second approach is an extended model which accounts for all observed data (i.e., not only the difference scores) and which can be used to

¹ In the traditional classical test theory approach, positive residual correlations will increase $\sigma_{pre, post}$ which will bias reliability downwards. In the item level classical test theory approach, positive residual correlations will decrease $\sigma_{E_{D_j}}^2$ (as $\sigma_{E_{D_j}}^2 = \sigma_{E_j^{(\text{pre})}}^2 + \sigma_{E_j^{(\text{post})}}^2 - 2\sigma_{E_j^{(\text{pre})}, E_j^{(\text{post})}}$) which will bias reliability upwards. This is also what was found in the simulations by Gu et al. The bias was however much larger for the traditional approach.

test for the presence of multidimensionality in the pretest and posttest item scores and for correlated errors and response shifts.

7.2.1 A Factor Model for Item Differences

A direct translation of the classical test theory item level approach into a factor model framework can be obtained by replacing the true score variables, $T_j^{(\text{pre})}$ and $T_j^{(\text{post})}$ in Eqs. 7.6 and 7.7 by the conditional mean in a common factor model, that is,

$$T_j^{(\text{pre})} = E\left(X_j^{(\text{pre})}\right) = b_j + a_j \xi^{(\text{pre})} \quad (7.10)$$

and

$$T_j^{(\text{post})} = E\left(X_j^{(\text{post})}\right) = b_j + a_j \xi^{(\text{post})}. \quad (7.11)$$

where a_j is a factor loading, b_j is an intercept, and $\xi^{(\text{pre})}$ and $\xi^{(\text{post})}$ are the unidimensional latent factors assumed to underlie the pretest and posttest scores, respectively. In addition, the measurement error variables $E_j^{(\text{pre})}$ and $E_j^{(\text{post})}$ from the classical test theory approach are replaced by the factor model residuals, $\delta_j^{(\text{pre})}$ and $\delta_j^{(\text{post})}$ (e.g., Bollen, 1989, p. 218; Sijtsma & Pfadt, 2021). In the resulting model, the item difference variables are then subject to

$$D_j = X_j^{(\text{post})} - X_j^{(\text{pre})} = a_j \left(\xi^{(\text{pre})} - \xi^{(\text{post})} \right) + \left(\delta_j^{(\text{pre})} - \delta_j^{(\text{post})} \right) = a_j \xi_D + \delta_{D_j} \quad (7.12)$$

that is, the model is a one-factor model on the difference scores with factor ξ_D modeling the latent differences between the pretest and posttest and δ_{D_j} with variance $\sigma_{\delta_{D_j}}^2$ containing the differences in residuals across the pretest and posttest. Within this one-factor model, reliability can be calculated using coefficient ω (McDonald, 1978, 1999) which is here denoted $\omega_{\text{difference}}$ to explicitly indicate that the reliability is based on the item level pretest and posttest difference scores D_j from Eq. 7.12 (and not on the raw pretest and posttest scores), that is,

$$\omega_{\text{difference}} = \frac{\left(\sum_{j=1}^J a_j \right)^2 \sigma_{\xi_D}^2}{\left(\sum_{j=1}^J a_j \right)^2 \sigma_{\xi_D}^2 + \sum_{j=1}^J \sigma_{\delta_{D_j}}^2}. \quad (7.13)$$

Note that in the factor analysis approach above, it is assumed that the pretest and posttest scores are unidimensional, while this assumption is not imposed in the

item level approach by Gu et al. (2021). However, as the models for the pretest and posttest scores from Eqs. 7.10 and 7.11 are not fit, this assumption does not necessarily need to hold. The assumption that does need to hold in the actual model in Eq. 7.12 is that the latent difference factor ξ_D is unidimensional, which is a comparable assumption to that of the Gu et al. approach (but not equivalent, see below).

7.2.1.1 Relation to Classical Test Theory

The factor model approach presented here is not equivalent to the classical test theory approach presented earlier. That is, the latent factor in the factor model from Eq. 7.12, ξ_D , and the true score from classical test theory in Eq. 7.8, T_D , are inherently different. That is, ξ_D is a unidimensional latent factor that accounts for the variance common to all item difference scores, D_j , while the true difference score T_D accounts for all systematic sources of variance without modeling its structure (Sijtsma & Pfadt, 2021; Sijtsma & Van der Ark, 2021; chapter 2). As T_D and ξ_D do not necessarily capture the same sources of variation in the data, E_{D_j} and δ_{D_j} are also not equivalent. That is, E_{D_j} represents random measurement error, while δ_{D_j} contains random measurement error and misfit. Therefore, due to these differences, the present approach is only one possible translation of the item level classical test theory approach to a factor model approach. Other possibilities may arise by replacing the true difference scores with different latent structures.

7.2.1.2 Categorical Item Scores

The model above is a linear factor model which can be estimated using maximum likelihood by assuming a multivariate normal distribution for D_j . However, especially at the item level, a normal distribution may not be appropriate for the difference scores as the pretest and posttest item scores are categorical in practice (dichotomous or ordinal). Commonly, variables with five or more ordered categories that are normally distributed can safely be analyzed using a normal linear factor model (Dolan, 1994; Rhemtulla et al., 2012). However, in the case of dichotomous items, the item difference scores will have a three-point scale at most (-1, 0, and 1), and the item difference scores of polytomous items may not be normally distributed.

To account for the ordinal (and possibly non-normal) nature of the difference scores, a distinction is made between D_j^* , which is the theoretical, normally distributed, item difference variable, and D_j which is the actually observed ordinal item difference score variable. To apply the factor model to D_j , it is assumed that the normally distributed D_j^* variable is subject to Eq. 7.12, that is,

$$D_j^* = a_j \xi_D + \delta_{D_j}. \quad (7.14)$$

The observed item difference scores D_j then arise by categorizing the continuous normal difference scores D_j^* at thresholds β_{jc} where $c = 0, \dots, M_j$ where M_j is the number of item difference scores for item j (i.e., the number of levels in D_j). Thus:

$$D_j = c \quad \text{if} \quad \beta_{jc} < D_j^* < \beta_{j_{c+1}} \quad \text{for} \quad c = 0, \dots, M_j \quad (7.15)$$

with $\beta_{j0} = -\infty$ and $\beta_{jM_j} = \infty$. Then, the full model for ordinal difference scores is given by Eqs. 7.14 and 7.15, with parameters: a_j and $\sigma_{\delta D_j}^2$ for all j , $\sigma_{\xi_D}^2$ and μ_{ξ_D} , and the thresholds, β_{jc} , for all j and all c except $c = 0$ and $c = M_j$. Not all parameters are uniquely identified, however. This is discussed below.

Note that coefficient ω in Eq. 7.13 is an estimate of the reliability of the observed item difference scores D_j by using variance and covariance estimates of the underlying normal variables D_j^* (i.e., polychoric variances and covariances). In factor analysis of ordinal variables, using polychoric (co)variances is preferred over assuming D_j to be normally distributed and using its observed (co)variances (Dolan, 1994). As mentioned above, if, in practice, the difference scores appear to be normal with five or more levels, one can assume that $D_j^* = D_j$ and fit the model in Eq. 7.12 directly to the observed difference scores (Dolan, 1994; Rhemtulla et al., 2012). Both options are explored in the simulation study and real data illustration later.

7.2.1.3 Correlated Errors and Response Shifts

Until now, the factor analysis approach addresses one aspect of the item level classical test theory approach by Gu et al. (2021). That is, in the present factor model for difference scores, departures from unidimensionality in the difference scores are straightforwardly detected by consulting model fit statistics. If these statistics indicate poor model fit, this is likely due to multidimensional difference scores, and the results of the model should be interpreted with care. The issue of correlated errors and response shifts is however not addressed in this approach.

Correlated Errors As discussed above, in practice, the errors from the item level classical tests theory may be correlated errors due to, for instance, a carry-over effect. Similarly as in the item level classical test theory approach, correlated errors will go unnoticed in the present model. That is, if the errors in the classical test theory approach are correlated, this will show up as a covariances between the item residuals $\delta_j^{(pre)}$ and $\delta_j^{(post)}$ in the factor model approach. Similarly as in the item level approach by Gu et al. (2021), these covariances will be absorbed in $\sigma_{\delta_D}^2$. This cannot be detected in the model fit, but it will bias the reliability estimates, similarly as discussed for the Gu et al. approach in footnote 1.

Response Shifts In the model in Eqs. 7.12 and 7.15, it is assumed that the measurement characteristics of the factor model are invariant for the pretest and posttest. That is, a_j and b_j are assumed to be the same for the pretest and the posttest. This corresponds to the assumption of measurement invariance (Meredith,

1993). This assumption is a prerequisite for a meaningful comparison of the latent difference factor, ξ_D in terms of the difference between $\xi^{(\text{pre})}$ and $\xi^{(\text{post})}$. In a pretest-posttest design, violations of measurement invariance can occur due to many reasons. For self-report questionnaires, one reason may be the occurrence of a so-called response shift (Oort et al., 2009). A response shift occurs if the posttest measures something psychometrically different from the pretest. That is, subjects may have recalibrated the scale on which they judge themselves (e.g., after the intervention they realize that they were much lower on the construct than they indicated before), they have reprioritized some aspects of the construct (e.g., after the intervention the subjects are appreciating some aspects of the construct more as before), or they have redefined the construct (e.g., after the intervention they have a different internal definition of what the construct is as before).

Thus, in the present factor model, it is assumed that there are no residual correlations and that there are no response shifts. In practice, it is important to test these assumptions and account for violations to ensure that the difference score reliability can be meaningfully interpreted. However, in the current one-factor model approach, these assumptions cannot be tested as the individual data of the pretest and posttest are not considered. Therefore, below, an extended factor analysis approach is considered in which both the assumption of measurement invariance and the presence of correlated residuals can be tested.

7.2.2 A Factor Model for Pretest-Posttest Scores

In this second approach to estimate difference score reliability, separate measurement models are considered for the pretest item scores and the posttest item scores:

$$X_j^{(\text{pre})} = b_j^{(\text{pre})} + a_j^{(\text{pre})}\xi^{(\text{pre})} + \delta_j^{(\text{pre})} \quad (7.16)$$

and

$$X_j^{(\text{post})} = b_j^{(\text{post})} + a_j^{(\text{post})}\xi^{(\text{post})} + \delta_j^{(\text{post})}. \quad (7.17)$$

where besides the residual variances $\sigma_{\delta_j^{(\text{pre})}}^2$ and $\sigma_{\delta_j^{(\text{post})}}^2$, the variances of the pretest and posttest latent factors, $\sigma_{\xi^{(\text{pre})}}^2$ and $\sigma_{\xi^{(\text{post})}}^2$, and the covariance between the pretest and posttest latent factors, $\sigma_{\xi^{(\text{pre})}, \xi^{(\text{post})}}$, are free parameters. Here we assume the pretest and posttest score to be unidimensional, but the model can straightforwardly be extended to include more factors. Note that in the presentation of the model, the measurement parameters are explicitly different from pretest to posttest, so that it can be tested whether they are equal (so that the assumption of measurement invariance is met). However, to establish the reliability of the difference scores D_j , which is denoted $\omega_{\text{pre} - \text{post}}$ for this model, measurement invariance needs to be

assumed to make the estimate of $\omega_{pre-post}$ meaningful, that is,

$$\omega_{pre-post} = \frac{\left(\sum_{j=1}^J a_j\right)^2 \left(\sigma_{\xi}^2(\text{pre}) + \sigma_{\xi}^2(\text{post}) - 2\sigma_{\xi(\text{pre}),\xi(\text{post})}\right)}{\left(\sum_{j=1}^J a_j\right)^2 \left(\sigma_{\xi}^2(\text{pre}) + \sigma_{\xi}^2(\text{post}) - 2\sigma_{\xi(\text{pre}),\xi(\text{post})}\right) + 2\sum_{j=1}^J \sigma_{\delta_j}^2} \quad (7.18)$$

where a_j is the factor loading of item j for the pretest and posttest and $\sigma_{\delta_j}^2$ is the residual variance of item j for the pretest and posttest.

7.2.2.1 Categorical Item Scores

Comparable to the above, it again holds that the dependent variables in the factor model, $X_j^{(\text{pre})}$ and $X_j^{(\text{post})}$, are commonly assumed to be multivariate normal variables. In the present case, where both the pretest and posttest are modeled, this assumption is even more problematic, as the pretest and posttest scores commonly do not have more than five levels in practice (i.e., in the case of a Likert scale questionnaire). In addition, in questionnaire data, scores are commonly skewed due to floor or ceiling effects. Therefore, the ordinal nature of the data is taken into account using the same approach as above, that is, it is assumed that the actual observed item scores, denoted $X_j^{(\text{pre})}$ and $X_j^{(\text{post})}$, are categorizations of the underlying normally distributed variables $X_j^{*(\text{pre})}$ and $X_j^{*(\text{post})}$ that follow Eqs. 7.16 and 7.17, respectively. Thus,

$$X_j^{(\text{pre})} = c \quad \text{if} \quad \beta_{jc}^{(\text{pre})} < X_j^{*(\text{pre})} < \beta_{j(c+1)}^{(\text{pre})} \quad \text{for } c = 0, \dots, M_j \quad (7.19)$$

and

$$X_j^{(\text{post})} = c \quad \text{if} \quad \beta_{jc}^{(\text{post})} < X_j^{*(\text{post})} < \beta_{j(c+1)}^{(\text{post})} \quad \text{for } c = 0, \dots, M_j \quad (7.20)$$

Note that in estimating $\omega_{pre-post}$ from Eq. 7.18, the thresholds need to be invariant across pretest and posttest to ensure a meaningful estimate of the factor model reliability.

7.2.2.2 Estimation and Identification

Here the estimation and identification of the two factor model approaches to difference score reliability are presented. Identification of the difference score factor model from Eqs. 7.12 and 7.15 can be done by fixing the variance and mean of the latent difference factor, $\mu_{\xi_D} = 0$ and $\sigma_{\xi_D}^2 = 1$. Next, the model can be fit to the observed item difference scores by maximum likelihood. However, if the observed

item difference scores are treated explicitly as ordinal using the underlying normal variable D_j in Eq. 7.15, the scale of D_j needs to be identified as well. To this end, the variance of D_j is fixed to 1, implying that the residual variances are equal to $\sigma_{\delta_{D_j}}^2 = 1 - a_j^2$ and that they need not to be estimated.

Identification of the pretest-posttest factor model from Eqs. 7.16, 7.17, 7.19, and 7.20 requires fixing the mean and variance of the pretest latent factor to, respectively, 0 and 1, that is, $\mu_{\xi^{(pre)}} = 0$ and $\sigma_{\xi^{(pre)}}^2 = 1$. The posttest latent factor mean and variance can be estimated and reflect the difference in, respectively, the mean and variance with respect to the pretest latent factor. Similarly as above, as the item scores are treated explicitly as ordinal using the underlying normal variables $X_j^{*(pre)}$ and $X_j^{*(post)}$ in Eqs. 7.19 and 7.20, the scales of $X_j^{*(pre)}$ and $X_j^{*(post)}$ are identified by fixing their variances to 1, implying that the residual variances are equal to $\sigma_{\delta_j}^2 = 1 - a_j^2$ and that they need not to be estimated.

7.2.2.3 Correlated Residuals

As discussed above, the presence of carry-over effects in the data results in correlated residuals between the pretest items and the posttest items. Therefore, in the model above, the residuals are allowed to be correlated. That is, $\sigma_{\delta_j \delta_j}$ denotes the residual covariance between the pretest scores and the posttest scores of item j . Note that even though the residual variances are fixed due to the ordinal nature of the data (see above), these residual covariances can be estimated for all items $j = 1, \dots, J$. The significance of the residual covariances can be tested by comparing the model fit of a model including residual covariances to the model fit of a model without residual covariances. In addition, individual residual covariances can be tested on significance by a Wald test.

7.2.2.4 Testing Measurement Invariance

A key assumption in calculating difference score reliability in factor models is that there is no response shift, meaning that the measurement parameters a_j , b_j , and β_{jc} are equal across the pretest and the posttest. This assumption cannot be tested in the factor model for difference score in Eqs. 7.12 and 7.15, but it can be tested in the pretest-posttest factor model from Eqs. 7.16, 7.17, 7.19, and 7.20. To this end, a series of increasingly restrictive models is fit to the pretest and posttest data.

The first model considered is referred to as “configural invariance.” In this model, all measurement parameters are free across the pretest and posttest, and the factor means and variances are fixed to 0 and 1, respectively, for identification. In the next model, “metric invariance,” the factor loadings are equated over the pretest and the posttest and the variance of the posttest latent factor is estimated (the

variance of the pretest latent factor should still be fixed to 1 for identification of the model; see above). Finally, the “factorial invariance” model is considered which is equivalent to the model presented above in Eqs. 7.16, 7.17, 7.19, and 7.20. That is, besides the factor loadings, the thresholds are also equated to be equal across the pretest and the posttest, and the mean of the posttest latent factor is estimated freely (the mean of the pretest factor should still be fixed to 0 for identification of the model).

If the “factorial invariance” model fits the best among the other competing models, measurement invariance is tenable. However, if one of the less restrictive models fit better, this indicates that there are differences in the measurement parameters across the pretest and posttest. In that case, this source of misfit should be inferred. If the misfit is due to a minor number of items, these can be removed and the reliability can be based on the remaining item.

7.3 Simulation Study

7.3.1 Design

To study the two factor model approaches above, a simulation study is considered. Data are simulated for five-point Likert items following the approach by Gu et al. (2021). That is, a graded response model (Samejima, 1969) is used to generate the item scores for the pretest and posttest. A sample size of $N = 500$ is used with either $J = 10$ or $J = 25$ items. The discrimination parameters for the items are set to increasing and equally spaced values between 1 and 2. The category threshold parameters are set as follows: For the items, a mean threshold is specified by equally spaced increasing values between -2 and 2 . For each item, the thresholds are then obtained by adding the mean threshold value to -2 , -0.75 , 0.75 , and 2 . As a result, the items are increasing in “item easiness” (or “item attractiveness” for personality items). The latent factor at the pretest is drawn from a normal distribution with mean 0 and variance 1. The latent factor at the posttest is calculated by adding a normally distributed variable with mean 0 and variance σ_{diff}^2 to the latent factor from the pretest. As a result the covariance between the latent factors equals 1 and the variance of the posttest latent factor equals $1 + \sigma_{\text{diff}}^2$. Following Gu et al., σ_{diff}^2 is equal to either 0.15 or 0.5.

The design also includes a manipulation of the residual correlations, which were either absent or present. To impose residual correlations, a carry-over effect was introduced in the data following the procedure by Gu et al. (2021). That is, for a random selection of 50% of the subjects in the data, the following transformation of

the posttest scores is conducted:

$$X_j^{(\text{post})} = \begin{cases} X_j^{\text{pre}} & \text{if } |X_j^{(\text{post})} - X_j^{\text{pre}}| = 1 \\ X_j^{\text{pre}} + 1 & \text{if } X_j^{(\text{post})} - X_j^{\text{pre}} \geq 2 \\ X_j^{\text{pre}} - 1 & \text{if } X_j^{(\text{post})} - X_j^{\text{pre}} \leq -2 \end{cases} \quad (7.21)$$

This corresponds to the strong carry-over effect condition in the simulation of Gu et al. Contrary to Gu et al., here, dimensionality of the pretest and posttest scores is not manipulated, as in a factor analysis framework this is less of interest (unidimensionality can easily be established using model fit diagnostics as is illustrated in the real data application below). Each cell in the design is replicated 100 times.

7.3.2 Dependent Variables

To the simulated data, the following approaches discussed in this chapter are applied:

- A. **CTT**: The sum score-based classical test theory approach (Eq. 7.5)
- B. **CTTT-D**: The item difference score-based classical test theory approach (Eq. 7.8)
- C. **FA-D-cat**: The categorical item difference score-based factor analysis approach (Eqs. 7.12, 7.15, and 7.13)
- D. **FA-D-con**: The continuous item difference score-based factor analysis approach (Eqs. 7.12 and 7.13)
- E. **FA-I-cov**: The item score-based factor analysis approach (Eqs. 7.16, 7.17, 7.19, 7.20, and 7.18) with residual covariances
- F. **FA-I**: The item score-based factor analysis approach (Eqs. 7.16, 7.17, 7.19, 7.20, and 7.18) without residual covariances

For the classical test theory approaches, λ_2 from Eq. 7.9 is used as the lower bound estimate of the reliability. In addition, note that for **FA-D-cat** and **FA-D-con** above, the item difference score factor analysis approach is applied by either assuming the difference scores to be normal (i.e., fitting Eq. 7.12 to the item difference scores) or by assuming the difference scores to be ordinal (i.e., by fitting the model subject to Eq. 7.15). In addition, the item score-based factor analysis approach is applied both with (**FA-I-cov**) and without residual covariances (**FA-I**). The factor models are estimated using the R-package “lavaan” (Rosseel, 2012) using maximum likelihood (for normal difference scores) or weighted least squares (for ordinal difference scores or for the ordinal pretest-posttest item scores). The classical test theory reliability coefficient λ_2 is estimated using R package “Lambda2” (Hunt, 2013).

7.3.3 Results

Tables 7.1, 7.2, and 7.3 contain, respectively, the bias, the root mean squared error, and the standard deviation of the reliability estimates across the replications in the study design. In addition, the results from the tables are graphically represented in the boxplots in Fig. 7.1. If there is no bias, the root mean squared error and the standard deviation are the same. As can be seen, bias is generally low and negative for all methods, meaning that the reliability is slightly underestimated for all methods. Only in the case of a carry-over effect, **FA-I** produces large bias. This can be attributed to the residual covariances in the data due to the carry-over effects which are unmodelled in **FA-I**. These covariances are absorbed in the factor covariance causing the factor covariance to be overestimated and the reliability to be underestimated (see Eq. 7.18). Indeed, for **FA-I-cov** such a severe bias does not occur.

The factor analysis approaches are generally somewhat less biased as compared to the classical test theory approach. With the **CTT-D** outperforming the **CTT**

Table 7.1 Bias of the reliability estimates with respect to the true reliability (true) for the different approaches

CF	J	Var.	True	Bias					
				CTT	CTT-D	FA-D-cat	FA-D-con	FA-I-cov	FA-I
No CF	10	Small	0.628	-0.159	-0.087	-0.060	-0.088	-0.045	-0.044
No CF	10	Large	0.849	-0.095	-0.066	-0.047	-0.064	-0.049	-0.048
No CF	25	Small	0.808	-0.078	-0.056	-0.038	-0.058	-0.029	-0.027
No CF	25	Large	0.934	-0.038	-0.031	-0.022	-0.031	-0.024	-0.023
CF	10	Small	0.628	-0.146	-0.090	-0.062	-0.092	-0.039	-1.238
CF	10	Large	0.849	-0.093	-0.066	-0.046	-0.063	-0.037	-0.239
CF	25	Small	0.808	-0.066	-0.055	-0.037	-0.058	-0.021	-0.37
CF	25	Large	0.934	-0.038	-0.031	-0.022	-0.031	-0.017	-0.094

Note. *CF* carry-over effect, *Var* the variance in the posttest scores (small or large)

Table 7.2 Root mean squared error of the reliability estimates for the different approaches

CF	J	Var.	Root mean squared error					
			CTT	CTT-D	FA-D-cat	FA-D-con	FA-I-cov	FA-I
No CF	10	Small	0.163	0.092	0.067	0.093	0.055	0.053
No CF	10	Large	0.097	0.068	0.049	0.066	0.052	0.050
No CF	25	Small	0.079	0.058	0.040	0.060	0.032	0.030
No CF	25	Large	0.039	0.031	0.023	0.032	0.024	0.024
CF	10	Small	0.150	0.094	0.067	0.096	0.055	1.320
CF	10	Large	0.094	0.067	0.048	0.065	0.041	0.242
CF	25	Small	0.068	0.057	0.040	0.060	0.027	0.376
CF	25	Large	0.038	0.031	0.023	0.032	0.019	0.095

Note. *CF* carry-over effect, *Var* the variance in the posttest scores (small or large)

Table 7.3 Standard deviation of the reliability estimates for the different approaches

CF	J	Var.	Standard deviation					
			CTT	CTT-D	FA-D-cat	FA-D-con	FA-I-cov	FA-I
No CF	10	Small	0.033	0.030	0.031	0.030	0.032	0.030
No CF	10	Large	0.017	0.015	0.014	0.016	0.016	0.014
No CF	25	Small	0.014	0.013	0.013	0.014	0.015	0.014
No CF	25	Large	0.007	0.006	0.006	0.007	0.006	0.006
CF	10	Small	0.031	0.027	0.027	0.028	0.038	0.465
CF	10	Large	0.015	0.013	0.012	0.014	0.017	0.039
CF	25	Small	0.015	0.015	0.014	0.015	0.017	0.069
CF	25	Large	0.007	0.006	0.006	0.006	0.008	0.018

Note. *CF* carry-over effect, *Var* the variance in the posttest scores (small or large)

approach in almost all conditions. The **FA-D-con**, in which the difference scores are considered continuous, performs comparable to the classical test theory difference score approach, **CTT-D**. In addition, **FA-D-cat** performs slightly better than **CTT-D** which is understandable as the data is generate according to the factor model.

Overall, **FA-I-cov** is associated with the smallest bias and smallest root mean squared error in the case of a carry-over effect, and **FA-I-cov** and **FA-I** have the smallest bias in the case of no carry-over effect and a small posttest score variance. In the case of no carry-over effect and a large posttest score variance, the factor analysis approaches produce comparable results. The standard deviation of the estimates is generally comparable across the methods, with the **FA-I-cov** having a slightly higher standard deviation. This is due to parameter estimation imprecision, as these approaches are statistically the most complex approaches with more parameters to be estimated as compared to the item difference score factor analysis approaches (**FA-D-cat** and **FA-D-con**). The standard deviations for **FA-I** are large in the case of a carry-over effect as compared to the other approaches due to the severe misfit in this model in this condition.

With respect to the manipulations in the design, the results are straightforward: more items result in less bias, and larger variance in the posttest scores results in less bias (except for the **FA-I** and **FA-I-cov**, it results in a slightly increased bias in the case of ten items). The effect of the carry-over effect is most notable for **FA-I-cov**, which is less biased in the case of a carry-over effect, and for **FA-I** which is more biased in the case of a carry-over effect, as discussed above. For **CTT**, bias seems to slightly decrease in the case of a carry-over effect, while for **CTT-D**, **FA-D-cat**, and **FA-D-con**, bias is comparable between the conditions with and without carry-over effect

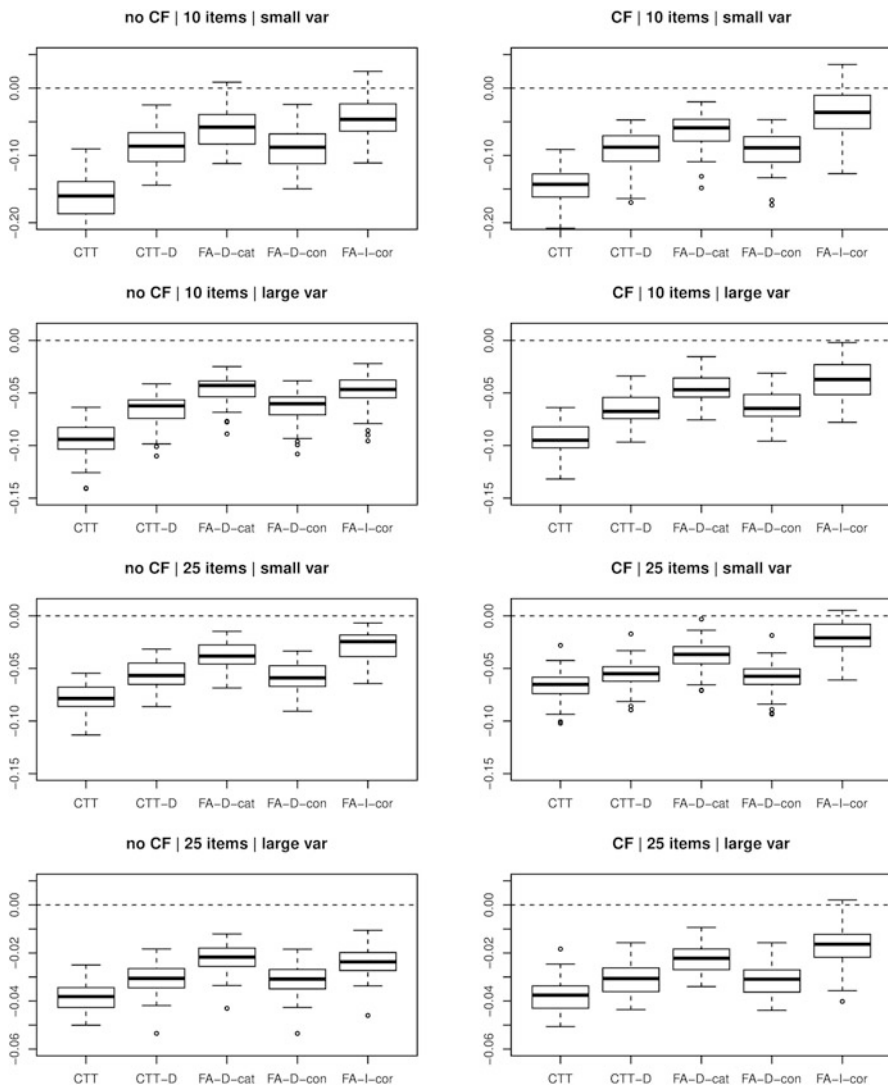


Fig. 7.1 Boxplots of the bias in the reliability estimates over the 100 replications from the simulation study. *CF* carry-over effect, *var* the variance in the posttest scores (small or large). In addition note that approach **FA-I** is not depicted as in the CF conditions, bias is large (see Table 7.1); these results fall mostly outside the graph. For the no CF condition, results are similar to **FA-I-cov**

7.4 Real Data Illustration

7.4.1 Data

In this section, data is analyzed from a study by Mackinnon et al. (2019) in which 263 subjects daily completed various items for 21 days. Aim of the study was to study the link between perfectionism and alcohol drinking. The data is available from Mackinnon et al. (2021). Here, five life satisfaction items from the Satisfaction with Life scale (Diener et al., 1985) are analyzed from the first day (pretest) and the last day (posttest). The items have seven-point Likert scale and comprise:

Q1: In most ways my life is close to my ideal.

Q2: The conditions of my life are excellent.

Q3: I was satisfied with my life.

Q4: I thought that, so far, I have gotten the important things I want in life.

Q5: I thought that, if I could live my life over, I would change almost nothing.

There was no intervention between these two measurement occasions selected for the present analysis. However, it is interesting to see whether life satisfaction increased or decreased in the 21-day period and, most importantly, what the reliability of the difference scores is. The present analysis is not a reanalysis of the Mackinnon et al. (2019) study. In that study, the authors did not use difference scores so the results from the present analysis have no implications for the original study. The present analyses are solely intended to illustrate the methodology from the present chapter.

7.4.2 Analysis

First, the approaches as studied in the simulation study above are applied to the pretest (day 1) and posttest (day 21) data from the Satisfaction with Life scale. Next, the presence of residual covariances is tested, and it is determined whether the assumption of measurement invariance is met. The modeling choices and details are the same as in the simulation study. If necessary, parameter estimates are tested using a 0.01 level of significance. In addition, the following fit indices are considered: The Comparative Fit Index (CFI; Bentler & Bonett, 1980) and Tucker-Lewis Index (TLI; Bentler & Bonett, 1980) for which values above 0.95 indicate acceptable model fit and values larger than 0.97 indicate good model fit (Schermelleh-Engel et al., 2003) and the Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993) for which values smaller than 0.08 indicate acceptable model fit and values below 0.05 indicate good model fit (Schermelleh-Engel et al., 2003). In the tests for measurement invariance, if a given model does not fit well as compared to other competing models, the modification indices are consulted to locate the source(s) of misfit. Modification indices are Lagrange

multiplier tests that indicate for each fixed parameter in the model how much the model fit will improve if that parameter is freed (in χ^2 (1) units). If a fixed parameter causes a substantial source of misfit, this will be evident from a large modification index for that parameter as compared to the other modification indices.

7.4.3 Results

The results of the different reliability approaches are in Table 7.4 for the full scale. For the factor analysis approaches, the table includes the CFI, TLI, and RMSEA model fit indices. As can be seen, the categorical difference score factor model **FA-D-cat** produces the largest reliability estimate. However, the RMSEA indicates that there is some source of misfit with a value above 0.08. In the continuous difference score factor model **FA-D-con**, the model fit is even worse. The pretest-posttest factor model with residual covariance **FA-I-cov** is the only model with an acceptable fit according to all indices. Reliability of the difference score in this model is estimated to be equal to 0.862. Reliability estimates for the other approaches are close, with the sum score-based classical test theory approach **CTT** producing the lowest reliability estimate. From the results of the **FA-I-cov** model, it appears that items 4 and 5 have a residual covariance between the pretest and posttest which are significant and equal to 0.158 (SE: 0.019) for item 4 and 0.227 (SE: 0.028) for item 5. In addition, in this model, the latent difference between posttest and pretest indicated no significant change: 0.088 (SD: 0.059).

In the factor model applied to the data, the assumption is made that the measurement model parameters are the same for the pretest and posttest. As discussed above, this measurement invariance assumption can be tested in the item scores-based factor model (**FA-I** and **FA-I-cov**). As there are significant residual covariances, the assumption of measurement invariance is tested in **FA-I-cov**. See Table 7.5 for the results. As can be seen, the configural model fits acceptable to good according to the fit indices; however, if the loadings are equated in the metric model, fit deteriorates which is mostly evident in the RMSEA. Judged by the modification indices (not tabulated), the loading of item 3 is freed across the pretest and posttest which improves model fit. However, still the model fits worse as compared to the

Table 7.4 Reliability estimates of the difference scores for the Satisfaction with Life scale

Approach	#par	χ^2	df	CFI	TLI	RMSEA	Full scale	Items 3 and 4 omitted
CTT							0.774	0.556
CTT-D							0.864	0.753
FA-D-cat	58	16.712	5	0.997	0.994	0.095	0.891	0.813
FA-D-con	11	78.249	9	0.970	0.939	0.127	0.878	0.802
FA-I-cov	43	147.731	62	0.999	0.999	0.073	0.862	0.765
FA-I	38	208.588	67	0.998	0.999	0.091	0.816	0.705

Table 7.5 Model fit measures of the different models to establish measurement invariance

Model	#par	χ^2	df	CFI	TLI	RMSEA
Configural invariance	76	61.780	29	1.000	0.999	0.066
Metric invariance	72	88.860	33	0.999	0.999	0.081
Metric invariance revised 1	73	75.476	32	0.999	0.999	0.073
Metric invariance revised 2	74	63.101	31	1.000	0.999	0.063
Factorial invariance	45	121.808	60	0.999	0.999	0.063

Note. #par: number of parameters in the model

configural model. Therefore, on the basis of the modification indices, the loading of item 4 is also freed across pretest and posttest. The resulting model has a similar fit as compared to the configural model and is therefore accepted. This implies that the loadings of item 3 and 4 are not invariant across the pretest and posttest. This may have caused the poor fit of the **FA-D-cat** model as a loading difference across pretest and posttest introduces multidimensionality in the difference scores. Next, equal thresholds are introduced, and the model fit improves slightly indicating that the assumption of equal thresholds is tenable.

As items 3 and 4 showed violations of measurement invariance, reliability is recomputed within each approach, but by not taking items 3 and 4 into account (i.e., for the classical test theory approaches, they are removed, and for the factor analysis approaches, their parameters are not taken into account in calculating reliability). Table 7.4 shows the resulting reliability coefficients which are smaller, but, at least for the **FA-I-cov** (which is the best fitting model), still acceptable (0.765). However, reliability is not as good as in the initial analysis including all items. In practice, if items violate measurement invariance (like items 3 and 4 in the present illustration), these items should not be used to assess change as the change on these items does not reflect change on the underlying factor for interest.

7.5 Discussion

In this chapter, two factor analysis approaches have been presented to estimate change score reliability. The first approach is a direct translation of the item level classical test theory approach by Gu et al. (2021). The other is an extended approach that enabled tests on measurement invariance and residual covariances. As appeared from the simulation study, the reliability estimates from the classical test theory approaches and the factor analysis approaches were close to each other, with the factor model estimates being slightly less biased overall. However, a direct comparison of the classical test theory approaches and the factor analysis approaches was not the aim of this chapter as both approaches differ intrinsically (as discussed in this chapter). In addition, in the simulation study, the data were generated using a factor model, putting the factor analysis approaches in an advantage. The main aim of this chapter was to present a factor model approach to

change score reliability to enable tests on dimensionality, measurement invariance, and residual covariances as these aspects have been shown to bias classical test theory approaches.

The first factor analysis approach presented makes use of the item difference scores, similarly as Gu et al. (2021). By doing so, strict tests on measurement invariance are not possible; however, dimensionality can be assessed. That is, as shown in the real data application, violations of measurement invariance can result in multidimensionality as the item difference scores contain variance due to the latent difference factor and due to the latent posttest factor (which measures something psychometrically different). In the item difference score factor analysis approach, this can be detected using goodness of fit measures like the CFI and RMSEA. The presence of residual correlations in the approach will however go unnoticed.

In the second factor analysis approach which uses both the pretest and posttest scores, measurement invariance and residual correlations can explicitly be tested. As discussed above, violations of measurement invariance may be due to a response shift. Response shifts, in turn, are due to the subjects recalibrating, reprioritizing, or redefining their internal representation of the construct being measured. Oort (2005) noted that if subjects rely on redefining, this will violate the configural invariance model in which items will have a different factor configuration on the posttest as compared to the pretest. If subjects rely on reprioritization, the metric invariance model will be violated as the size of the factor loadings will differ between pretest and posttest. Recalibration will result in violations of intercept invariance (in the case of uniform effects) and in violations of the invariance of the residual variances (in the case of nonuniform effects); see also Fokkema et al. (2013). However, as in this chapter, the focus was on ordinal data, the model does not contain free residual variances or intercepts, but thresholds instead. These thresholds pick up both the uniform and nonuniform effects (as the thresholds and residual variances are not uniquely defined; see Takane & De Leeuw, 1987). As in the present real data analysis two factor loadings were found to differ across the pretest and posttest of the life satisfaction scale, it can be concluded that—for these items—, subjects have reprioritized the life satisfaction construct. Without these items, the difference score reliability turned out to be smaller. Using a classical test theory approach, it would have been challenging to detect this.

References

- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588–606. <https://doi.org/10.1037/0033-2909.88.3.588>
- Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage. <https://doi.org/10.1177/0049124192021002005>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J., & Furby, L. (1970). How we should measure “change”: Or should we? *Psychological Bulletin*, *74*(1), 68–80. <https://doi.org/10.1037/h0029382>
- Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13
- Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309–326. <https://doi.org/10.1111/j.2044-8317.1994.tb01039.x>
- Dunn, T. J., Baguley, T., & Brunsten, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. <https://doi.org/10.1111/bjop.12046>
- Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520–531. <https://doi.org/10.1037/a0031669>
- Gu, Z., Emons, W. H., & Sijtsma, K. (2018). Review of issues about classical change scores: A multilevel modeling perspective on some enduring beliefs. *Psychometrika*, *83*(3), 674–695. <https://doi.org/10.1007/s11336-018-9611-3>
- Gu, Z., Emons, W. H., & Sijtsma, K. (2021). Estimating difference-score reliability in Pretest–Posttest settings. *Journal of Educational and Behavioral Statistics*, *46*(5), 592–610. <https://doi.org/10.3102/1076998620986948>
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. <https://doi.org/10.1007/BF02288892>
- Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, *64*(2), 144–150. <https://doi.org/10.1037/0021-9010.64.2.144>
- Hunt, T. (2013). *Lambda4: Collection of internal consistency reliability coefficients* (Version 3.0) [Computer software]. CRAN. <https://CRAN.R-project.org/package=Lambda4>
- Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research*, *47*(1), 121–150. <https://doi.org/10.3102/00346543047001121>
- Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). The University of Wisconsin Press.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Adison-Wesley.
- Mackinnon, S. P., Ray, C. M., Firth, S. M., & O’Connor, R. M. (2019). Perfectionism, negative motives for drinking, and alcohol-related problems: A 21-day diary study. *Journal of Research in Personality*, *78*, 177–188. <https://doi.org/10.1016/j.jrp.2018.12.003>
- Mackinnon, S. P., Ray, C. M., Firth, S. M., & O’Connor, R. M. (2021). Data from “Perfectionism, negative motives for drinking, and alcohol-related problems: A 21-day diary study”. *Journal of Open Psychology Data*, *9*(1), 1. <https://doi.org/10.5334/jopd.44>
- McConnel, K., Strand, I. E., & Valdes, S. (1998). Testing temporal reliability and carry-over effect: The role of correlated responses in test-retest reliability studies. *Environmental and Resource Economics*, *12*(3), 357–374. <https://doi.org/10.1023/A:1008264922331>
- McDonald, R. P. (1978). Generalizability in factorable domains: “Domain Validity and Generalizability”. *Educational and Psychological Measurement*, *38*(1), 75–79. <https://doi.org/10.1177/001316447803800111>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Erlbaum.
- Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*(3), 293. <https://doi.org/10.1037/1082-989X.1.3.293>

- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Oort, F. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, 14(3), 587–598. <https://doi.org/10.1007/s11136-004-0830-y>
- Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, 62(11), 1126–1137. <https://doi.org/10.1016/j.jclinepi.2009.03.013>
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. <https://doi.org/10.1037/a0029315>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Monograph No. 17). Psychometric Society. <https://doi.org/10.1007/BF03372160>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23–74. <http://www.mpr-online.de>
- Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, 86(4), 843–860. <https://doi.org/10.1007/s11336-021-09789-8>
- Sijtsma, K., & Van der Ark, L. A. (2021). *Measurement models for psychological attributes*. Chapman and Hall/CRC. <https://doi.org/10.1201/9780429112447>
- Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science & Medicine*, 48(11), 1507–1515. [https://doi.org/10.1016/s0277-9536\(99\)00045-3](https://doi.org/10.1016/s0277-9536(99)00045-3)
- Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393–408. <https://doi.org/10.1007/BF02294363>
- Williams, B. J., & Kaufmann, L. M. (2012). Reliability of the go/no go association task. *Journal of Experimental Social Psychology*, 48(4), 879–891. <https://doi.org/10.1016/j.jesp.2012.03.001>

Chapter 8

Handling Missing Data in Principal Component Analysis Using Multiple Imputation



Joost R. van Ginkel

Abstract Principal component analysis (PCA) is a widely used tool for establishing the dimensional structure in questionnaire data. Whenever questionnaire data are incomplete, the missing data need to be treated prior to carrying out a PCA. Several methods exist for handling missing data prior to carrying out a PCA. The current chapter first discusses the most recent developments regarding the treatment of missing data in PCA. Next, of these methods, the method that is most promising both from a theoretical and practical point of view will be discussed in more detail, namely, multiple imputation. Finally, some extensions of multiple imputation to other PCA-related techniques or to statistics within PCA beyond the basics are discussed, and some general recommendations regarding the use of PCA on multiply imputed datasets in different statistical software packages will be given.

8.1 Introduction

One important part of establishing the psychometric properties of a test or questionnaire is determining its dimensional structure. Oftentimes measurement instruments measure different aspects of the same psychological construct. For example, a questionnaire may measure different ways in which one can be religious (Hills et al., 2005) or different aspects of schizotypal personality disorder (Mata et al., 2005).

Although establishing the dimensional structure of a measurement instrument is mostly done in personality assessment, there are also situations in educational settings where dimensionality of a measurement instrument may be relevant. For example, in a school setting, one may be interested in students' attitudes towards different types of bullying (Boulton et al., 1999) or different aspects of students' well-being (Borgonovi & Pál, 2016). As the developer of such measurement instruments, you may want to know whether its items indeed measure the specific aspect

J. R. van Ginkel (✉)
Methodology and Statistics, Leiden University, Leiden, The Netherlands
e-mail: jginkel@fsw.leidenuniv.nl

of the trait that they are intended to measure. In such cases a statistical technique is used that establishes which items measure which aspect of the underlying construct. One widely used technique for this purpose is *principal component analysis* (PCA).

In practice, many datasets that are used for determining the dimensional structure of questionnaires suffer from missing data. When data are incomplete, this complicates the use of PCA or any analyses that are aimed towards determining the dimensional structure. When missing data are not properly handled, erroneous conclusions may be drawn about dimensional structure of the measurement instrument. It is therefore important that missing data are properly treated prior to determining the dimensional structure.

The current chapter is going to focus on a situation where one is interested in determining the dimensional structure of a test or questionnaire using PCA in an incomplete dataset. In the first part of this chapter, an overview of the most recent developments of missing data handling in PCA will be given. In this overview, several methods for handling missing data in PCA are going to be discussed. The second part will focus on the method that is the most promising one both from a theoretical and practical point of view in more detail: multiple imputation. The chapter will end with some extensions of missing data handling in PCA to statistics within PCA beyond the basics and to PCA-related techniques, and some general recommendations regarding the use of PCA on multiply imputed datasets in several statistical software packages are given.

8.2 Principal Component Analysis

Within a questionnaire, different subsets of items may exist that each are supposed to measure a different aspect of the same construct. Such a subset is also called a subscale. In PCA, the goal is to reduce a large number of continuous variables J to a smaller number of components, K . Although theoretically the variables need to be continuous, in practice PCA is regularly applied to items measured on a Likert scale.

Suppose \mathbf{Z} is the standardized dataset consisting of the responses of I respondents to J items. In PCA, by means of a singular value decomposition, \mathbf{Z} is decomposed as:

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}' \quad (8.1)$$

Here, \mathbf{U} is a column wise orthonormal $N \times J$ matrix, \mathbf{V} is a column wise orthonormal $J \times J$ matrix, and $\mathbf{\Lambda}$ is a $J \times J$ diagonal matrix with the singular values on the main diagonal. The singular values are the square roots of the eigenvalues. An important part of the output in PCA that gives insight in how the items in the data are related to the different underlying components is the $J \times J$ component matrix. This matrix is computed as $\mathbf{A} = N^{-1/2}\mathbf{V}\mathbf{\Lambda}$ and contains the correlations between the variables and the components. These correlations coincide with the regression

coefficients (loadings) from multivariate multiple regression of the item scores on the principal components.

In the original singular value decomposition, there are as many components as there are variables. However, usually only the first few components explain a substantial portion of the variance of the variables in \mathbf{Z} . Additionally, given that a goal in PCA is to reduce the original number of variables J to a smaller set of dimensions K ($K < J$) and given that in PCA the dimensions are represented by the components, usually only a smaller number of components K are used for interpretation (there are several ways for determining K . See, for example, Furr, 2018, pp. 85–92). The resulting reduced component matrix is denoted by \mathbf{A}_K ($J \times K$).

For interpretational purposes the resulting \mathbf{A}_K matrix may be rotated using either Varimax rotation or Oblique rotation (Harman, 1976). The rotated component matrix is denoted \mathbf{A}_K^* .

8.3 Missing Data

As already mentioned in the introductory section, in the data collection process, it may happen that not all respondents provide answers to all the questions in the questionnaire. Reasons for this may be that a respondent finds a question too personal, that (s)he accidentally skipped a question, (s)he did not understand the question, and so on. When respondents have not answered all the questions, this results in a dataset with missing data.

When data are incomplete, this might have consequences for the PCA that is carried out next. Before the PCA can be carried out, the missing data need to be handled. Several ways to deal with missing data in PCA exist (to be discussed later on), ranging from very simple to highly advanced. However, each of these methods makes either explicit or implicit assumptions about the underlying process that caused the missing data, also called the *missingness mechanism*. Rubin (1976) and Little and Rubin (2002) defined three main missingness mechanisms, namely, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). As these missingness mechanisms are extensively described by Rubin (1987), Little and Rubin (2002), and various other literature on missing data, they will only briefly be discussed here.

8.3.1 Missingness Mechanisms

When the data are MCAR, there is no relation between the missing values and any observed or unobserved information. Consequently, the missing data are randomly scattered across the dataset. Under MAR, missing data may depend on observed data but not on unobserved data. It could be, for example, that within different age

groups, respondents have different amounts of missing data on the questions in the questionnaire. If, however, age is observed for all respondents and within each age group the missing data are randomly scattered across the data, then the missingness is MAR. Finally, MNAR is any missingness mechanism that does not qualify as either MCAR or MAR. Thus, under MNAR the missingness depends either on a variable that was not included in the data collection process (e.g., the older people get, the more missing data they have, and age not being observed) or on the value of the missing score itself (people with higher scores on an item in the questionnaire being more likely not to answer the item than people with low scores), or both.

8.3.2 *Methods for Handling Missing Data*

Methods for dealing with missing data in PCA range from ad hoc to highly advanced. Among the ad hoc procedures are *listwise deletion* and *pairwise deletion*; a few examples of advanced methods are *missing data passive* (Meulman, 1982; Takane & Oshima-Takane, 2003), *regularized PCA* (Josse et al., 2009), *EM-covariances* (Bernaards & Sijtsma, 2000), and *multiple imputation* (Rubin, 1987; Van Ginkel & Kroonenberg, 2014). Although the methods mentioned here are not exhaustive, with the exception of listwise deletion and pairwise deletion, they all have in common that they are advanced in the sense that they all carry out the PCA in a statistically sound way, without throwing away any data.

In the abovementioned references, usually the performance of only one of these methods was compared with the performance of other less advanced methods (such as substituting the variable mean for each missing value) or with different variants of the same method. However, none of these studies compared all of these advanced methods with each other. Van Ginkel et al. (2014) did a simulation study in which they compared all of the abovementioned methods. Before discussing the results of their study, each of these methods will be discussed in more detail first. In so doing they will be categorized into three categories, namely, *traditional methods*, *simultaneous methods*, and *sequential methods*.

8.3.2.1 **Traditional Methods**

The traditional methods described by Van Ginkel et al. (2014) are listwise deletion and pairwise deletion. Listwise deletion deletes every case with at least one missing value on any of the variables in the PCA from the analysis. Since usually more data points are thrown away than there are missing data points, listwise deletion is very wasteful. An additional problem of listwise deletion is that in general, unbiased results of statistical analyses are only guaranteed when the data are MCAR. However, in PCA, component loadings are intrinsically biased. This has to do with the fact that they are bound to -1 and $+1$, as in normal correlations (Fisher, 1915). Consequently, in PCA the question is not whether loadings are biased as a

result of listwise deletion, but how much more biased they are than without missing data.

Like listwise deletion, pairwise deletion deletes cases with missing data for the calculation of the component loadings, but in doing so it uses more observed information than listwise deletion does. Pairwise deletion calculates the component loadings from a PCA in a slightly different way than is described in Sect. 8.2. Rather than carrying out a singular value decomposition on the standardized dataset, pairwise deletion computes the component loadings by performing an eigenvalue decomposition on the correlation matrix (technical details are discussed in Tabachnick & Fidell, 2001, pp. 591–595). In doing so it also deletes cases with missing values, but does this for each variable pair for which a correlation is computed, separately. Consequently, in pairwise deletion more information is used than in listwise deletion.

Although pairwise deletion uses more information from the data than listwise deletion does, an implicit assumption is still that the data are MCAR. An additional disadvantage of pairwise deletion is that since each correlation is based on different cases, combinations of correlations may occur that together form a correlation matrix that is not positive semi-definite. Consequently, computational problems may occur when computing the component loadings.

8.3.2.2 Simultaneous Methods

Van Ginkel et al. (2014) discussed two methods that estimate the loadings of the PCA and handle the missing data in the process, namely, *missing data passive* (Meulman, 1982; Takane & Oshima-Takane, 2003) and *regularized PCA* (Josse et al., 2009). Since both methods estimate the PCA and in the process also handle the missing data while not throwing away any information, these methods were referred to as *simultaneous methods*.

The idea of missing data passive is that a weight matrix of 1's (observed data) and 0's (unobserved data) is used in a weighted homogeneity analysis, a categorical form of PCA. Regularized PCA, on the other hand, is based on PCA using weighted least squares (Kiers, 1997; Grung & Manne, 1998). In weighted least squares, after filling starting values for the missing data, an iterative algorithm is used that alternates between a regression analysis predicting the component scores from the current estimates of the loadings and a regression analysis predicting the loadings from the current estimates of the component scores. At each iteration, the estimates for the missing data are updated. Regularized PCA is based on the same principle. The difference with weighted least squares is that regularized PCA uses a smoothing procedure for estimating the missing data in the process. This smoothing procedure is especially useful when many components are extracted as weighted least squares may break down in case of many components.

The simultaneous methods have two theoretical advantages over pairwise deletion. Firstly, they do not throw away data like pairwise deletion does. Secondly, as long as the missing data are related to variables that take part in the PCA, using

these methods will not introduce any additional bias in the component loadings as a result of deviations from MCAR.

8.3.2.3 Sequential Methods

Lastly, Van Ginkel et al. (2014) discussed two methods that treat the missing data separately from the calculation of the component loadings: *EM-covariances* (Bernaards & Sijtsma, 2000) and *multiple imputation* (Rubin, 1987). In EM-covariances, first an expectation-maximization algorithm (EM; Dempster et al., 1977) is used to obtain full information maximum likelihood estimates of the means and covariances of the data under the assumption that the data are multivariate normally distributed. Next, the covariances of the variables that are part of the PCA are converted to correlations, and an eigenvalue decomposition of this correlation matrix is carried out to obtain the component loadings.

EM-covariances has the same theoretical advantages over pairwise deletion that missing data passive and regularized PCA have. However, whereas missing data passive and regularized PCA can only handle MAR mechanisms where the missing data depend on variables that are included in the PCA, EM-covariances can also handle MAR mechanisms where the missingness depends on variables outside the PCA, as long as they are included in the maximum likelihood estimation of the covariance matrix.

Multiple imputation is perhaps the most widely recommended method for dealing with missing data. This procedure works in three steps. In the first step, the missing data are estimated multiple (M) times according to a statistical model that accurately describes the structures present in the data. This results in M complete versions of the incomplete dataset, which only differ in the estimates for the missing data. In the second step, the statistical analysis of interest is applied to each of the M completed datasets, resulting M different outcomes of the same analysis (in the current context, a PCA). Finally, the results of the M analyses are combined into one overall result, using specific calculations, denoted *combination rules* (for the specific PCA context, combination rules will be discussed in Sects. 8.5.1 and 8.5.2).

Like EM-covariances, multiple imputation can handle any MAR mechanism, regardless of whether the missingness depends on a variable within the PCA or outside the PCA. However, an additional advantage of multiple imputation is that the multiply imputed data can be used for almost any type of statistical analysis other than PCA, whereas the means and covariance matrices of EM-covariances can only be used as the input for analyses that use means and covariances.

8.3.2.4 Which Method for Handling Missing Data in PCA Is the Preferred One?

In this subsection a short summary of the results found by Van Ginkel et al. (2014) is given. Based on the results and on the theoretical properties of each method, a recommendation is given on which method is generally the best one to use.

To determine the performance of each method, Van Ginkel et al. (2014) studied three quality measures in their simulation study, namely, the root mean squared bias (RMSB) of the component loadings, the mean bias (MB) of the component loadings, and the average number of items assigned to the incorrect component, denoted the *classification error* (CE). The RMSB, MB, and CE were defined as follows: Suppose that a_{jk}^* is the population component loading of item j on Varimax rotated component k and $\hat{a}_{jk,d}^*$ is the corresponding loading for the incomplete simulated dataset d ($d = 1, \dots, D$) in a specific condition of the simulation study (specific missing data handling method, specific percentage of missingness, etc.). For the specific condition, the RMSB is:

$$\text{RMSB} = \frac{1}{D} \sum_{d=1}^D \sqrt{\frac{\sum_{j=1}^J \sum_{k=1}^K (\hat{a}_{jk,d}^* - a_{jk}^*)^2}{JK}}, \tag{8.2}$$

and the MB is:

$$\text{MB} = \frac{1}{JKD} \sum_{d=1}^D \sum_{j=1}^J \sum_{k=1}^K (\hat{a}_{jk,d}^* - a_{jk}^*). \tag{8.3}$$

As for the CE, define f as the component number of the component for which it holds that

$$a_{jf}^* = \max \left(|a_{j1}^*|, \dots, |a_{jK}^*| \right)$$

and g as the component number of the component for which it holds that

$$\hat{a}_{jg,d}^* = \max \left(|\hat{a}_{j1,d}^*|, \dots, |\hat{a}_{jK,d}^*| \right).$$

Next, based on guidelines by Comrey and Lee (1992) that state that loadings below 0.32 should not be interpreted, define:

$$\begin{aligned} w_{j,d} &= 0 \text{ if } \max \left(|a_{j1}^*|, \dots, |a_{jK}^*| \right) < 0.32 \text{ and } \max \left(|\hat{a}_{j1,d}^*|, \dots, |\hat{a}_{jK,d}^*| \right) < 0.32 \\ w_{j,d} &= 0 \text{ if } \max \left(|a_{j1}^*|, \dots, |a_{jK}^*| \right) > 0.32 \text{ and } \max \left(|\hat{a}_{j1,d}^*|, \dots, |\hat{a}_{jK,d}^*| \right) > 0.32 \\ &\text{and } f = g \\ w_{j,d} &= 1 \text{ otherwise.} \end{aligned}$$

For the specific condition, the CE is:

$$CE = \frac{1}{D} \sum_{d=1}^D \sum_{j=1}^J w_{j,d}. \quad (8.4)$$

The results of the traditional methods will be discussed first. Van Ginkel et al. (2014) studied the performance of all methods under both MCAR, MAR, and MNAR. The bias of the individual component loadings was not studied so it remains unclear how much bias in component loading deviations from MAR introduced for the advanced methods and how much bias deviations from MCAR introduced for listwise deletion and pairwise deletion. However, it became clear from the study that listwise deletion did not perform well on either the *RMSB*, the *MB*, or the *CE*, regardless of the missingness mechanism. Additionally, for high percentages of missing data, listwise deletion was not even feasible because after removing the incomplete cases, no or too few complete cases were left to analyze. In short, based on the results of Van Ginkel et al. (2014), listwise deletion is not recommended for PCA.

As for pairwise deletion, Van Ginkel et al. (2014) found that this method actually gave satisfactory results on all three quality measures, regardless of the missingness mechanism. Additionally, computational problems did not occur in the situations studied by Van Ginkel et al. (2014). However, the latter does not mean that these problems cannot occur in practice, so using pairwise deletion in practice may not always be feasible.

Regarding the simultaneous methods, Van Ginkel et al. (2014) found that, firstly, missing data passive generally gave results that were similar to pairwise deletion with respect to the outcome measures and that missingness mechanism did not have a substantial effect on the performance of missing data passive. Regularized PCA, on the other hand, produced results that were slightly worse than those of pairwise deletion and missing data passive. Thus, despite their theoretical advantages over pairwise deletion, they do not seem to show in the quality measures in the study by Van Ginkel et al. (2014).

Finally, regarding the sequential methods, Van Ginkel et al. (2014) found that regarding the outcome measures multiple imputation and EM-covariances performed similar to pairwise deletion. Thus, despite the theoretical advantages of multiple imputation and EM-covariances over the other methods, this does not really seem to show in the quality measures either. This leaves us with the question which method is the preferred one.

Of all the methods discussed in the previous subsections, multiple imputation is the method that is most preferred *from a theoretical point of view* because it will not introduce additional bias in component loadings under any MAR mechanism. Not considering lower benchmark listwise deletion, pairwise deletion is the least preferred method from a theoretical point of view because it assumes MCAR, and it may run into computational problems. However, Van Ginkel et al. (2014) showed that although multiple imputation was one of the better performing methods, it did not perform any better than pairwise deletion. Furthermore, pairwise deletion

(together with listwise deletion) is the simplest method for handling missing data in PCA (it is included in most statistical software packages and does not require any additional preprocessing of the data). This raises the question whether pairwise deletion should not be preferred over any missing data handling method for PCA at all times, including multiple imputation. Not quite, as being simple and performing well on a number of outcome measures may not necessarily be the only criteria for preferring a missing data method in PCA over others. There are other things that may have to be taken into consideration as well.

Firstly, even though it did not occur in the simulation study by Van Ginkel et al. (2014), in practice computational problems may still occur using pairwise deletion. Secondly, even when computational problems do not occur, then still the question is what the sample size is that the PCA solution is based on as some correlations are computed for different cases and different number of cases than others.

Thirdly, sometimes researchers may be interested in confidence intervals of principal component loadings. Van Ginkel and Kiers (2011) developed ways to construct bootstrap confidence intervals for component loadings in multiply imputed datasets (more on this in Sect. 8.4) and showed that these ways performed well regarding coverage of the population loadings. For pairwise deletion there is no way to construct bootstrap confidence intervals of the component loadings.

Finally, and most importantly, a practical advantage of multiple imputation over all other methods for handling missing data including pairwise deletion is that multiple imputation provides the researcher a complete dataset which can be used for other statistical analyses as well. In practice, a dataset is almost never subjected to one single statistical analysis, so it is desirable to have a general solution for all analyses that are carried out on the dataset, such that all analyses on these datasets are comparable regarding sample size, regarding the cases used, and regarding the data points (both observed and imputed). When not imputing the data and analyzing only the usable data however, for some analyses listwise deletion will be applied (and for each of these analyses, different cases may be used, depending on which variables are included in the specific analysis), for other analyses full information maximum likelihood will be applied, and yet for other analyses, pairwise deletion (as in PCA) will be applied. This will make the statistical analyses mutually incomparable.

Additionally, while pairwise deletion may give good results for PCA, this is not necessarily the case for other analyses that are applied to the dataset. It has been well established that multiple imputation performs better regarding bias and coverage of parameters than methods based on deleting data (listwise/pairwise deletion). Consequently, when a researcher decides not to impute the data, conclusions regarding PCA may be valid, but conclusions based on other statistical analyses on the same dataset may not.

In short, although from the study of Van Ginkel et al. (2014) we cannot conclude that multiple imputation necessarily recovers the PCA solution better than pairwise deletion does, there are numerous other advantages of multiple imputation over pairwise deletion in PCA. For the remainder of this chapter we are hence going to take the standpoint that multiple imputation is to be recommended most for handling

missing data in PCA. Hence, we are going to get into more detail about multiple imputation in the context of PCA in the next section.

8.4 Multiple Imputation in Principal Component Analysis

As already said in Sect. 8.3.2.3, multiple imputation works in three steps: (1) the imputation step, where multiple estimates for the missing data are generated; (2) the analysis step, where each of the resulting M imputed datasets is analyzed using the statistical analysis of interest; and (3) the combination of the M results into one overall result. Various methods for generating multiple estimates of the missing data in step 1 have been developed, and various texts have been written on them (e.g., Schafer, 1997; Van Buuren, 2018). The general process of generating multiple imputed values for the missing data is not tied to PCA as an analysis for the data, but is generally the same for all statistical analyses that follow after the data have been multiply imputed. Consequently, technical details regarding the process of generating multiple imputed values are not further discussed here. The interested reader is referred to Van Buuren (2018).

In the context of PCA, the second step in the multiple imputation process is carrying out a PCA on each of the M complete versions of the incomplete dataset. This step has already been explained in Sect. 8.2 so this step will not be discussed here either. This leaves us with the third and final step of the multiple imputation process: the combination of M PCA results into one overall PCA result. Van Ginkel and Kroonenberg (2014) discussed combination techniques for the results of PCA in multiply imputed data, which will be discussed next.

8.4.1 Combining the Component Loadings

8.4.1.1 The Problem of Traditional Combination Rules When Applied to PCA

Once a PCA has been obtained from each of the M imputed datasets, this leaves us with M sets of component loadings. The question is how these component loadings are combined into one overall set of component loadings. Rubin (1987) defined combination rules for a parameter estimate with its statistical test and confidence interval. An overall parameter estimate is obtained by averaging the M estimates of the parameter. Considering a component loading $a_{jk,m}$ on variable j on component k to be a parameter estimate of imputed dataset m , a direct application of Rubin's combination rules for parameter estimates would come down to averaging the M component loadings $a_{jk,m}$.

Van Ginkel and Kroonenberg (2014) argued that averaging component loadings across M imputed datasets has three potential problems. Firstly, the order of the

components may not be the same for all M imputed datasets. For example, in one imputed dataset, a set of items may load highest on the first component, while in another imputed dataset, this same set may load highest on the second component. This may especially happen when two adjacent components have near equal variance.

Secondly, many questionnaires contain both indicative items (a higher score means a higher amount of the underlying construct) and contraindicative items (a higher score means a lower amount of the underlying construct). When a specific subscale of a questionnaire contains about as many indicative items as contraindicative items, it could happen that in one or more imputed datasets, the signs of the loadings are reversed compared to those of the other imputed datasets. When averaging these loadings, their signs may cancel each other out, resulting in an average loading lower than the average of the absolute values.

A third disadvantage is that even when sign changes of loadings switching of the order of components do not occur among the M $\mathbf{A}_{K,m}$ matrices, then still the M matrices are not optimally aligned as a result of rotational freedom. Because of this rotational freedom, the average solution is computed across solutions that have more variation among each other than necessary (e.g., Chatterjee, 1984; Markus, 1994; Milan & Whittaker 1995; Linting et al. 2007).

8.4.1.2 Using Generalized Procrustes Analysis for Combining the Component Loadings

A procedure that can resolve all of the three abovementioned problems is Generalized Procrustes analysis (Ten Berge, 1977; Gower, 1975). Generalized Procrustes analysis was originally proposed to derive one overall component solution from several ones, not necessarily obtained from multiply imputed data (e.g., from several different studies). However, Van Ginkel and Kroonenberg (2014) proposed this procedure to explicitly combine the results of several PCA solutions obtained from M imputed datasets. In a simulation study, they showed that this method gave better results regarding *RMSB* (see Eq. (8.2)) than averaging of component loadings did.

In the context of M PCA solutions obtained from M imputed datasets, generalized Procrustes analysis works as follows. Suppose that we have unrotated component matrix $\mathbf{A}_{K,m}$ of imputed dataset m ($m = 1, \dots, M$). We need an orthogonal $K \times K$ rotation matrix \mathbf{T}_m for each of the M imputed datasets that minimizes the sum of squared distances between the transformed loading matrices, given by:

$$f(\mathbf{T}_1, \dots, \mathbf{T}_M) = \sum_{i < j} \text{tr}(\mathbf{A}_{K,i}\mathbf{T}_i - \mathbf{A}_{K,i}\mathbf{T}_i)'(\mathbf{A}_{K,j}\mathbf{T}_j - \mathbf{A}_{K,j}\mathbf{T}_j). \quad (8.5)$$

The rotation matrices $\mathbf{T}_1, \dots, \mathbf{T}_M$ are obtained using a procedure that is a generalization of the classical orthogonal Procrustes problem (Green, 1952; Gower, 1971). In the classical Procrustes problem, we have two matrices \mathbf{A} and \mathbf{B} where \mathbf{A}

needs to be optimally rotated to \mathbf{B} . The required rotation matrix for this problem is found as follows: suppose \mathbf{QLV}' is the singular value decomposition of matrix $\mathbf{A}'\mathbf{B}$. The rotation matrix \mathbf{T} is obtained by:

$$\mathbf{T} = \mathbf{QV}' \quad (8.6)$$

Finally, \mathbf{A} can be optimally rotated to \mathbf{B} by post-multiplying \mathbf{A} by \mathbf{T} .

When optimally rotating M component matrices towards each other, we can use an algorithm by Ten Berge (1977, p. 272). Suppose t is the iteration number, and starting at $t = 1$, the algorithm has the following steps:

Step 0: Set $\mathbf{T}_m = \mathbf{I}$ for $m = 2, \dots, M$.

Step 1: Rotate $\mathbf{A}_{K,1}$ optimally to $\mathbf{B} = \sum_{m=2}^M \mathbf{A}_{K,m} \mathbf{T}_m$ using rotation matrix \mathbf{T}_1 as computed in the right-hand side of Eq. (8.6), yielding $\mathbf{A}_{K,1} \mathbf{T}_1^{(t)}$.

Step 2: Rotate $\mathbf{A}_{K,2}$ optimally to $\mathbf{B} = \mathbf{A}_{K,1} \mathbf{T}_1^{(t)} + \sum_{m=3}^M \mathbf{A}_{K,m} \mathbf{T}_m$, yielding $\mathbf{A}_{K,2} \mathbf{T}_2^{(t)}$.

Step M : Rotate $\mathbf{A}_{K,M}$ optimally to $\sum_{m=1}^{M-1} \mathbf{A}_{K,m} \mathbf{T}_m^{(t)}$, yielding $\mathbf{A}_{K,M} \mathbf{T}_M^{(t)}$.

Step $M + 1$: Rotate $\mathbf{A}_{K,1} \mathbf{T}_1^{(t)}$ optimally to $\mathbf{B} = \sum_{m=2}^M \mathbf{A}_{K,m} \mathbf{T}_m^{(t)}$, yielding $\mathbf{A}_{K,1} \mathbf{T}_1^{(t+1)}$.

Next, the steps 2– M are repeated, where t increases with 1 at each iteration, until convergence. Once convergence has been achieved, the mean of all transformed solutions, also denoted the *centroid* solution $\mathbf{A}_{K,C}$, is used as the pooled PCA solution for the M imputed datasets. Like a PCA solution in complete data, $\mathbf{A}_{K,C}$ can be rotated either with an orthogonal or an oblique transformation.

8.4.2 Uncertainty About the Component Loadings

In the traditional way in which PCA is used, usually no statistical tests or confidence intervals are computed. There are procedures for confidence intervals of population component loadings (more on this in Sect. 8.5), but normally PCA is mainly used without any statistical testing.

However, in multiple imputation uncertainty is created about parameter estimates by the fact that for each imputed dataset the imputed values differ and that this results in slightly different sets of PCA loadings for each imputed dataset. Although $\mathbf{A}_{K,C}$ gives an impression of what the actual sample loadings without missing data would have been, there is still uncertainty about this centroid solution as a result of the variation of the imputed values.

Van Ginkel and Kroonenberg (2014) discussed a procedure to show variation in the component loadings as a result of imputation uncertainty. Using this procedure a loading plot of one component against the other is created, which shows both the centroid solution represented by dots and the uncertainty of the centroid solution

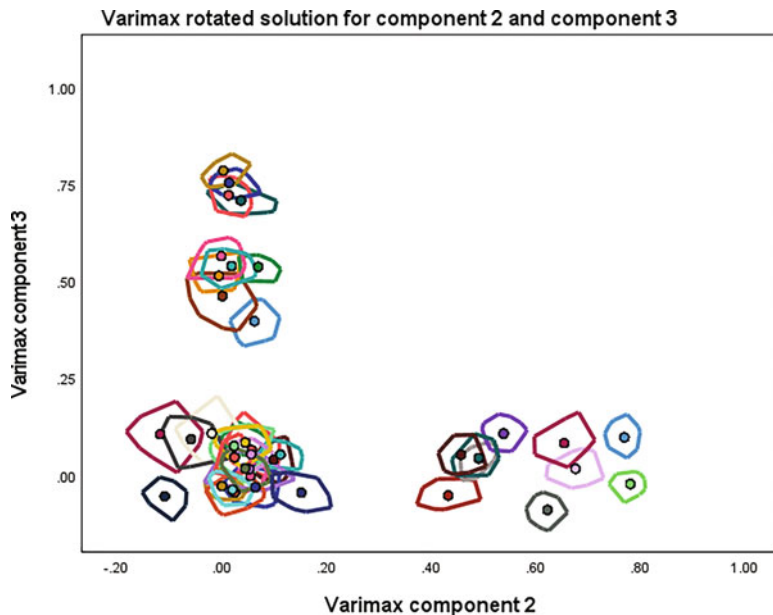


Fig. 8.1 Loading plot of a Varimax rotated four-component solution of components 2 and 3, applied to a multiply imputed dataset with $M = 100$ imputations. The loading plot shows both the centroids and their convex hulls

represented by areas surrounding the dots. These areas are called convex hulls. Figure 8.1 displays a loading plot that includes both the centroid solution of the M PCAs and the convex hulls.

The surface of the convex hulls may serve as a measure of uncertainty about the PCA loadings. These surfaces may be computed in the following way. Each convex hull may be decomposed as several triangles. Suppose a triangle has three sides, namely, a , b , and c , and we define $s = (a + b + c)/2$. See Fig. 8.2. By using Heron's rule dating back to before 200 BC, the surface of one triangle can be determined as $\sqrt{[s(s-a)(s-b)(s-c)]}$. Doing this for all triangles that the convex hull is composed of, and adding up the surfaces, the total surface of the convex hull is obtained.

It should be noted that the convex hulls do not in any way intend to represent some kind of confidence intervals of the population loadings with a specific coverage percentage. All the convex hulls do is give the reader some visual impression of where the uncertainty in the PCA solution lies as a result of the missing data. A loading with a large convex hull is estimated with more uncertainty than a loading with a small convex hull, and the larger a convex hull is, the more cautious we must be regarding the interpretation of its loading. However, in order to assign some more absolute meaning to the convex hulls, Van Ginkel and Kroonenberg (2014) also studied what percentage of the $J \times K$ sample loadings that would be obtained if no data were missing is covered by the convex hulls under various circumstances. What they found was that under $M = 100$ imputations, the

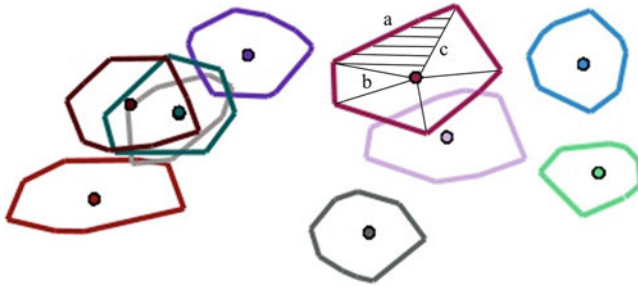


Fig. 8.2 Surface of one triangle in one of the convex hulls

convex hulls usually capture about 80% of the loadings that would be obtained if no data were missing, with percentages of missing data up to 15%. Based on these results, they gave a rough guideline to use $M = 100$ imputations if the researcher wants about 80% of the true sample loadings to fall within the corresponding convex hulls.

Finally, it should be noted that it is possible to use the convex hulls in the form of confidence intervals, using convex hull peeling (Green, 1981) or confidence ellipses (e.g., Josse et al., 2011). However, these confidence intervals do not make any statistical inference about a population loading, only about the true sample loading if no data were missing. A procedure for constructing confidence intervals of the population loadings will be discussed next.

8.5 Extensions

8.5.1 Confidence Intervals of the Component Loadings

As already mentioned in Sect. 8.4.2, in complete data there is the possibility of constructing confidence intervals of population component loadings. Analytical lower and upper bounds of confidence intervals have been derived by various authors (Girshick, 1939; Anderson, 1963; Archer & Jennrich, 1973; Ogasawara, 2000, 2002). However, these analytical confidence intervals have either been derived under the assumption that the data are multivariate normally distributed (Girshick, 1939; Anderson, 1963; Archer & Jennrich, 1973; Ogasawara, 2000), or they require a large sample size (Ogasawara, 2002).

Alternatively, bootstrap confidence intervals may be used for component loadings (Chatterjee, 1984; Efron & Tibshirani, 1994; Kiers, 2004; Lambert et al., 1990, 1991; Linting et al., 2007; Lorenza-Seva & Ferrando, 2003; Markus, 1994; Milan & Whittaker, 1995; Raykov & Little, 1999). Timmermans et al. (2007) studied two bootstrap procedures for component loadings in a simulation study, namely, the percentile method and the bias-corrected and accelerated (BCa) method (Efron,

1987). They found that both bootstrap procedures give better results regarding coverage of the component loadings than analytic methods.

Van Ginkel and Kiers (2011) proposed procedures to combine the bootstrap confidence intervals of both the percentile method and the BC_a method. Suppose that in the complete data case B , bootstrap samples are drawn for constructing confidence intervals of the loadings in \mathbf{A}_K , and the central $(1 - 2\alpha)$ part of the cumulative bootstrap distribution is the confidence interval. Van Ginkel and Kiers (2011) used the centroid solution $\mathbf{A}_{K,C}$ as the component matrix (see Sect. 8.4.1.2). Next, they drew B bootstrap samples from each of the M imputed datasets and used the central $(1 - 2\alpha)$ part of the total of $B \times M$ bootstrap samples as the confidence interval. They did this for both the percentile method and the BC_a method. In a simulation study, they investigated the statistical properties of their proposed procedures, and they turned out to produce coverage percentages close to the theoretical percentages, for various confidence widths (90%, 95%, and 99% coverage). The interested reader is referred to their paper.

8.5.2 Three-Mode Analysis

Three-mode analysis (e.g., Kroonenberg, 2008) is an extension of principal component analysis. It is used in datasets that consist of three different modes, for example, respondents (first mode) and questions on a questionnaire (second mode) at several different time points (third mode). The PCA model can be extended to a situation with three modes in several ways. The three most well-known extensions for three-mode data are the Tucker2 model (Tucker, 1972), the Tucker3 model (Tucker, 1966), and the Parafac model (Harshman, 1970; Carroll & Chang, 1970).

What all three models have in common is that they replace the singular value matrix \mathbf{A} in Eq. (8.1) with a three-dimensional core array that also models the properties of the third mode, represented by different slices. Additionally, while in PCA \mathbf{A} is always a square diagonal matrix, in the Tucker2 and Tucker3 model, the number of rows, columns, and slices of the core array are not necessarily the same. This implies that each mode (respondents, variables, time points) may be summarized by a different number of components. Furthermore, while the PCA model in Eq. (8.1) only has a matrix containing the scores of each respondent on the components (\mathbf{U}) and a matrix with scores of each variable on the components (\mathbf{V}), the Parafac and Tucker3 model also contain a matrix with scores of the third mode on the components.

Kroonenberg and Van Ginkel (2012) proposed rules for combining the results of the Tucker2 model in multiply imputed datasets. These combination rules are similar to the proposed combination rules discussed in Sect. 8.4.1.2. They involve applying generalized Procrustes analysis to both the three-mode equivalent of matrix \mathbf{U} and of matrix \mathbf{V} and by calculating the core matrix from both these two matrices and the M imputed datasets using matrix algebra. For the exact procedure, see Kroonenberg and Van Ginkel (2012).

Van Ginkel and Kroonenberg (2017) found that multiple imputation in combination with generalized Procrustes analysis produced good results of three-mode analysis in terms of RMSB (Eq. (8.2)) as compared to generalized least squares (see Sect. 8.3.2.2), the default method for handling missing data in three-mode analysis. It is, however, hard to tell what the specific influence of the combination techniques is on the RMSB, as for three-mode analysis there are no other combination techniques available than the one proposed by Kroonenberg and Van Ginkel (2012), to compare the procedure with.

8.6 Implementation in Software

Nowadays, most standard statistical software packages have included at least some procedure for creating multiply imputing incomplete datasets. Thus, when applying a PCA to an incomplete dataset, the question is not so much how to find a software package that can multiply impute the data as there are various options for that. The question is more which software package to use for combining the results of PCA on an incomplete dataset once it has been multiply imputed.

The software program `3WayPack` (The Three-Mode Company, 2021) is a freeware program that can be used for several three-mode models. The package also includes an option of using generalized Procrustes analysis. The program requires plain text as input, which is not really convenient when PCA results of multiply imputed datasets are printed in software specific output as they need to be converted to plain text first.

Alternatively, one can use the `shapes` package in R (Dryden & Mardia, 2016). This package can perform generalized Procrustes analysis. However, this package is more generally meant for the statistical analysis of landmark shapes and just happens to be also usable for combining results of PCA applied to multiply imputed dataset.

If one wants to stay completely within the framework of PCA on multiply imputed datasets, then the SPSS macro `GPA.sps` (Van Wingerde & Van Ginkel, 2021) may be used. This macro has been developed for applied researchers who use SPSS for their basic analyses and who want to combine the results of PCA within SPSS. The macro reads PCA output that has been saved to an SPSS data file, performs the calculations, and provides the (possibly Varimax rotated) matrix $\mathbf{A}_{K,C}$ in a new output. Plots with convex hulls as shown in Fig. 8.1 can also be printed.

8.7 Limitations and Final Considerations

Finally, a few limitations within the framework of PCA of multiply imputed datasets, and some points to take into consideration, will be discussed. As pointed out in this chapter, combination rules for component loadings in multiply imputed

datasets have been developed and investigated (e.g., Van Ginkel & Kiers, 2011; Van Ginkel & Kroonenberg, 2014; Van Ginkel et al., 2014). However, in PCA usually more outcomes are used and/or interpreted than only the component loadings.

For example, the component scores of the persons may need to be used for further analysis. At the moment not much has been written on how to compute component scores for multiply imputed datasets. Although not explicitly stated in their paper, Buisman et al. (2020) computed component scores for each imputed dataset m by standardizing the data to \mathbf{Z}_m and using $\mathbf{V}_m = \mathbf{Z}_m \sqrt{N} \mathbf{A}_{K,C}$. It has not been investigated, however, how this ad hoc solution performs in terms of bias in subsequent statistical analyses with these component scores.

As a second example, no combination rules have been defined for the proportion of variance accounted for by the extracted components. One could construct a pooled $\mathbf{\Lambda}$ matrix using a similar procedure for constructing the core three-way array in three-mode analysis discussed in Sect. 8.5.2 (also, see Kroonenberg & Van Ginkel, 2012). Next, the first K singular values of the pooled $\mathbf{\Lambda}$ could be used for getting a measure for the total amount of explained variance. At present the theoretical properties of such a solution have not been derived nor investigated. Consequently, it is currently unknown how closely such an estimate of the proportion of explained variance resembles the proportion of explained variance that would have been obtained if the data had been complete.

In short, there are still things that remain to be developed and investigated regarding the pooling of estimates and statistics within PCA applied to multiply imputed data. This is more generally a problem of multiple imputation. Rubin (1987) provided only very general combination rules for statistical analyses that can be applied when a parameter estimate or a set of parameter estimates is tested for significance. For some statistics and analyses that do not directly fit into that framework, additional combination rules have been developed since Rubin (1987), but for other statistics and analyses, there is still work to be done regarding combination rules. Whenever applied researchers are interested in statistics or analyses for which no combination rules are available yet, they are often inclined to set aside multiple imputation as a method for handling their missing data altogether.

However, Van Ginkel et al. (2020) argue that even when combination rules for specific analyses and statistics are lacking, it may not always be harmful to use something ad hoc. Even without a theoretical justification, ad hoc solutions can still give a rough but reasonable indication of what the actual statistic would have been without missing data. Additionally, since PCA is usually (but not always) used without any statistical testing, one cannot draw erroneous conclusions as a result of type I or type II errors. Even when something as simple as averaging $\mathbf{\Lambda}$'s across imputed datasets is done, this will probably still give a good indication of how many of the components contribute substantially to the explained total variance and which do not.

In summary, when a PCA needs to be carried out on an incomplete dataset, multiple imputation may be a good tool to handle the missing data. Although pairwise deletion does not necessarily give worse results than multiple imputation, multiple imputation comes with many other advantages, such as all analyses being

applied to the dataset being comparable regarding sample size and cases being included in the analyses. Besides, pairwise deletion has the disadvantage that computational problems may occur. Estimates of component loadings in multiply imputed datasets can readily be computed using generalized Procrustes analysis. Other statistics in PCA may not have combination rules as of yet, but using some quick-and-dirty procedures may not be harmful for the given purposes of PCA.

References

- Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1), 122–148. <http://www.jstor.org/stable/2991288>
- Archer, C. O., & Jennrich, R. I. (1973). Standard errors for rotated factor loadings. *Psychometrika*, 38(4), 581–592. <https://doi.org/10.1007/BF02291496>
- Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and em methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate behavioral research*, 35(3), 321–364. https://doi.org/10.1207/S15327906MBR3503_03
- Borgonovi, F., & Pál, J. (2016). *A framework for the analysis of student well-being in the pisa 2015 study*. 140. <https://doi.org/10.1787/5jlpszwghvnb-en>
- Boulton, M. J., Bucci, E., & Hawker, D. D. (1999). Swedish and English secondary school pupils' attitudes towards, and conceptions of, bullying: Concurrent links with bully/victim involvement. *Scandinavian Journal of Psychology*, 40(4), 277–284. <https://doi.org/10.1111/1467-9450.404127>
- Buisman, R., Pittner, K., Tollenaar, M. S., Lindenberg, J., van den Berg, L., Compier-de Block, L., van Ginkel, J. R., Alink, L., Bakermans-Kranenburg, M. J., Elzinga, B. M., & van IJzendoorn, M. H. (2020). Intergenerational transmission of child maltreatment using a multi-informant multi-generation family design. *PLoS One*, 15(3), e0225839. <https://doi.org/10.1371/journal.pone.0225839>
- Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of “Eckart-Young” decomposition. *Psychometrika*, 35(3), 283–319. <https://doi.org/10.1007/BF02310791>
- Chatterjee, S. (1984). Variance estimation in factor analysis: An application of the bootstrap. *British Journal of Mathematical and Statistical Psychology*, 37(2), 252–262. <https://doi.org/10.1111/j.2044-8317.1984.tb00803.x>
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Psychology Press. <https://doi.org/10.4324/9781315827506>
- 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 (Methodological)*, 39(1), 1–22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
- Dryden, I. L., & Mardia, K. V. (2016). *Statistical shape analysis: With applications in R*. Wiley.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185. <https://doi.org/10.2307/2289144>
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap* (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9780429246593>
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4), 507–521. <https://doi.org/10.2307/2331838>
- Furr, R. M. (2018). *Psychometrics: An introduction* (3rd ed.). Sage.
- Girshick, M. A. (1939). On the sampling theory of roots of determinantal equations. *The Annals of Mathematical Statistics*, 10(3), 203–224. <https://doi.org/10.1214/aoms/1177732180>

- Gower, J. C. (1971). Statistical methods of comparing different multivariate analyses of the same data. In F. R. Hodson, D. G. Kendall, & P. Tautu (Eds.), *Mathematics in the archaeological & historical sciences* (pp. 138–149). Edinburgh University Press.
- Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1), 33–51. <https://doi.org/10.1007/bf02291478>
- Green, B. F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, 17, 429–440. <https://doi.org/10.1007/BF02288918>
- Green, P. J. (1981). Peeling bivariate data. In V. Barnett (Ed.), *Interpreting multivariate data* (pp. 3–19). Wiley.
- Grung, B., & Manne, R. (1998). Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 42(1), 125–139. [https://doi.org/10.1016/S0169-7439\(98\)00031-8](https://doi.org/10.1016/S0169-7439(98)00031-8)
- Harman, H. H. (1976). *Modern factor analysis*. University of Chicago Press.
- Harshman, R. A. (1970). Foundation of the PARAFAC procedure: Models and condition for an “explanatory” multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16, 1–84.
- Hills, P., Francis, L. J., & Robbins, M. (2005). The development of the Revised Religious Life Inventory (RLI-R) by exploratory and confirmatory factor analysis. *Personality and Individual Differences*, 38(6), 1389–1399. <https://doi.org/10.1016/j.paid.2004.09.006>
- Josse, J., Husson, F., & Pagès, J. (2009). Gestion des données manquantes en Analyse en Composantes Principales. *Journal de la société française de statistique*, 150(2), 28–51. http://www.numdam.org/item/JSFS_2009__150_2_28_0/
- Josse, J., Pagès, J., & Husson, F. (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, 5(3), 231–246. <https://doi.org/10.1007/s11634-011-0086-7>
- Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, 62(2), 251–266. <https://doi.org/10.1007/BF02295279>
- Kiers, H. A. L. (2004). Bootstrap confidence intervals for three-way methods. *Journal of Chemometrics*, 18(1), 22–36. <https://doi.org/10.1002/cem.841>
- Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Wiley.
- Kroonenberg, P. M., & van Ginkel, J. R. (2012). Combination rules for multiple imputation in three-way analysis illustrated with chromatography data. *Current Analytical Chemistry*, 8(2), 224–235. <https://doi.org/10.2174/157341112800392544>
- Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1990). Assessing sampling variation relative to number-of-factors criteria. *Educational and Psychological Measurement*, 50(1), 33–48. <https://doi.org/10.1177/0013164490501004>
- Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research*, 26, 421–434.
- Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Kooij, A. J. (2007). Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological Methods*, 12(3), 359–379. <https://doi.org/10.1037/1082-989X.12.3.359>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley. <https://doi.org/10.1002/9781119013563>
- Lorenza-Seva, U., & Ferrando, P. J. (2003). IMINCE: An unrestricted factor-analysis-based program for assessing measurement invariance. *Behavior Research Methods, Instruments, and Computers*, 35, 318–321.
- Markus, M. T. (1994). Bootstrap confidence regions for homogeneity analysis; the influence of rotation on coverage percentages. In R. Dutter & W. Grossmann (Eds.), *Compstat*. Physica. https://doi.org/10.1007/978-3-642-52463-9_38
- Mata, I., Mataix-Cols, D., & Peralta, V. (2005). Schizotypal personality questionnaire-brief: Factor structure and influence of sex and age in a nonclinical population. *Personality and Individual Differences*, 38(5), 1183–1192. <https://doi.org/10.1016/j.paid.2004.08.001>
- Meulman, J. J. (1982). *Homogeneity analysis of incomplete data*. DSWO Press.

- Milan, L., & Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(1), 31–49. <https://doi.org/10.2307/2986193>
- Ogasawara, H. (2000). Standard errors of the principal component loadings for unstandardized and standardized variables. *The British Journal of Mathematical and Statistical Psychology*, 53(2), 155–174. <https://doi.org/10.1348/000711000159277>
- Ogasawara, H. (2002). Concise formulas for the standard errors of component loading estimates. *Psychometrika*, 67(2), 289–297. <https://doi.org/10.1007/BF02294847>
- Raykov, T., & Little, T. D. (1999). A note on procrustean rotation in exploratory factor analysis: A computer intensive approach to goodness-of-fit evaluation. *Educational and Psychological Measurement*, 59(1), 47–57. <https://doi.org/10.1177/0013164499591004>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Pearson.
- Takane, Y., & Oshima-Takane, Y. (2003). Relationships between two methods for dealing with missing data in principal component analysis. *Behaviormetrika*, 30(2), 145–154. <https://doi.org/10.2333/bhmk.30.145>
- Ten Berge, J. M. F. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42(2), 267–276. <https://doi.org/10.1007/BF02294053>
- The Three-Mode Company Home Page. (2021). Retrieved 15 October 2021, from <https://three-mode.leidenuniv.nl/>
- Timmerman, M. E., Kiers, H. A. L., & Smilde, A. K. (2007). Estimating confidence intervals for principal component loadings: A comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology*, 60(2), 295–314. <https://doi.org/10.1348/000711006X109636>
- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3), 279–311. <https://doi.org/10.1007/BF02289464>
- Tucker, L. R. (1972). Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika*, 37(1, Pt. 1), 3–27. <https://doi.org/10.1007/BF02291410>
- Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9780429492259>
- Van Ginkel, J. R., & Kiers, H. A. L. (2011). Constructing bootstrap confidence intervals for principal component loadings in the presence of missing data: A multiple-imputation approach. *British Journal of Mathematical and Statistical Psychology*, 64(3), 498–515. <https://doi.org/10.1111/j.2044-8317.2010.02006.x>
- Van Ginkel, J. R., & Kroonenberg, P. M. (2014). Using generalized procrustes analysis for multiple imputation in principal component analysis. *Journal of Classification*, 31(2), 242–269. <https://doi.org/10.1007/s00357-014-9154-y>
- Van Ginkel, J. R., & Kroonenberg, P. M. (2017). Evaluation of multiple-imputation procedures for three-mode component models. *Journal of Statistical Computation and Simulation*, 87(16), 3059–3081. <https://doi.org/10.1080/00949655.2017.1355368>
- Van Ginkel, J. R., Kroonenberg, P. M., & Kiers, H. A. L. (2014). Missing data in principal component analysis of questionnaire data: A comparison of methods. *Journal of Statistical Computation and Simulation*, 84(11), 2298–2315. <https://doi.org/10.1080/00949655.2013.788654>
- Van Ginkel, J. R., Linting, M., Rippe, R., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, 102(3), 297–308. <https://doi.org/10.1080/00223891.2018.1530680>
- Van Wingerde, B., & van Ginkel, J. (2021). SPSS syntax for combining results of principal component analysis of multiply imputed data sets using generalized procrustes analysis. *Applied Psychological Measurement*, 45(3), 231–232. <https://doi.org/10.1177/0146621621990757>

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 licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence 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

Quantifying the Bias of Non-linear Equating and Score Transformations



Matthias von Davier and Brian Clausner

Abstract This paper shows that using non-linear functions for equating and score transformations leads to consequences that are not commensurable with classical test theory (CTT). More specifically, a well-known theorem from calculus shows that the expected value of a non-linearly transformed variable does not equal the transformed expected value of this variable. Translated to CTT this implies that the transformed observed test score does not have an unbiased expectation, i.e., is different from the transformed true score. In order to quantify the bias, second-order Taylor expansions are used in this work to show that non-linear equating and scale transformations do not only lead to variability of SEMs but also to predictable bias in the expected values of the transformed observed scores. In line with Lord's finding that is often described as "Equating is either unnecessary or impossible," this bias due to non-linear equating vanishes either for perfectly reliable tests, or if the equating function is indeed linear, i.e., the tests are congeneric.

9.1 Introduction

When test scores of a new test are transformed so that they can be compared with scores on an old test form or some reference score scale, both linear and non-linear functions may be applied. Linear functions have the advantage that transformations of sums of scores can be carried out either on the sum or on the components of the sum, and results are identical. Non-linear transformations may be considered if the distributions of the two forms to be compared differ substantially.

M. von Davier (✉)
Boston College, Chestnut Hill, MA, USA
e-mail: vondavim@bc.edu

B. Clausner
NBME, Philadelphia, PA, USA
e-mail: bclausner@nbme.org

The research presented in this paper explores the effects of using non-linear functions for transforming fallible, that is, not perfectly reliable, test scores. In particular, it is a well-known effect of non-linear functions used in scale transformation that the conditional standard error of measurement will be affected differentially. One of the simplifying (but not necessary) assumptions of CTT is a common error variance, which is also a customary assumption in more general linear models. Once a non-linear transformation is applied, this equal error variance assumption no longer holds (e.g., Feldt and Qualls, 1998; Kolen et al., 1992; Woodruff et al., 2013), and conditional standard errors of measurement (CSEM) need to be estimated. The results presented here are in line with the research around CSEM in that they are based on the same assumptions, in particular when looking at CSEM estimates based on first-order Taylor expansions (sometimes referred to as delta method). However, we take these results to the next level by looking at higher-order terms and show that non-linear equating and scale transformations do not only lead to variability of SEMs but also to bias in the expected values of the transformed observed scores. This bias only vanished for perfectly reliable tests, as well as for linear transformation and equating functions.

A basic assumption of any valid scale transformations would be that they (if correctly specified) are functions that map the true scores on the source scale (e.g., the new test form) onto the true score of the target scale (the old test form). If a perfectly reliable test score of a person u is transformed, it should be mapped on the true score of the target scale. This corresponds to the true equating function. If there would be the ability of observing a test taker u 's scores independently over and over on both test forms, the true score on the new scale T_n should map to the true score on the old test form T_o for any test taker, if the two tests are indeed exchangeable.

However, this is no longer true if tests are not perfectly reliable and non-linear equating or transformations are used. This potentially more consequential effect will be derived mathematically and exemplified empirically in the subsequent sections: The expected value of an observed score, given a true score on the new form, does not equal the transformed true score under non-linear transformations.

The present paper shows how this conditional bias due to unreliability of the test scores can be estimated using standard results from calculus. An example that uses an approximate continuous non-linear function based on a concordance between ACT and SAT shows the order of magnitude for different score levels as they are affected by non-linearity and different levels of reliability.

9.2 Notation

Let Ω denote a population of test takers, and let $u \in \Omega$ denote a randomly drawn test taker from this population. Let X_n and X_o denote random variables (test scores) defined on Ω . We will use X_f with $f \in \{o, n\}$ to denote both new and old test forms at once when introducing notation and derived variables from now on. We will assume these test scores are real-valued and defined on a compact interval $X_n, X_o \in \mathbb{R}$ and

will consider functions $g: \mathbb{R} \mapsto \mathbb{R}$ that map the scores on the new form onto the interval on which the old form scores are defined. For each test taker u , assume there is a distribution $\phi_n''(X)$ and $\phi_o''(X)$ with means $E(X_n | u)$ and $E(X_o | u)$ and conditional variances $V(X_n | u)$, $V(X_o | u)$. This assumption allows for a potentially different distribution of test scores for each test taker. Across test takers $u \in \Omega$, one can define

$$E(X_f) = \int_{\Omega} \pi(u) E(X_f | u) du \quad (9.1)$$

as the marginal mean of test form $f \in \{o, n\}$ and

$$E[V(X_f | u)] = \int_{\Omega} \pi(u) V(X_f | u) du \quad (9.2)$$

as the average conditional variance (or marginal measurement error), and

$$V[E(X_f | u)] = \int_{\Omega} \pi(u) [E(X_f | u) - E(X_f)]^2 du \quad (9.3)$$

the variance of conditional means. Together, these variance components establish the total variance

$$V(X_f) = E[V(X_f | u)] + V[E(X_f | u)] \quad (9.4)$$

for test scores X_n and X_o of test forms $f \in \{o, n\}$.

9.3 Classical Test Theory: Which Tests Can Be Equated?

The assumptions of CTT can be expressed in terms of conditional expectations of observed scores and deviations (errors) from these expectations. Note that CTT can be viewed as first-order IRT (Holland and Hoskens 2003). Also note that a slightly stronger set of assumptions commensurate with CTT related no-DIF and monotonicity assumptions leads to models that are equivalent to IRT models (von Davier, 2017). Many of the basic desiderata for test quality are the same between IRT and CTT: Item scores are expected to be monotonically increasing with increasing trait level and are expected to be unaffected by variables other than the trait level. In CTT the trait level of person u is called the true score, $T(u)$ or T_u , while in IRT, the trait level is referred to as ability or latent variable θ_u . The true score is the expectation of the sum of item scores given u , that is, $T_f(u) = E(X_f | u)$, where old and new test forms $f \in \{o, n\}$ are considered. Lord (1980) argues that tests

can be equated if for any two test-takers $u, w \in \Omega$, the following equivalency holds:

$$T_o(u) = T_o(w) \leftrightarrow T_n(u) = T_n(w). \quad (9.5)$$

In words, two tests can be equated if any two respondents $u, w \in \Omega$ who have the same true score (conditional expectation) on the new test form also have the same true score (conditional expectation) on the old test form and vice versa. This is one of Lord's (1980) conditions that are prerequisite for two tests to be equatable; in addition, the conditional distributions of the error around the true score should be the same for equated (i.e., transformed from source test) observed scores and observed (on target test) scores, and this should be population independent.

9.4 Effects of Linear and Non-linear Transformations

When aiming at providing a transformation that produces equivalent values for the old (X_o -based) observed and true score variables using the new (X_n -based) score, the goal is to come as close as possible to the above equivalency of true scores. One way to achieve this is to require that a transformation g of sum score X_n to an equivalent score $g(X_n) = \tilde{X}_o$ should have the property

$$E[g(X_n | u)] = E[T_n + e_n | u] = T_o(u) = E(X_o | u). \quad (9.6)$$

Or, at the level of true scores, one could postulate $g[T_n(u)] = T_o(u)$ for all $u \in \Omega$. It appears that both properties are desirable. The transformed true score $g[T_n(u)]$ should equal the corresponding true score on the target scale $T_o(u)$ for all test takers u . If there was no error variable, the tests are perfectly reliable, and this property would be the only one to consider. However, we only have observed scores X_f at our disposal, so we would also want to aim at the criterion that the expected value of the transformed observed scores $E[g(X_n | u)]$ should equal the true score on the target scale as well. It turns out both conditions are equivalent properties as long as function $g()$ is linear as will be shown below.

9.4.1 The Linear Case

If the transformation $g()$ is linear, there are constants A, B with

$$\tilde{X}_o = g(X_n) = AX_n + B \quad (9.7)$$

and by plugging in the true score plus error decomposition, we find

$$f(T_n + e_n) = AT_n + Ae_n + B \quad (9.8)$$

and finally, for all $u \in \Omega$ we have

$$E[f(X_n|u)] = AE(T_n | u) + AE(e_n | u) + B = AE(T_n | u) + B \quad (9.9)$$

and hence

$$E[g(X_n | u)] = g[E(X_n | u)] = g[T_n(u)] = T_o(u). \quad (9.10)$$

In the linear case, either one of the above conditions implies the other. In addition, the transformed true score and error variable are uncorrelated. This can be easily verified since

$$V_\Omega(g(X_n)) = A^2[V_\Omega(T_n) + V_\Omega(e_n) + 2cov_\Omega(T_n, e_n)] \quad (9.11)$$

and $cov_\Omega(T_n, e_n) = 0$ (see Appendix A) yields the result

$$V_\Omega(\tilde{X}_o) = A^2[V_\Omega(T_n) + V_\Omega(e_n)]. \quad (9.12)$$

9.4.2 Non-linear Transformations and CTT

For non-linear functions $g()$, assume that the expected (true) score on the new form is mapped to the expected (true) score on the old form (Lord, 1980). Then we have $g(T_n(u)) = T_o(u)$ for all $u \in \Omega$. Under mild conditions, the transformation or equating function $g()$ can be split into a linear and a non-linear part (von Davier, 2008), a result implied directly by the definition of the first derivative $g'(x)$ as a limit of differences, as a “linear-part” function of a differentiable function can be defined by $g(x)$ is $h_{x_0}(x) = g'(x)(x - x_0) + g(x_0)$ for some constant x_0 (e.g., Forster, 1984). Then, $g(x) - h_{x_0}(x)$ is the “non-linear” part. Taking this argument further, standard calculus results imply that if $g()$ is k -times differentiable, with $g^{(l)}$ denoting the l^{th} derivative with respect to x , we can write $g(x) = g_u^{[k]}(x) + O(k)$ with a function

$$g_u^{[k]}(x) = \sum_{l=0}^k \frac{g^{(l)}(T)}{l!} (x - T)^l \quad (9.13)$$

that approximates $g()$ around some point $T \in \mathbb{R}$. This function is the well-known Taylor series (Taylor, 1715) approximation of $g()$ of order k . The remainder $O(k)$ denotes a quantity that is negligible in a neighborhood of T . Setting $T = T_n(u)$ and

inserting $X_n(u) = T_n(u) + e_n(u)$ into the equation yields

$$g_u^{[k]}(X_n(u)) = T_o(u) + \sum_{l=1}^k \frac{g^{(l)}(T_n(u))}{l!} (e_n(u))^l \quad (9.14)$$

Looking at the non-linear part using this Taylor expansion around $T_n(u)$ shows that the transformed score cannot be split in the true score on the transformed scale and an error that does not depend on $T_n(u)$. Examining

$$e_o^*(u) \approx \sum_{l=1}^k \frac{g^{(l)}(g^{-1}(T_o(u)))}{l!} (e_n(u))^l \quad (9.15)$$

and changing variables using the invertible true equating functions g and g^{-1} for mapping $T_o(u)$ onto $T_n(u)$, and vice versa, illustrate the issue.

Very often, the first-order Taylor series is quite informative and can be used to approximate the variance of the transformed quantity based on the first derivative (e.g., Wolter, 2007). In the case of a transformed test score, an approximate expression of the conditional error variance is derived by using the first-order Taylor series which is the linear function

$$g_T^{[1]}(x) = g(T) + g^{(1)}(T) * (x - T). \quad (9.16)$$

Then, the approximate conditional variance of $e_o^*(u)$ can be calculated as

$$V(e_o^* | u) \approx g^{(1)}(X_n(u)) V(e_n | u) \quad (9.17)$$

using the observed score $X_n(u)$ as a plug-in estimator of the true score $T_n(u)$ and assuming the higher-order Taylor terms to be negligible. However, unless g is linear (which it is not by the non-linearity assumption made in this section), there are derivatives of order $l > 1$ in the expression that are non-vanishing.

9.4.2.1 Effects on the Expectation of the Transformed Error Term

As shown above, the first-order term of the Taylor series approximation can be used to derive an approximate expression for the variance of the transformed variable, i.e.,

$$V(e_o^* | u) \approx g^{(1)}(X_n(u)) V(e_n | u). \quad (9.18)$$

This is an exact expression if $g()$ is linear and then $g^{(1)}(X_n(u)) = A_{T_n(u)}^2$. However, if $g()$ non-linear, higher-order Taylor terms are non-vanishing. Using the second-order (quadratic) term of the Taylor series, $\frac{1}{2}g^{(2)}(T_n(u)) [e_n(u)]^2$, we find

that the expectation of the transformed error term may not be equal to zero. More specifically, we have

$$E \left(\frac{1}{2} g^{(2)}(T_n(u)) [e_n(u)]^2 \right) = \frac{1}{2} g^{(2)}(T_n(u)) V(e_n(u)) \neq 0 \tag{9.19}$$

whenever $\frac{1}{2} g^{(2)}(T_o(u)) \neq 0$ and $V(e_n(u)) > 0$. Again, this expression is exact if $g(\cdot)$ contains only polynomial terms of first and second order; the expectation of the transformed error may be affected by additional terms if the non-linear transformation $g(\cdot)$ contains higher-order terms.

9.4.2.2 Jensen’s Inequality

After going through all this work, a colleague¹ who still remembers his advanced math education (in contrast to us) mentioned another well-known theorem as an alternative, and potentially simpler, explanation for this bias of non-linear equating functions. However this theorem is more general and cannot be directly applied to estimate the magnitude of the bias, for which we still need the Taylor series approximation.

Jensen’s (1906) inequality states that for [non-linear, locally] convex functions, the expected value of the transformed observed score can be shown to be larger than the transformed expectation. More specifically, for convex $g(\cdot)$, we have

$$E [g(X_n | u)] > g[E(X_n | u)] = g[T_n(u)] = T_o(u). \tag{9.20}$$

and if $g(\cdot)$ is concave, we have

$$E [g(X_n | u)] < g[E(X_n | u)] = g[T_n(u)] = T_o(u). \tag{9.21}$$

This means that the expectation of the transformed score is larger (smaller) than the transformed true score. For a concave equating function, the Jensen gap is positive (larger than zero), and for concave equating functions, the expected transformed score has a negative Jensen gap.

9.5 Examples

The linear case discussed above is one example where the transformed observed score can be decomposed in transformed true score and an error term with vanishing expectation and variance that is given by the customary equations for transformed

¹ Braun (2021), personal communication.

linear variables. Another example is polynomials; we will use and study the consequences of the case where the second-order term is non-vanishing.

9.5.1 Quadratic Forms

If $g(X) = aX^2 + b$, i.e., the transformation is a quadratic function. Assume $a \neq 0$, and $V(e_n^2 | u)$ for all $u \in \Omega$. Then, the conditional expectation of the transformed observed score is given by

$$E(g(X_n) | u) = [aT_n^2(u) + b] + a2T_n E(e_n | u) + aE(e_n^2 | u) \quad (9.22)$$

and further

$$E(g(X_n) | u) = g(T_n(u)) + aV(e_n^2 | u) \quad (9.23)$$

since

$$E(e_n^2 | u) = V(e_n | u) + [E(e_n | 0)]^2 \quad (9.24)$$

while $E(e_n | u) = 0$. Note that this implies

$$E(g(X_n) | u) \neq g(T_n(u)) = T_o(u) \quad (9.25)$$

since $aV(e_n^2 | u) \neq 0$. However, we have

$$g(E(X_n | u)) = g(T_n(u)) = T_o(u). \quad (9.26)$$

In consequence, we have a result that shows the error term of the transformed to have an expectation that differs from zero and hence that the expected value of the transformed observed score is unequal to the transformed true score. In summary, by setting

$$e_o^*(u) = a2T_n(u)e_n(u) + ae_n^2(u) \quad (9.27)$$

and noting that $T_o(u) = g(T_n(u))$, we obtain

$$g(X_n(u)) = g(T_n(u) + e_n(u)) = T_o(u) + e_o^*(u) \quad (9.28)$$

where

$$E(e_o^*(u)) = aV(e_n^2 | u) \neq 0. \quad (9.29)$$

Table 9.1 Score concordance between ACT and SAT as provided in ACT (2009)

ACT	SAT	diff	ACT	SAT	diff
36	1600	40	23	1070	40
35	1560	50	22	1030	40
34	1510	50	21	990	40
33	1460	40	20	950	40
32	1420	40	19	910	40
31	1380	40	18	870	40
30	1340	40	17	830	40
29	1300	40	16	790	50
28	1260	40	15	740	50
27	1220	40	14	690	50
26	1190	30	13	640	50
25	1150	40	12	590	60
24	1110	40	11	530	

9.5.2 ACT-SAT Concordance

The College Board provides score transformation tables referred to as *concordances* for the two US college entrance exams ACT² and SAT (e.g., ACT, 2009; Dorans, 1999). These tables translate a discrete score variable obtained on one of the tests and provide an equivalent score on the other test. While it needs to be pointed out that both tests are not exchangeable in the sense that they do not exactly measure the same skills, the composite score on the ACT and the combined verbal and math score on the SAT do correlate at $r_{\text{ACT, SAT}} = 0.92$ (Dorans, 1999). The following table provides the concordance between ACT and SAT as reported in ACT (2009).

It can be seen in Table 9.1 the differences between ACT scores are one point, whereas the differences between transformed SAT scores are 40 points for most adjacent scores, while some differences are 50 points, in particular for the upper and lower score regions. Hence the conversion is slightly non-linear. The graphical representation of the concordance shows that the deviations from linearity are small.

In this example, we use the ACT score as the X_n variable and the SAT as the X_o so that the task is to transform ACT scores onto the SAT score scale. Just for the purpose of this exercise, we use the s.d. of the ACT provided by Dorans (1999), $S(\text{ACT}) = 4.86$, and calculate using as the lower bound of the reliability the

² The acronyms ACT and SAT are household names in the USA as well as for many international students. ACT stands for “American College Test,” and SAT has no meaning as an acronym. The SAT acronym originally stood for “Scholastic Aptitude Test,” but as the test evolved, the acronym’s meaning was dropped (<https://blog.collegeboard.org/difference-between-sat-and-psat>).

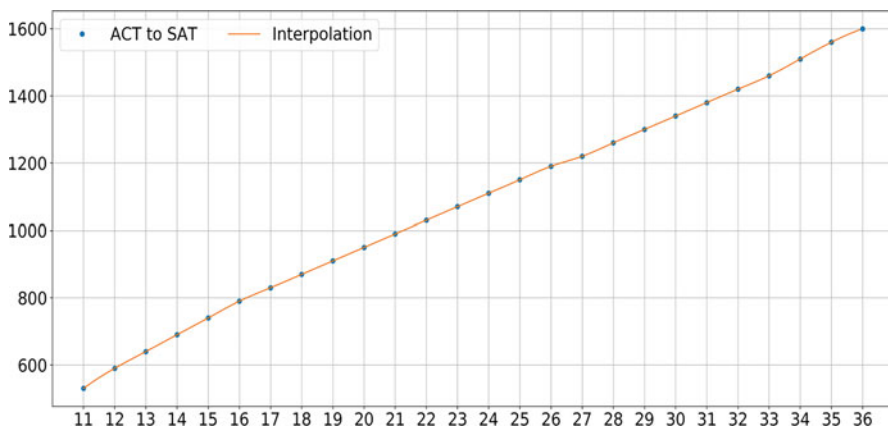


Fig. 9.1 Continuation of the concordance table on the interval $\Omega_{\text{ACT}} = [11, 36]$ using natural cubic splines

correlation between ACT and SAT given as $r_{\text{ACT, SAT}} = 0.92$ an upper boundary of the marginal error variance for the ACT as

$$S^2(e_{\text{ACT}}) = (1 - 0.92) 4.86^2 \approx 1.89.$$

For higher levels of reliability, a lower error variance is obtained, and the bias of the non-linear transformation is reduced, as will be demonstrated below. For the example at hand, we base explorations on the publicly available concordance Table 9.1 and use a reverse-engineered approximate continuous function in this exercise.

We use a commonly applied smoothing method to continuize the concordance Table 9.1. This continuation using natural cubic splines results in a smooth function defined on the interval $\Omega_{\text{ACT}} = [11, 36]$ and depicted in Fig. 9.1. It is not important at this point whether we have found the “correct” continuization, as the point to be made is based only on the observation that the concordance table is not perfectly fit by a linear model as is easily verified by the variability of the differences between the transformed values on the SAT scale in Table 9.1. The differences of the transformed SAT scale scores vary between 30 and 60 points, with most being 40 points, and hence force variability in the slope of any continuous transformation function that goes through the discrete points provided in the concordance table.

The function depicted in Fig. 9.1 is twice continuously differentiable; in fact it is infinite times piecewise differentiable so that a second-order Taylor approximation can be used to derive expressions for approximate variance of the error of the transformed SAT score scale. As indicated above, ACT (2009) reports a standard deviation of $s_{\text{ACT}} = 4.86$, and for subsequent calculations, it will be assumed that the error variance is the same across the ACT score range. Consider three different levels of ACT score reliability, so that the implied error variances are 1.89, 1.18, and 0.47, respectively (Table 9.2).

Table 9.2 Different levels of marginal error variance for three different levels of reliability assumed for the ACT

rel_{ACT}	$V_{ACT}(e)$
0.92	1.89
0.95	1.18
0.98	0.47

The three plots in Fig. 9.1 show the bias of the non-linear scale transformation based on a local Taylor series approximation that uses the function values across a score range of 3, which includes three ACT score points of the concordance table centered around the transformed score, plus the interpolated values from the second closest score points for half the interval.³ This choice ensures that the changes in slope are traced locally in the approximate Taylor series approach used here. The continuization and the local Taylor approximation were conducted using readily available tools from Python’s `scipy` package (Jones et al., 2001).

The simulation of observed scores is using the CTT score decomposition $X = T + e$ and generates observed scores based on the error variance levels given above and the 26 ACT scores ranging from 11 to 36. The simulations were conducted 20 times per level of reliability with a sample size of 5200 (200 error terms per each of the 26 observed ACT score levels) per simulated database; the simulations were conducted using a Python script that can be made available by the authors upon request.

In the figures, the bias based on a simulated observed score variable (50 data sets with $26 \times 200 = 5200$ observed scores) with one of the three different levels of reliability is given as well as the approximate Taylor-based estimate of the bias of the transformed observed score, calculated as the difference of the expected transformed observed score from the transformed expected (true) score.

Why this effect can indeed be regarded as a bias of the non-linear transformed score becomes evident by comparing different levels of reliability. With increasing reliability, the difference vanishes; in other words the bias becomes larger with decreasing reliability of the scale score that is to be transformed.

It is evident from Fig. 9.2 that the bias of the transformed observed score is reduced as the reliability of the new test form score increases. Also, the bias is smaller in the middle of the score range where the function, as verified by the concordance Table 9.1 and Fig. 9.1, deviates less from linearity than in the extremes of the score range.

³ Three score points are located within a closed interval of length 2 on the ACT scale, and 5 points are contained in a closed interval of 4.

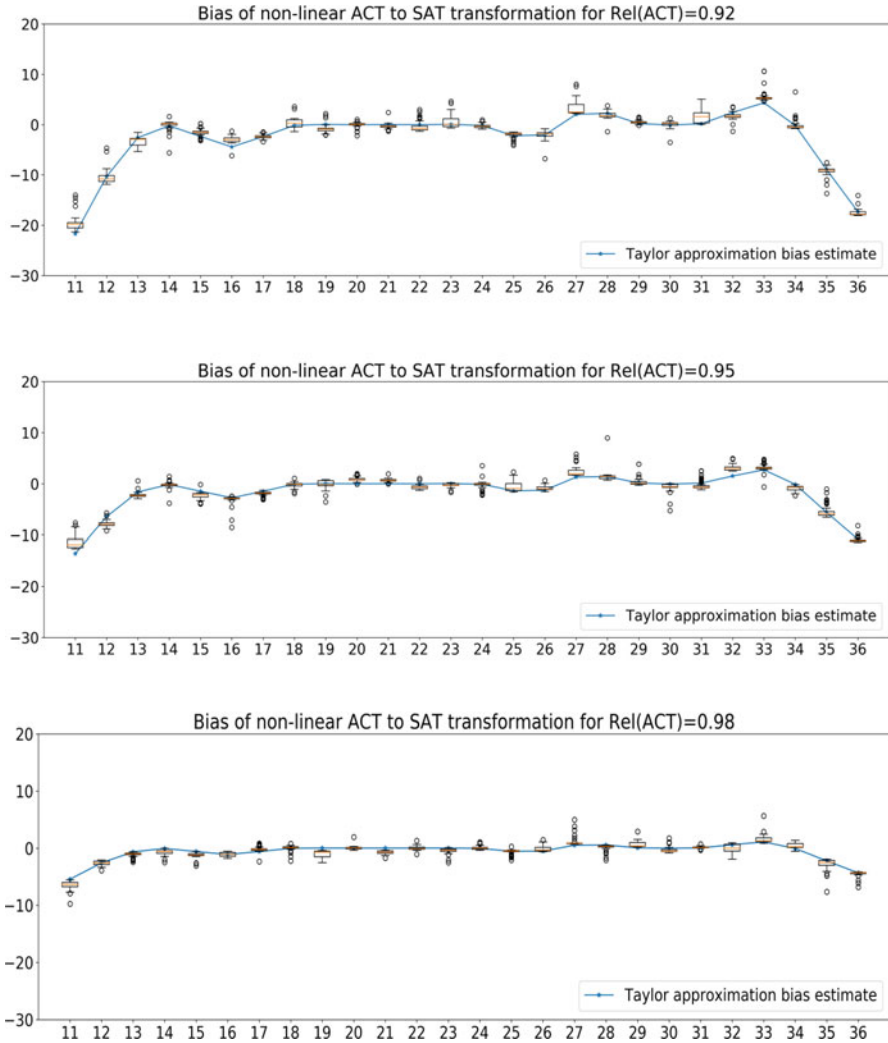


Fig. 9.2 Graphs of estimated bias, based on second-order Taylor approximation and simulation based using a marginal error variance of $V_{ACT}(e) = 1.89, 1.18, 0.47$ that corresponds to a reliability of 0.92, 0.95, 0.98 (figures in that order)

9.6 ITED-ACT Concordance

The example presented in this section uses an approximate transformation that was derived from graphs displayed in an article by Yin et al. (2004) that compares linear and equi-percentile concordances between the Iowa Test of Educational Development (ITED) and the ACT.

We are examining the concordance between ACT (M), a 60-item test, and ITED (QT), a 40-item test as given on page 283 of Yin et al. (2004). It appears that psychometric quality indicators are not readily available for the ITED; therefore, we rely on numbers taken from a somewhat seasoned publication: We assume a reliability of 0.88 for ITED (QT) and a standard deviation of 31 (e.g., Hendricks, 1967). The main point here is not critiquing any particular testing program but to show a general effect of non-linear equating and scale transformations using reasonable numbers for quantities such as the standard error of measurement. We used an interval of 8 ITED points which again corresponds to including 3 value pairs at a minimum, plus adjacent points through the spline interpolation. Since there is no tabular or functional form available, the Google Chrome web application WebPlotDigitizer (Rohatgi, 2015) was used to transform the concordance plots into numerical form that can be processed with spreadsheets and statistical software. For the linear and the equi-percentile concordance, about 90 value pairs were digitized from the graphs on page 283 in Yin et al. (2004) and subsequently smoothed using a 5-point moving average with weights (1,2,3,2,1) to reduce jitter caused by the step function character of the equi-percentile transformation as well as by inaccurate digitizing. This smoothing will overall reduce the reported bias, but will also avoid reporting bias that is caused by errors in digitizing. The linear concordance was shown not to produce bias and is hence expected to only show the level of error introduced by digitizing a linear function by manually collecting value pairs with WebPlotDigitizer.

In addition to the linear and the equi-percentile tables, a third approximate concordance was generated using a piecewise linear function with one change in slope. The value pairs for this function were also smoothed with the weighted moving average as described above. Figs. 9.3, 9.4, and 9.5 show the results first for the linear (but manually digitized) concordance, second for the equi-percentile concordance, and third for the piecewise linear (smoothed) concordance.

The smoothed digitized linear concordance does not appear to deviate from a straight line. However, the approximate Taylor series-based bias as well as the bias estimate using transformed observed scores do show some small variability around zero. This is likely to be due to the manual digitalization process that introduced slight variations in the slope of the digitized function.

Figure 9.4 shows the same exhibits for the digitized equi-percentile concordance. The linear case may serve as a baseline here for comparison so that the magnitude of the bias terms of this non-linear concordance can be put in perspective.

A visual inspection of the equi-percentile concordance, even in its smoothed version, shows regions of rapid changes in slope in the extremes of the score distribution. As seen in the corresponding regions in the bias plot, these changes in the slope are also associated with larger biases of transformed observed scores. An increase in slope is associated with a positive bias, whereas a decrease is associated with a negative bias of the transformed score.

Figure 9.5 shows this effect in an exemplary constructed concordance that only involves one region of increase in slope and otherwise uses a constant slope before and after the (non-linear) region of change. Using this simple setup, we can focus

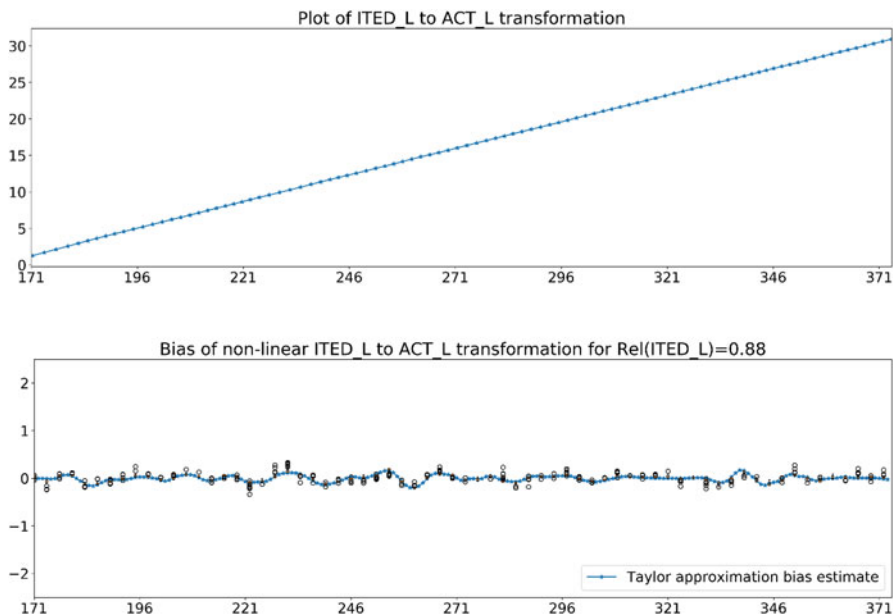


Fig. 9.3 Smoothed linear concordance using digitized values from Yin et al. (2004), p. 283, and Taylor approximate bias as well as empirical bias using simulated observed scores. The simulated ITED observed scores were generated for every third ITED score level and correspond to a reliability of 0.88

on how the non-linearity introduces bias in the transformed scores unless these are perfectly reliable.

In Fig. 9.6, we zoom in to a score range where the piecewise linear transformation depicted in Fig. 9.5 shows large bias values in positive and negative directions due to two changes of the slope in different directions and within a short score interval. It can be seen from the rectangular magnification of this crucial section that the approximate bias calculated based on the second order Taylor term closely matches the estimated bias based on simulated observed scores.

This last example shown in Figs. 9.5 and 9.6 demonstrates that the bias indeed vanishes in the linear regions but is obvious and positive in the score region where the slope changes, i.e., where the transformation is non-linear. Note that in Figs. 9.4 and 9.5, there are regions where the bias exceeds one point on the ACT scale. This indicates that the use of transformations with regions of non-linearity may indeed have unintended consequences in score concordance and equating applications.

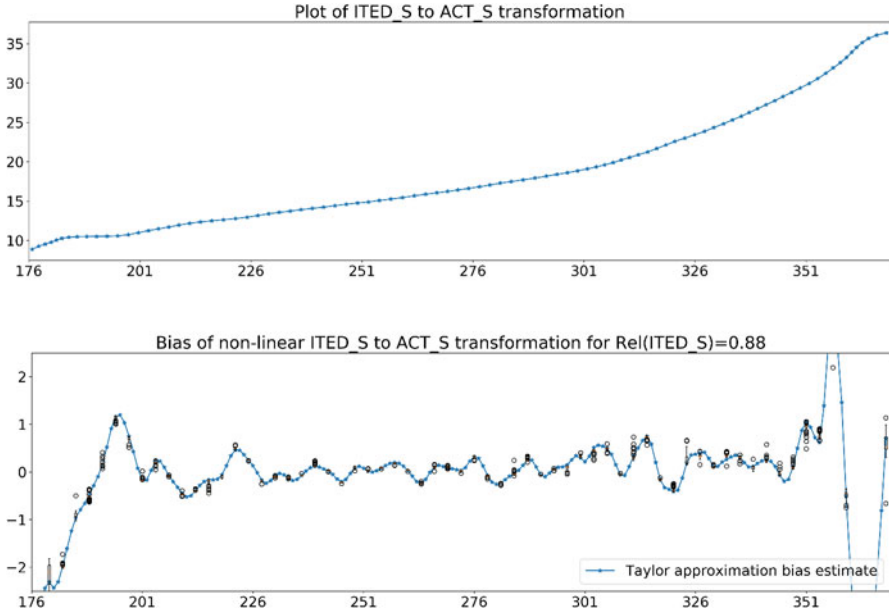


Fig. 9.4 Smoothed equi-percentile concordance using digitized values from Yin et al. (2004), p. 283, and Taylor approximate bias as well as empirical bias using simulated observed scores. The simulated ITED observed scores correspond to a reliability of 0.88

9.7 Discussion

This paper shows that linear transformations of observed scores retain the central building blocks of classical test theory, if these held with respect to the original test scores.

First, as it was shown in past research that if there is variance homogeneity in the error variable, non-linear transformations produce a dependency of the conditional error variance on the true score, this is a well-known result that is used together with the Taylor approximation to derive conditional standard errors of measurement (e.g., Feldt and Qualls (1998); Kolen et al., 1992; Woodruff et al., 2013).

Second, when using non-linear transformations in order to perform a scale transformation or an equating from a new test form X^n to an old test form X^o , the transformed sum of true score and error may no longer exhibit a feature central to CTT, namely, that error terms on the transformed scale are vanishing in expectation. If an additive decomposition as used in CTT is performed, the resulting error variable has no longer a zero expectation on the transformed scale. This introduces a bias into the transformed observed scores that is mainly driven by lack of reliability. For perfectly reliable measures, the bias vanishes, but for tests that have non-vanishing error variances, there is bias that can be well approximated by the second-order term of a local Taylor approximation of the non-linear transformation function.

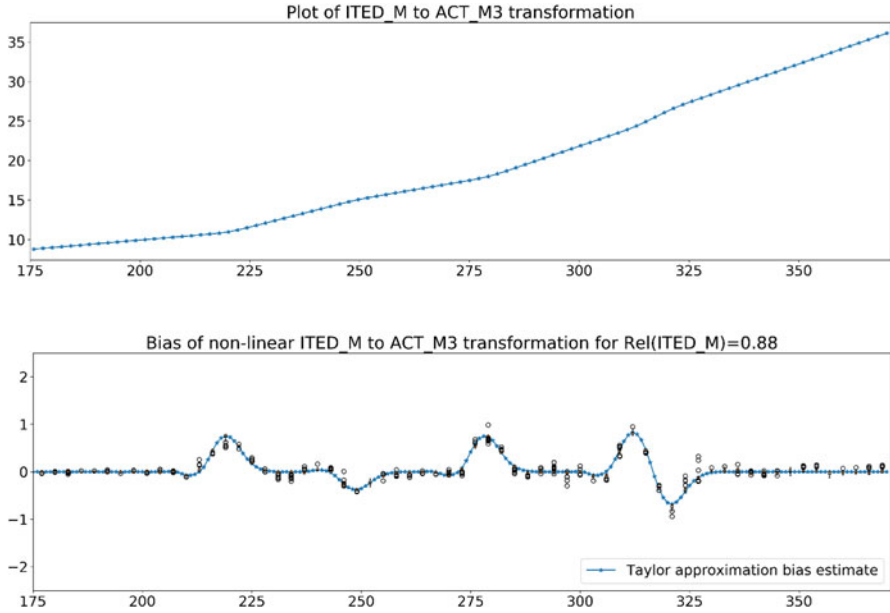


Fig. 9.5 Smoothed piecewise linear concordance using digitized values from Yin et al. (2004), p. 283, and Taylor approximate bias as well as empirical bias using simulated observed scores. The simulated ITED observed scores were generated for every third ITED raw score and correspond to a reliability of 0.88

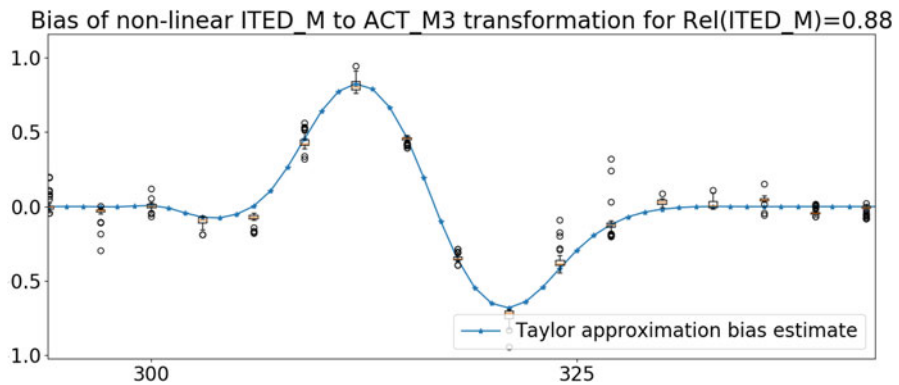


Fig. 9.6 Detail of the Taylor approximate bias function as well as empirical bias using simulated observed scores for the piecewise linear transformation depicted in Fig. 9.5

How to deal with this bias is an obvious question. Can it be ignored, or are there ways around it? Here is a list of suggestions that are ordered in the order of preference of the authors, but not necessarily by practicality:

1. Produce and use only tests with (close to) perfect reliability for equating and scale transformations.
2. Use only linear transformations for equating, concordance tables, and scaling.
3. Reduce non-linear transformation bias by developing an approximate bias correction.

Note that bias corrections are a well-researched topic in ML estimation, including IRT modeling (Warm, 1989; Firth, 1993; von Davier, 1995), and can be readily applied. However, it appears that the awareness that non-linear transformation of scale scores produce biased is not widespread. When researching results on non-linear functions of test scores, the only reference we came across was Lockwood and McCaffrey (2015) who do discuss bias when using test scores that relate to outcome variables in non-linear ways. In their application, they look at polynomial terms in regression models, for example, of the value-added model type, but do not utilize the expressions developed in this article for quantifying the bias associated with non-linear transformations.

In summary, if transformed scale scores are expected to follow central results of classical test theory, only certain types of variable transformations appear to be permissible. In particular, if the error variance was assumed to be homogeneous, a non-linear transform introduces conditional error variances that vary across the score range. Maybe even more concerning is that non-linear scale transformations introduce bias in the expected value of the transformed scores, as shown in this paper. The magnitude of this bias depends on the reliability of the scale that undergoes transformation as well as on the local changes in slope of the transformation function. This bias can be approximated by standard methods and could be approximately corrected for (using second derivatives around the observed score rather than the unobserved true scores). However, given that the foundation of CTT is linear in true score and error, one could consider linear transformations the natural domain of permissible transformations. Given that the bias vanishes for linear transformations as well as for perfectly reliable scales, it appears that linear scale transformations should be the preferred ones for equating, concordances, and scale transformations.

References

- ACT, Inc. (2009). *ACT-SAT concordance table*. <https://research.collegeboard.org/sites/default/files/publications/2012/7/researchnote-2009-40-act-sat-concordance-tables.pdf>
- Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (College Board Research Reports 99-01). The College Board. <https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/j.2333-8504.1999.tb01800.x>
- Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education*, 11(2), 159–177. https://doi.org/10.1207/s15324818ame1102_3
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38. <https://doi.org/10.2307/2336755>

- Forster, O. (1984). *Analysis 1. Differential- und Integralrechnung einer Veränderlichen* [Calculus 1. Univariate differential and integral calculus]. Vieweg & Sohn.
- Hendricks, J. (1967). *The Iowa tests of educational development as predictors of academic success at Utah State University*. All Graduate Theses and Dissertations. <https://doi.org/10.26076/9f3e-393b>
- Holland, P. W., & Hoskens, M. (2003). Classical Test Theory as a first-order Item Response Theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68, 123–149. <https://doi.org/10.1007/BF02296657>
- Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes [On convex functions and inequalities between mean values]. *Acta Mathematica*, 30(1), 175–193. <https://doi.org/10.1007/BF02418571>
- Jones, E., Oliphant, T., & Peterson, P. (2001). *SciPy: Open source scientific tools for Python* [Computer Software]. <http://www.scipy.org/>
- Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, 29(4), 285–307. <http://www.jstor.org/stable/1435086>
- Lockwood, J. R., & McCaffrey, D. F. (2015). Should nonlinear functions of test scores be used as covariates in a regression model? In R. W. Lissitz & H. Jiao (Eds.), *Value added modeling and growth modeling with particular application to teacher and school effectiveness* (pp. 1–36). Information Age Publishing.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Rohatgi, A. (2015). *Web plot digitizer* [Computer Software]. <https://automeris.io/WebPlotDigitizer>
- Taylor, B. (1715). *Methodus incrementorum directa et inversa* [Direct and inverse increments method]. William Innys.
- von Davier, M. (1995). *Winnmira user manual*. Chapter on person parameter estimation using WLE. IPN: Kiel University. <http://208.76.80.46/~svfklumu/wmira/winnmiramanual.pdf>
- von Davier, A. (2008). New results on the linear equating methods for the non-equivalent-groups design. *Journal of Educational and Behavioral Statistics*, 33(2), 186–203. <http://www.jstor.org/stable/20172112>
- von Davier, M. (2017). CTT and No-DIF and ? = (Almost) Rasch Model. In M. Rosén, K. Yang Hansen, & U. Wolff (Eds.), *Cognitive abilities and educational outcomes: A Festschrift in Honour of Jan-Eric Gustafsson* (pp. 249–272). Springer. https://doi.org/10.1007/978-3-319-43473-5_14
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450. <https://doi.org/10.1007/BF02294627>
- Wolter, K. (2007). *Introduction to variance estimation* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-35099-8>
- Woodruff, D., Traynor, A., Cui, Z., & Fang, Y. (2013). *A comparison of three methods for computing scale score conditional standard errors of measurement* (ACT Research Report No. 2013-7). American College Testing Program. <https://files.eric.ed.gov/fulltext/ED555593.pdf>
- Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement*, 28(4), 274–289. <https://doi.org/10.1177/0146621604265034>

Chapter 10

Examination of Test Characteristics’ Effect on Coefficient α and Coefficient ω



Terry Ackerman, Ye Ma, and Richard Luecht

Abstract In this study, five factors were simulated to determine their effect on three measures of reliability: coefficient α , coefficient ω , and the true scale reliability as defined in a classical test theory context as the ratio of true score variance over observed score variance. The factors were the number of items, the level of item discrimination, the number of dimensions, the correlations among dimensions, and the location of the items in relationship to the latent ability score distribution. In all higher-order dimensional conditions, simple structure was assumed. The data were generated using the multidimensional item response theory compensatory two-parameter logistic model. As expected, when the number of items, the magnitude of the item discriminations, and the correlations among the dimensions increased, the reliability correspondingly increased. Noticeable differences were observed across all higher dimensionality conditions with ω values being significantly lower than α , a finding which could have been an artifact of the simulated simple structure.

10.1 Background

Reliability is one of the hallmark measures of an assessment’s quality. It is a necessary condition for validity. Several authors have noted that a test’s reliability is a function of the scores on a test, not the test itself or multiple forms of a test (Brennan, 2001; Thompson & Vacha-Haase, 2000). There are a host of measures

T. Ackerman (✉)
University of Iowa, Iowa City, IA, USA
e-mail: terry-ackerman@uiowa.edu

Y. Ma
Amazon Web Services (AWS), Chicago, Illinois, USA
e-mail: ymcheryl@amazon.com

R. Luecht
University of North Carolina at Greensboro, Greensboro, NC, USA
e-mail: rmluecht@uncg.edu

that have been developed to estimate reliability (Feldt & Brennan, 1989; Kane, 1996). The basic definition of reliability is based on the classical test theory assumption that for individual i , a test score X_i is the sum of two unobservable and uncorrelated components, T_i , a true score and measurement error, E_i :

$$X_i = T_i + E_i. \quad (10.1)$$

Reliability is then defined as the squared correlation between the observed test scores and the corresponding unobserved true scores which can be shown to be equal to the ratio of true score variance, σ_T^2 , to total observed score variance, σ_X^2 :

$$\rho_{TX}^2 = \frac{\sigma_T^2}{\sigma_X^2} \quad (10.2)$$

As noted by Sijtsma (2009a, b), over the years, the one standard reliability index that researchers and psychologists have adopted is coefficient alpha (Cronbach, 1951), further referred to as α . Although Cronbach's name is tied to the statistic, this measure can be traced through the works of Kuder and Richardson (1937), who published a version of α for dichotomous items—the KR-20 coefficient. Hoyt (1941) proposed an equivalent statistic using of analysis of variance with dichotomous responses.

Finally, Guttman (1945) derived a series of reliability coefficients. One coefficient, denoted as λ_3 , was equivalent to α .

Assuming a test composed of J -items, where a random variable, Y_j , represents a score on item j , and the total score on the test for an examinee is defined as the sum,

$$X = \sum_{j=1}^J Y_j, \quad (10.3)$$

α for a group of examinees can be expressed as

$$\alpha = \frac{J}{(J-1)} \left[1 - \frac{\sum_{j=1}^J \sigma_{Y_j}^2}{\sigma_X^2} \right] \quad (10.4)$$

where $\sigma_{Y_j}^2$ represents the item variances and σ_X^2 is the variance of the total scores.

If the item scores are standardized, the formula for α can be expressed in terms of the mean of the inter-item correlations, $\bar{\rho}$; that is,

$$\alpha = \frac{J\bar{\rho}}{1 + (J-1)\bar{\rho}}, \quad (10.5)$$

or equivalently as the average of the inter-item covariances, $\overline{\sigma_{YY}}$,

$$\alpha = \frac{J \left(\frac{\overline{\sigma_{YY}}}{\sigma_X^2} \right)}{1 + (J - 1) \left(\frac{\overline{\sigma_{YY}}}{\sigma_X^2} \right)}. \quad (10.6)$$

It should be noted that α also approximates the mean of all possible Spearman-Brown split-half coefficients (Spearman, 1910; Brown, 1910) where the split-half coefficients, r_{12} , are adjusted, pairwise Pearson product-moment correlations between the two half-test scores:

$$r_{\text{split-half(SB)}} = \frac{2r_{12}}{1 + r_{12}}. \quad (10.7)$$

Coefficient α equals the mean of the split-half coefficients when the standard deviations of all possible halves are equal and smaller when the standard deviations are heterogeneous (Cortina, 1993). Feldt and Brennan (1989) and Lord and Novick (1968) further noted that α will be equal to the mean of all split-half correlations when the split-half correlations are calculated by the Flanagan-Rulon formula:

$$r_{\text{split-half(FR)}} = \frac{4r_{12}s_1s_2}{s_T^2}, \quad (10.8)$$

where s_1 and s_2 are the standard deviations of each half and s_T^2 is the variance of the total test (Flanagan, 1937; Rulon, 1939).

Many researchers have criticized the pervasive use of α (Green, et al., 1977; Green and Yang, 2009; Rodriguez & Maeda, 2006; Sijtsma, 2009a, b) or even wrote about the shortcomings of the statistic and its interpretations (Cronbach & Shavelson, 2004; Ten Berge & Socan, 2004). One drawback is the ubiquitous interpretation of α as a measure of internal consistency. Internal consistency is a characteristic of the test items, not the test, and does not reflect the length of the test (i.e., the pattern of inter-item covariances). Another caveat is that calculations of α can yield values that are outside the range of possible values of the score reliability that should be derivable from a single test administration (Cho & Kim, 2015; Sijtsma, 2009a).

It is often thought that α requires the test to be unidimensional and that it can be used as a measure signifying the degree of multidimensionality. Cronbach (1951) did address the test dimensionality issue when he wrote that for a test:

to be interpretable, . . . it is not essential that all the items be factorially similar. What is required is that a large proportion of the test variance be attributable to the first principal factor running through the test.

Several authors have noted that multidimensional tests can exhibit high values of α (Davenport, et al., 2015; Davison & Davenport, 2015). When a test has been empirically demonstrated to be multidimensional, it is important the test developer

be able to articulate the meaning of the composite scale which α is characterizing (e.g., that the total test score is a weighted linear composite of two or more subscores by design). In any case, it has been well documented that a multidimensional test does not necessarily have a lower α than a unidimensional test.

Friedman and Weisberg (1981) demonstrated that if all the inter-item correlations are positive, the first principal component eigenvalue is approximately proportional to the average correlation of the J items

$$\lambda_1 \approx 1 + (J - 1)\bar{r}. \quad (10.9)$$

Using this relationship, α can be approximated as

$$\alpha \approx \frac{J\bar{r}}{\lambda_1}. \quad (10.10)$$

Another approach that tries to capture the underlying possibly multidimensional nature is to assess reliability using a factor-analytic approach such as coefficient ω_h (McDonald, 1985, 1999; Zinbarg et al., 2005), further referred to as ω_h . The subscript h denotes that this measure of reliability is derived from the hierarchical factor analytic model. That is, it is assumed that all items measure a common factor that accounts for a major proportion of variance in the scaled scores. In addition, it is assumed that each item measures a unique skill uncorrelated with the common scale. For the purposes of this study, we used a bifactor model in which all items load on a general factor and on a unique factor. All unique factors are uncorrelated. The ω_h statistic used is calculated as

$$\omega_h = \frac{\left(\sum_{j=1}^J \lambda_{gj}\right)^2}{\sigma_X^2}, \quad (10.11)$$

where λ_{gj} are the factor loadings on the general factor.

The goal of this research is to examine and compare the performance of α and ω_h under several different test conditions including the correlations between dimensions, number of items, discrimination power of the items, and whether the difficulty of the items is optimal given the ability distribution of the examinees.

The response data were generated using the compensatory multidimensional two-parameter IRT model (M2PL) (Reckase, 2009). The M2PL can be expressed as

$$p_j(\theta) = P(u_j = 1|\theta) = \frac{1}{1 + e^{-(\sum_{k=1}^m a_{jk}\theta_{ik} + d_j)}}, \quad (10.12)$$

where $\theta = (\theta_1, \theta_2, \dots, \theta_k, \dots, \theta_m)$ is a m -length vector of the latent scores with elements indexed as θ_{ik} (the score of person i on dimension k), a_{jk} is a discrimination for item j on dimension k , respectively, and d_j is an intercept term denoting the composite difficulty of each item. The MDISC index is the multidimensional analog

to unidimensional discrimination parameter, a . It is a composite discrimination index for each that can be expressed as

$$\text{MDISC}_i = \sqrt{\sum_{k=1}^m a_{jk}^2} \quad (10.13)$$

where a_{jk}^2 is defined above.

10.2 Research Design

This is a simulation study. The response data were generated under prescribed testing conditions with multiple replications. Three coefficients were computed for each data set and then the comparative results aggregated across replications: (i) ρ_{TX}^2 , the true scale reliability when the true score and error variances are known (through simulation), (ii) α (Eq. 10.4), and (iii) ω_h (Eq. 10.11). This design demonstrates how logically influential test design considerations such as test length, item discrimination, and the homogeneity of items relative to the population mean(s) impact those three reliability coefficients. The study included five completely crossed design factors:

- Number of items ($J = 24, J = 48$)
- Levels of MDISC (low MDISC, 0.4–0.8; moderate MDISC, 0.8–1.2; high MDISC, 1.2–1.6)
- Number of dimensions ($m = 1, 2, 3, 4$)
- Location of mean item difficulty ($d = 0, 1$) given the examinee distribution will always be centered at the origin
- Correlation of abilities ($\rho = .0, .5$)

The sample size for each simulation was fixed at 1000 randomly generated examinees sampled from a standard normal univariate or multivariate normal distribution centered at the origin for each simulation. Each condition was further replicated 100 times to provide empirical sampling distributions of each reliability coefficient for comparative purposes.

10.3 Reliability Estimation and Evaluation

Three reliability coefficients were calculated for each of the simulated data sets: the true scale reliability, ρ_{TX}^2 , coefficient α , and ω_h (based on a fitted bi-factor model). As noted earlier, the true scale reliability was calculated using Eq. 10.1 where the

true score variance is the variance of the expected scores of the N -examinees over J -items:

$$\sigma_T^2 = \sigma^2 \left(\sum_{i=1}^N \sum_{j=1}^n P(u_i = 1 | \theta_{i1}, \theta_{i2}, a_{j1}, a_{j2}, d_j) \right) \quad (10.14)$$

using the generated $N \times m$ matrix, θ , and the $J \times (m + 1)$ matrix of generated item parameters. The raw score variance is calculated using the total score for each person and including all the items in the test. The α and ω_h were calculated using the corresponding functions in the R package **psych** (Revelle, 2021). That package calculates the three reliabilities given in Eqs. 10.4 and 10.11. In aggregate, there were 96 conditions ($2 \times 3 \times 4 \times 2 \times 2$), and each condition was replicated 100 times to provide empirical sampling distributions of the three coefficients. In particular, the means and standard deviations of those sampling distributions were computed across the 100 replications per condition, and graphical visualizations were created using the R package **ggplot2** (Wickham, 2016). All the simulations, data management, and analytical aspects of this study were carried out using R (R Core Team, 2021).

10.4 Results

The 5 design factors produced 96 simulation test design conditions. These factors were expected to have direct or indirect impact on the three reliability indices, ρ_{TX}^2 , α , and ω_h . The impact of the number of items (test length) on reliability is well-known given the extensive body of research on the Spearman-Brown formula (e.g., Angoff, 1953; Traub, 1997),

$$\rho_{XX'}^* = q\rho_{XX'} / [1 + (q - 1)\rho_{XX'}] \quad (10.15)$$

where $\rho_{XX'}$ is the original reliability index and q is the ratio of new to original (old) test lengths. In contrast, the average MDISC (composite item discrimination) and item location were generated to either *offset* or *match* to the population centroids' impact the contribution of each item to the score variance (e.g., Gulliksen, 1950). These two factors also directly and indirectly reflect on item quality—especially the item discrimination parameters and MDISC, which act as weights for the latent scores. Finally, the number of underlying dimensions and the correlation between those dimensions represent the dispersion of the measurement *signal* across the apparent latent structures representing the item covariances. Including these latter two conditions in the simulation directly speaks to the motivation for ω_h , that is, to have a reliability index that responds to untended or idiosyncratic dimensionality, or to a test that includes multiple dimensions by design and perhaps reports the total score as weighted linear composite of subscores. Increasing the dimensionality

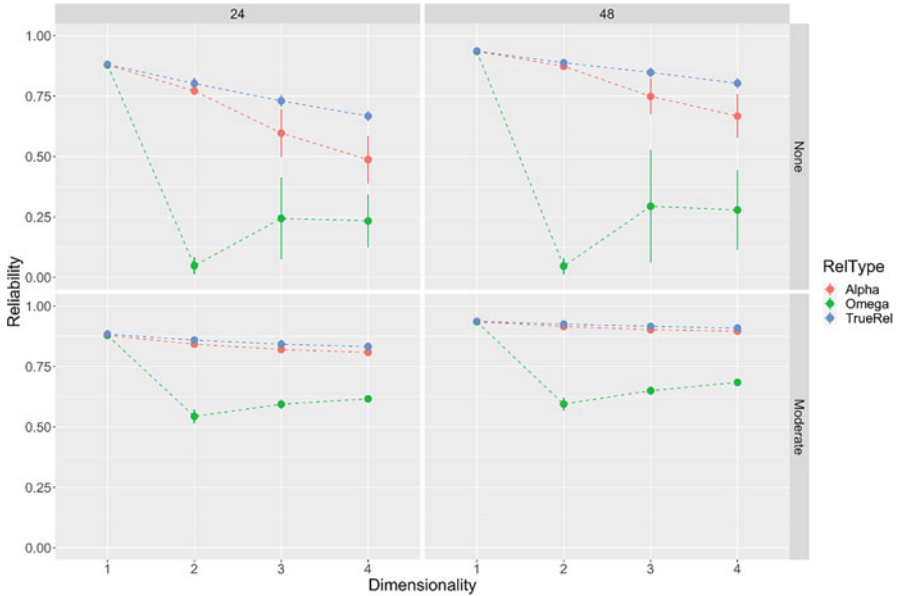


Fig. 10.1 Summary of reliability coefficients for high MDISC and item difficulty matched to the population proficiency score centroids: $\mu(d) - \mu(\theta_k) = 0$ (100 replications per condition)

and covariance(s) among the underlying factors should disperse the “measurement signal” relative to a reported total score.

For the most part, these factors produced results that met expectations. Figures 10.1, 10.2, 10.3, 10.4, 10.5 and 10.6 include “trellis” or faceted multi-plots that embed a bivariate plot conditioned on the number of items (columns) and the magnitude of correlation between the underlying dimensions or factors (*none* implies a zero correlation between the factors; *moderate* implies a correlation of .5 between all factors). The number of dimensions is shown along the horizontal axis for each plot, and the vertical axis represents the magnitude of the correlation. The three plotted outcomes in each cell of the multi-plot denote the three reliability indices: ρ_{TX}^2 , α , and ω_h . These results are summarized as the mean and standard error of the reliability coefficients across 100 replications per combination of simulation conditions.

As Fig. 10.1 shows (high MDISC, with the mean item difficulty matched to the population centroids, $\mu(d) - \mu(\theta_k) = 0$ for all k), there is a noticeable increase in the ρ_{TX}^2 and α coefficients as the test length increased from 24 to 48 items, and a decrease in the coefficients as the number of dimensions increased from 1 up to 4 due to the amount of total test score signal dispersion among the dimensions. The three coefficients are all highly similar in the unidimensional case ($m=1$) with α and ω_h essentially being identical. The coefficients only start to decline as the total score signal is dispersed across two or more underlying factors. Note that the

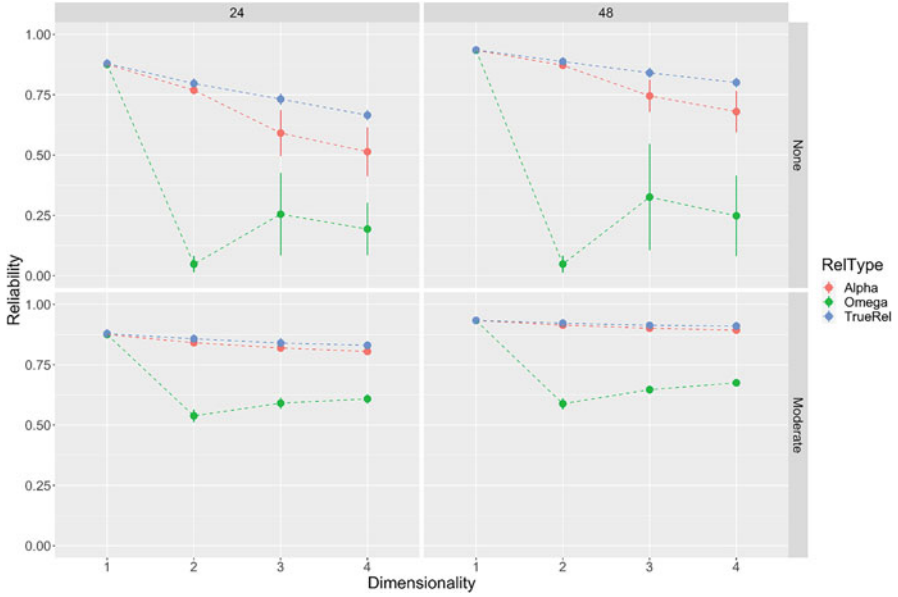


Fig. 10.2 Summary of reliability coefficients for high MDISC with item difficulty offset from the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 1$ (100 replications per condition)

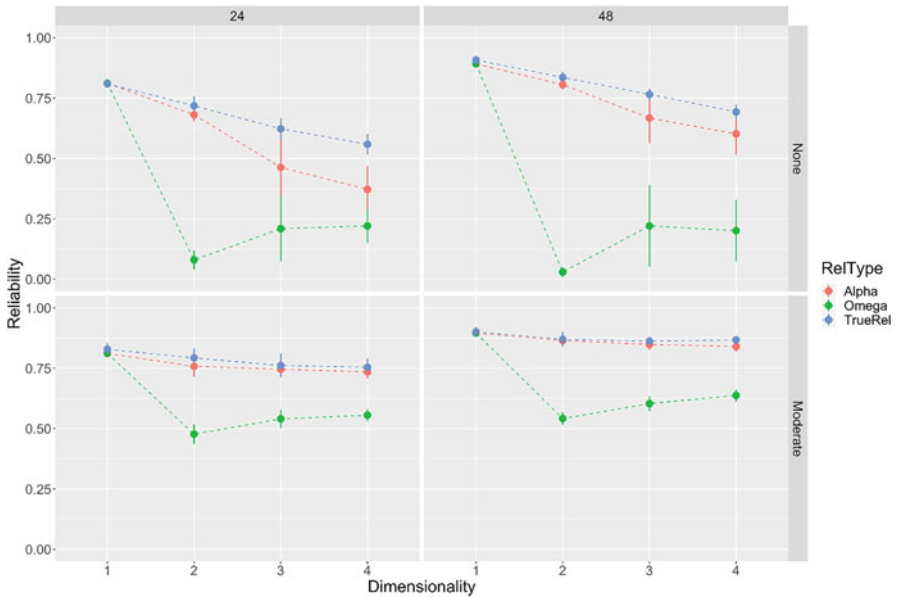


Fig. 10.3 Summary of reliability coefficients for moderate MDISC and item difficulty matched to the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 0$ (100 replications per condition)

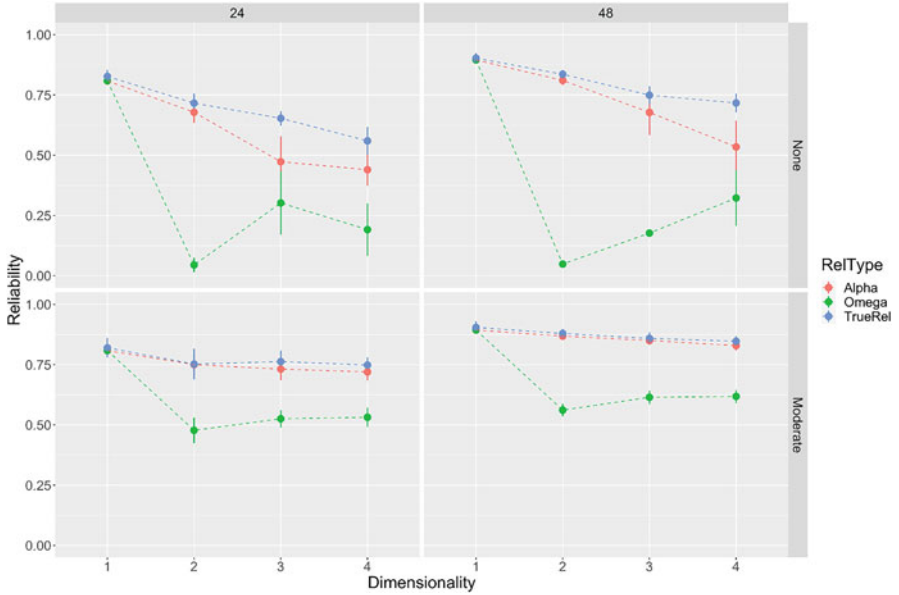


Fig. 10.4 Summary of reliability coefficients for moderate MDISC with item difficulty offset from the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 1$ (100 replications per condition)

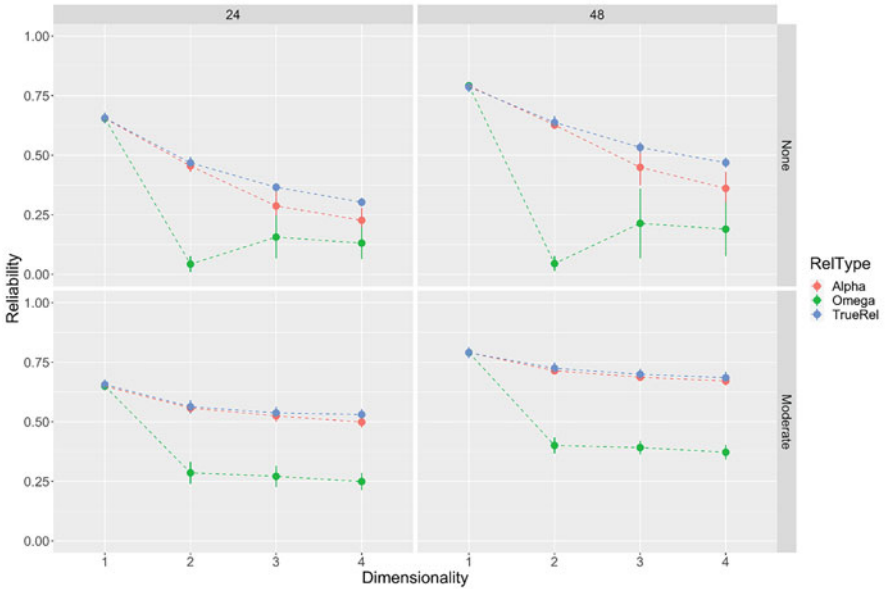


Fig. 10.5 Summary of reliability coefficients for low MDISC and item difficulty matched to the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 0$ (100 replications per condition)

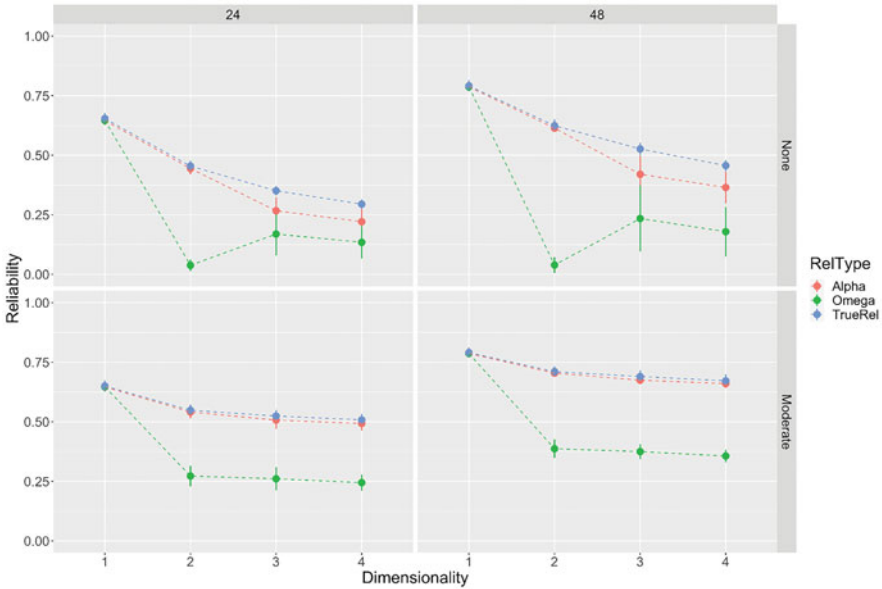


Fig. 10.6 Summary of reliability coefficients for low MDISC with item difficulty offset from the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 1$ (100 replications per condition)

zero-correlation condition is rather unrealistic in a practical sense¹, but provides a reasonable baseline under “maximum dispersion” conditions. Interestingly, the mean values of ω_h tend to somewhat track with the inter-factor correlations (.0 = none or .5 = moderate).

Figure 10.2 (high MDISC, with the mean item difficulty offset from the population centroids, $\mu(\theta_k) - \mu(d) = 1$ for all dimensions) shows a pattern that is very consistent with Fig. 10.1. Cronbach’s α values tend to be smaller than the “true reliabilities” with known true scores, ρ_{TX}^2 . This likely reflects some sampling error when estimating the item error variances (see Eq. 10.3). The ω_h coefficients, again, somewhat track with the magnitude of the inter-factor correlations, although the mean values are also confounded by the high MDISC present in the items.

Figure 10.3 (moderate average MDISC, with the mean item difficulty matched to the population centroids, $\mu(\theta_k) - \mu(d) = 0$ for all dimensions) begins to show an interesting pattern where the mean α and ρ_{TX}^2 values respond to the reduced composited item discrimination, but the ω_h coefficients do not.

Figure 10.4 (moderate average MDISC, with the mean item difficulty offset from population centroids, $\mu(\theta_k) - \mu(d) = 1$ for all dimensions) confirms the coefficient

¹ In practice, it would be very rare to encounter a test designed to measure two or more underlying traits with NO covariance between the traits. Even tests measuring distinctly different traits like mathematics and English language arts tend to positively correlate in the moderate range.

patterns of Fig. 10.3; that is, the ω_h coefficients respond more to the amount of total score signal dispersion than to the reduced composite item discrimination. The mean α and ρ_{TX}^2 values respond to the reduced composited item discrimination and, to a lesser degree, to the signal dispersion across dimensions.

Figures 10.5 and 10.6 show an overall decline in mean α and ρ_{TX}^2 values proportional to both the low average MDISC values and the dimensional dispersion of the total score signal. Interesting, and similar to Figs. 10.3 and 10.4, the latter dispersion has less impact across the increasing number of dimensions than under the high discrimination condition. Increasing the test length helps to somewhat offset the decline in the reliability coefficients, but the recommendation to write high-quality items and monitor that the level of composite item discrimination remains as high as possible seems to be good advice.

10.5 Conclusion

In this study, we varied testing conditions that we felt would influence the performance of the three reliability coefficients: (1) true reliability, (2) Cronbach's α , and (3) ω_h . As the number of items was doubled from 24 to 48, there was the expected proportional increase in reliability. Likewise, as the discrimination of the items, MDISC, increased, the magnitude of the reliability coefficients also unilaterally increased. The simulation response data were generated relative to an underlying multidimensional simple structure for three of the four simulation conditions. As the correlations between the multidimensional latent abilities increased from 0 to .5, thus "collapsing" the latent space—the reliability coefficients also proportionally increased. The effect of increasing the average difficulty of the items, that is, increasing the amount of offset between the location of maximum measurement information relative to the centroid of the examinee ability, joint latent distributions did not induce any prominent change in reliability.

The simulation condition that appeared to demonstrate the greatest impact on the reliability coefficients was multidimensionality. As the number of dimensions increased, coefficient ω dropped considerably in comparison to the true scale reliability and coefficient α . This was anticipated because ω_h was computed using the sum of the loadings on the general factor in the hierarchical, orthogonal bifactor model, where all factors are uncorrelated. Because the data were generated using simple structure, the loadings on the unique factors were higher than the loadings on the general factor, creating significant dispersion in the measurement "signal"—specifically, inducing "noise" relative to the general factor. That is, the R-packages that were used estimated ω_h using the bi-factor model versus a common factor or component model.

In the unidimensional case, α and ω were always equal. In some cases, these coefficients exceeded the true scale reliability. As dimensionality increased, α like the ρ_{TX}^2 decreased though not nearly as much ω . It appeared that α was not affected as much as ω_h by the increase in dimensionality. There was one notable

inconsistency. In the two-dimensional case, ω was consistently lower than in the three- and four-dimensional cases across all conditions. This may have been a function of the sampled item discrimination parameters.

It seems clear that testing practitioners must be advised always to conduct a thorough dimensionality analysis of their test results relative to the intended, reported score scale(s) and further evaluate the dimensionality analysis outcomes in terms of the test specification so that they can articulate the meaning of the observed score scale. Only evaluating a reliability coefficients or standard errors of measurement is not sufficient.

Future research will extend the current research to incorporate factorially complex item structures where the multidimensionality may relate to nuisance dimensions of idiosyncratic characteristics of the items (i.e., items loadings on both intended and unintended factors underlying the data). We also plan to examine reliability from a multidimensional IRT perspective and relate more directly to the concept of a unidimensional composite of intended multidimensional traits (i.e., Wang's (1985) reference composite). Lastly, we plan to experiment with the formulation of ω_h and determine if additional information about dimensionality and its effect on reliability can be delineated for testing practitioners.

References

- Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika*, *18*(1), 1–14. <https://doi.org/10.1007/BF02289023>
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *1904–1920*, *3*(3), 296–322. <https://doi.org/10.1111/j.2044-8295.1910.tb00207.x>
- Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295–317. <https://doi.org/10.1111/j.1745-3984.2001.tb01129.x>
- Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230. <https://doi.org/10.1177/1094428114555994>
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. <https://doi.org/10.1037/0021-9010.78.1.98>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. <https://doi.org/10.1177/0013164404266386>
- Davenport, E. C., Davison, M. L., Liou, P.-Y., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice*, *34*(4), 4–9. <https://doi.org/10.1111/emip.12095>
- Davison, M. L., & Davenport, E. C. (2015, April 15–19). *Coefficient α and dimensionality* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Chicago, IL, United States.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (pp. 105–146). Macmillan.
- Flanagan, J. C. (1937). A proposed procedure for increasing the efficiency of objective tests. *Journal of Educational Psychology*, *28*(1), 17–21. <https://doi.org/10.1037/h0057430>

- Friedman, S., & Weisberg, H. F. (1981). Interpreting the first eigenvalue of a correlation matrix. *Educational and Psychological Measurement*, 41(1), 11–21. <https://doi.org/10.1177/001316448104100102>
- Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 37(4), 827–838. <https://doi.org/10.1177/001316447703700403>
- Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, 74(1), 121–135. <https://doi.org/10.1007/s11336-008-9098-4>
- Gulliksen, H. (1950). *Theory of mental tests*. Wiley. <https://doi.org/10.1037/13240-000>
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–282. <https://doi.org/10.1007/BF02288892>
- Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6(3), 153–160. <https://doi.org/10.1007/BF02289270>
- Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education*, 9(4), 355–379. https://doi.org/10.1207/s15324818ame0904_4
- Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, 2(3), 151–160. <https://doi.org/10.1007/BF02288391>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- McDonald, R. P. (1985). *Factor analysis and related methods*. Psychology Press. <https://doi.org/10.4324/9781315802510>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press. <https://doi.org/10.4324/9781410601087>
- R Core Team. (2021). *R: A language and environment for statistical computing*. [Computer Software]. R Foundation for Statistical Computing, <http://www.R-project.org/>
- Reckase, M. D. (2009). *Multidimensional item response theory*. Springer. <https://doi.org/10.1007/978-0-387-89976-3>
- Revelle, W. (2021). *Psych R package* (Version 2.1.3) [Computer Software]. <https://personality-project.org/r/psych/>
- Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, 11(3), 306–322. <https://doi.org/10.1037/1082-989X.11.3.306>
- Rulon, P. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9(1), 99–103.
- Sijtsma, K. (2009a). On the use, misuse, and very limited usefulness of Cronbach's α . *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, 74(1), 169–173. <https://doi.org/10.1007/s11336-008-9103-y>
- Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, 3(3), 271–295. <https://doi.org/10.1111/j.2044-8295.1910.tb00206.x>
- 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(4), 613–625. <https://doi.org/10.1007/BF02289858>
- Thompson, B., & Vacha-Haase. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, 60(2), 174–195. <https://doi.org/10.1177/0013164400602002>
- Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16(4), 8–14. <https://doi.org/10.1111/j.1745-3992.1997.tb00603.x>
- Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT*. Unpublished manuscript, University of Iowa.
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. <https://doi.org/10.1007/978-3-319-24277-4>
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω_H : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 123–133. <https://doi.org/10.1007/s11336-003-0974-7>

Chapter 11

Methods for Estimating Conditional Standard Errors of Measurement and Some Critical Reflections



Wilco H. M. Emons

Abstract Educational assessments can have far-reaching consequences for individuals. To allow test users to make valid decisions, it is important to provide evidence about the uncertainties in the observed scores on which the individual decisions are based. In this chapter we examine standard errors of measurement defined for specific score groups, which are referred to as *conditional standard errors of measurement*. In particular, we study the foundations of the ANOVA method proposed by Feldt et al. (Appl Psychol Meas 9:351–361, 1985) within the context of classical test theory. In addition, we suggest some variations and study their practical usefulness including sample size requirements.

11.1 Methods for Estimating Conditional Standard Errors of Measurement and Some Critical Reflections

Educational tests are widely used to make decisions about individuals in various educational settings. Examples include low-stakes classroom situations in which the teacher has to make instructional decisions, as well as high-stakes placement decisions and decisions about whether or not to provide additional support such as remedial teaching. However, test scores have measurement errors and thus come with uncertainties. To allow test users to make valid decisions, it is important to provide evidence about the uncertainties in the observed scores on which the individual decisions are based (Hopster-Den Otter et al., 2019). Withholding information about the precision may give test users the impression that scores are more accurate than they are. There are different ways to incorporate precision in test score reports. For example, instead of only reporting a single value, one can report a confidence interval around the observed score to express the uncertainties involved

W. H. M. Emons (✉)

Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands
e-mail: W.H.M.Emons@tilburguniversity.edu

(Harvill, 1991; Sijtsma & Van der Ark, 2020). Test users can then make informed decisions whether the precision of the tests they want to use is sufficient for their application envisaged (Sijtsma & Emons, 2011). Information about score precision may also be used for, for example, setting cutoffs on selection tests (e.g., personnel selection) whereby taking into account the expected false positives and negatives due to measurement errors.

In practice, it is customary to express measurement precision using the *standard error of measurement* (SEM) from classical test theory (CTT; Lord & Novick, 1968). The SEM, denoted by σ_E , is defined as

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}, \quad (11.1)$$

where σ_X is the (population) standard deviation of the test scores X and $\rho_{XX'}$ is the (population) test-score reliability. In real data applications, σ_X is replaced by the observed standard deviation of test scores (S_X) obtained in a representative sample; and $\rho_{XX'}$ is replaced by an estimate of the test score reliability – most often coefficient (Cronbach's) α – obtained in that same sample. The resulting sample value for the SEM is then used as a measure of precision for all individuals to whom the test is applied. However, there are good reasons to doubt that the group-level SEM is appropriate for individual-level decisions at different points on the sum-score scale. It is generally acknowledged that measurement precision is person specific (Guttman, 1945; Lek & Van de Schoot, 2018; Lord & Novick, 1968) and that measurement precision varies along the score scale (e.g., Feldt et al., 1985). The SEM that results from Eq. (11.1) is actually the *average* precision by which the persons in the target population are measured (e.g., Lord, 1980; Mellenbergh, 1996). We use this average as our “best guess” for the individual-level precision.

Instead of using a single group-level SEM for all individuals, it would be more appropriate to assess the precision for each testee individually and use that for further decision making. Unfortunately, it is impossible to test individuals repeatedly under the same conditions, not even twice (Sijtsma, 2009). As a result, we do not have empirical access to these person-specific measurement errors. However, it is possible to assess the SEM for individuals within specific homogeneous subgroups. Previous research has suggested that measurement precision varies across ability levels, where the precision is typically higher at either endpoint of the score scale and smaller in the middle (Feldt et al., 1985). Therefore, the most obvious choice is to create groups of individuals having the same sum score. The group-specific SEM, which is the average precision for the individuals in the specific group, may serve as a more accurate predictor of the person-specific precision than an overall global average.

Standard errors of measurement defined for specific score groups are referred to as *conditional standard errors of measurement* (CSEM; Holland & Hoskens, 2003; Lee et al., 2000; Lek & Van de Schoot, 2018; Nicewander, 2019; Raju et al., 2007). However, because in subgroups with constant X we have $\sigma_X = 0$ and reliability $\rho_{XX'}$ undefined (Sijtsma & Van der Ark, 2020), obtaining the CSEMs is unfortunately more complicated than simply applying Eq. (11.1) in subgroups. Several well-

established methods for estimating CSEMs from a single administration have been proposed (see, for example, Brennan, 1998; Feldt et al., 1985; Lek & Van de Schoot, 2018; Qualls-Payne, 1992; Sijtsma & Van der Ark, 2020; Thorndike, 1951; Woodruff et al., 2013). These methods can be grouped into three classes: first, methods that split a single test into two or more parallel parts (Lek & Van de Schoot, 2018; Thorndike, 1951; Woodruff, 1990). The variances of the scores across parallel parts, taking into account differences in test length, serve as indicators of precision. A practical problem with these methods is that defining parallel halves often turns out to be unsuccessful. Failure to find parallel parts makes these methods questionable. The second class subsumes methods that use a probabilistic model for the item scores. Examples include the CSEMs based on a binomial or compound binomial model and those based on item response theory models (e.g., Lee et al., 2000). The validity of the CSEMs *derived* from a fitted IRT model is contingent on adequate model fit, and this fit must be ascertained everywhere along the trait scale where important decisions are made.

Third, and perhaps somewhat less well-known, are methods based on a two-way mixed (repeated measures) ANOVA (Feldt et al., 1985) and generalizability theory (Brennan, 2001). The ANOVA method in particular is an interesting method. As shown below, the method is intimately related to internal consistency estimates for the reliability, which places the method in a unified framework with lower-bound estimates for reliability. To be specific, Feldt et al.'s (1985) ANOVA method applied to a random (unconditional) sample would produce the same SEM as when using Eq. (11.1) with coefficient α substituting $\rho_{XX'}$. The ANOVA method in fact directly estimates the average measurement precision that *defines* coefficient α . This also means that there is no need to first compute α and then reconstruct the SEM. Being a direct expression of the average measurement precision, the ANOVA approach allows computing non-zero SEMs in groups in which everyone has the same sum score. Moreover, coefficient α is just one of a series of lower bounds to the reliability, suggesting that Feldt et al.'s (1985) approach can be generalized by using estimates of the measurement precision that are induced by other lower-bound reliability measures from the Guttman's λ -series.

The ANOVA method also has some convenient computational advantages, which makes it a particularly attractive method from a practical point of view. First, the method is easy to implement and is suitable for dichotomous, polytomous, or continuous scores. Second, there is no need to group the items into parallel parts. Third, the ANOVA method does not require a calibrated and fitting statistical model for the item responses. Of course there is no such thing as a free lunch. A possible drawback of the method is the fact that large numbers are required per score group to get stable estimates. However, to put things into perspective, given the impact educational assessments may have on people's lives, we expect test publishers to pay serious attention to this. Moreover, this is certainly not unique for the ANOVA method and applies also model-based methods such as IRT-based methods. However, and perhaps surprisingly, little is known about the minimal sample-size requirements for the ANOVA method in relation to, for example, item quality (i.e., difficulties and discrimination) and test length. More insight into

sampling requirements is needed to be able to weigh this method against other CSEM methods.

This chapter examines and elaborates on Feldt et al.'s ANOVA-based CSEM by connecting them to some well-known lower-bound reliability indices. This is done having three goals in mind: Enhance understanding of the ANOVA method and variants thereof, advancing the theoretical foundations for preferring one method over the other, and gain more insight into the practical possibilities and limitations of the method. To accomplish these goals, this chapter is organized as follows. First, we explain the key theoretical concepts. Second, we derive some general expressions for the CSEMs within the framework of lower-bound estimates for the reliability. Third, we present the results of simulations on the accuracy of the methods under varying conditions for sample size, test length, and item quality. Fourth, we present simulation results focusing on the added value of CSEMs within the context of individual change assessment. Finally, we discuss the results and limitations.

11.2 Theoretical Background

11.2.1 Measurement Precision Versus Reliability

CTT assumes that observed test scores are affected by random influences that are person and occasion specific (Lord & Novick, 1968). By implication, if we were able to test the same person repeatedly under identical conditions, we would observe a *distribution* of test scores. Lord and Novick (1968) refer to this hypothetical distribution as the *propensity distribution*. Every arbitrary person v in the population can be characterized by his or her own propensity distribution. The mean of the propensity distribution is the *true score*, denoted τ , which is the focal quantity in practical person measurement (see Borsboom, 2005, Chap. 2 for a critical conceptual discussion on true scores). The difference between the observed score X and τ is the error of measurement, denoted E . The variance of the individual test-score distribution reflects the error variance and thus also the person-specific measurement precision. Let $\sigma_{E_v}^2$ denote the person-level measurement error variance. The smaller the error variance $\sigma_{E_v}^2$, the more precise we can measure τ_v .

Ideally, we would like to estimate the person-specific precision for each testee individually, but, as noted above, that is practically impossible. Therefore, it is customary to use the standard error of measurement (SEM), which is the *average* precision in a representative sample from the focal population. Most textbooks explaining the SEM do so by using Eq. (11.1) as their starting point (e.g., Allen & Yen, 2002; Sijtsma & Van der Ark, 2020). This practice may unintentionally suggest that measurement precision is a property that *follows* from the test score reliability, but the opposite is actually the case. Reliability *follows* from the ratio of two *independent* properties: measurement precision (*within*-subject variability) and population heterogeneity (*between*-subject variability). In particular, test score

reliability can be expressed as

$$\rho_{XX'} = 1 - \frac{\mathcal{E}(\sigma_{E_v}^2)}{\sigma_X^2} = 1 - \frac{\mathcal{E}(\sigma_{E_v}^2)}{\sigma_T^2 + \mathcal{E}(\sigma_{E_v}^2)}, \tag{11.2}$$

(e.g., Lord & Novick, 1968; Mellenbergh, 1996). Eq. (11.2) shows that reliability is high if *on average* the within-subject variance is much smaller than between-subject differences in the true scores, σ_T^2 . Random within-subject fluctuations will then have little influence on the relative position of individuals in a group upon retesting, and the result is a high test-retest correlation. Equation (11.2) also emphasizes that reliability is population dependent (Lord & Novick, 1968; Thompson, 2003).

Test-score reliability is in most cases assessed using coefficient (Cronbach’s) α . However, coefficient α is just one of a series of lower bounds to the reliability, including the Guttman’s λ s (Guttman, 1945) and the greatest lower bound (*glb*; e.g., Woodhouse & Jackson, 1977). The latter is the highest possible lower-bound to the reliability that can be estimated from empirical data. Closer inspection of these reliability coefficients (e.g., Sijtsma & Van der Ark, 2020, chap. 2; Oosterwijk, 2016) shows that they all have a structure similar to Eq. (11.2). In particular, let λ_l ($l = 1, \dots, 6$) be one of the Guttman’s reliability indices. Let Σ_{X_j} be the item-level variance-covariance matrix for a test that consists of J items. The lower-bound reliability indices can be expressed as

$$\lambda_l = 1 - \frac{\text{tr}(\Sigma_{X_j}) - C_{\lambda_l}(\Sigma_{X_j})}{\sigma_X^2}, \tag{11.3}$$

where $\text{tr}(\Sigma_{X_j})$ is the trace (sum of diagonal elements) of Σ_{X_j} and $C_{\lambda_l}(\Sigma_{X_j})$ is a summary of the item variance-covariance matrix.

Three aspects are particularly important here. First, different choices for C_{λ_l} define the different lower bound reliability indices. Examples follow below. Second, the numerator in Eq. (11.3) is an estimate of the *squared* SEM, that is,

$$\hat{\sigma}_E^2 = \text{tr}(\Sigma_{X_j}) - C_{\lambda_l}(\Sigma_{X_j}), \tag{11.4}$$

which is directly computed from the Σ_{X_j} . As an aside we may note that Eq. (11.4) shows how lower-bound methods to the reliability essentially work; they use the covariances to make inferences about the part that is explainable from inter-individual differences in true scores and subtract that part from the sum of the observed item variances. The underlying idea is that in CTT, the inter-item covariances in a random sample only reflect true-score differences because errors are uncorrelated by construction and each item constitutes an independent observation. The practical importance of Eq. (11.4) is that we do not need to compute the reliability first and then reconstruct the SEM, but we can derive direct expressions for the SEM as a function of the item-level variance-covariance matrix.

More importantly, Expression (11.4) does *not* include $\hat{\sigma}_X^2$. Hence, Eq. (11.4) is also applicable in groups having the same sum score x ; hence, Eq. (11.4) circumvents the problem involved when using Eq. 11.1 in conditional sum-score groups (i.e., $\hat{\sigma}_X^2 = 0$ and reliability is undefined; see Sijtsma & Van der Ark, 2020, p. 80).

To illustrate the matters at hand, consider, for example, Guttman’s λ_3 , which is equivalent with coefficient (Cronbach’s) α . Coefficient λ_3 is obtained by defining

$$C_{\lambda_3}(\mathbf{\Sigma}_{X_j}) = \frac{1}{J-1} \left[\mathbf{1}^T \mathbf{\Sigma}_{X_j} \mathbf{1} - \text{tr}(\mathbf{\Sigma}_{X_j}) \right], \tag{11.5}$$

where $\mathbf{1}$ is a $J \times 1$ (column) vector of ones, such that $\mathbf{1}^T \mathbf{\Sigma}_X \mathbf{1}$ equals the sum of all elements in $\mathbf{\Sigma}_X$. Notice that $\mathbf{1}^T \mathbf{\Sigma}_{X_j} \mathbf{1}$ also equals the variance of the sum score X (i.e., S_X^2). The corresponding squared SEM equals

$$\sigma_E^2(\lambda_3) = \text{tr}(\mathbf{\Sigma}_{X_j}) - \frac{1}{J-1} \left[\mathbf{1}^T \mathbf{\Sigma}_{X_j} \mathbf{1} - \text{tr}(\mathbf{\Sigma}_{X_j}) \right], \tag{11.6}$$

which when substituted in Eq. (11.2) together with σ_X^2 gives λ_3 . Hence, Eq. (11.6) shows the estimator of the average measurement precision that defines λ_3 . Expression (11.6) also can be used to compute the CSEMs.

Before going further into CSEMs based on other λ s, let us first have a brief look at the relationship with Feldt et al.’s ANOVA method. It is well-known that the Type III intraclass correlation for the mean score, commonly denoted as ICC(3, J) (Shrout & Fleiss, 1979), from a one-way repeated measures ANOVA equals coefficient α and thus also equals λ_3 (Hoyt, 1941; Maxwell & Delaney, 2004, p. 566; Shrout & Fleiss, 1979). The ICC is given by

$$ICC(3, J) = 1 - \frac{MS_{N \times J}}{MS_s} = \lambda_3 = \alpha, \tag{11.7}$$

where $MS_{N \times J}$ is the mean squares for the interaction term and MS_s for the between-subject differences (Maxwell & Delaney, 2004, chap. 11). It can be shown that

$$J \times MS_{N \times J} = \sigma_E^2(\lambda_3) \tag{11.8}$$

(see the Appendix for the proof; see also Woodruff, 1990, p. 194, and Jarjoura, 1986 who relates the ICC to the KR-20 index of reliability). Hence, there is a direct link between Feldt et al.’s ANOVA approach and the CSEMS/SEMS that follows from using coefficient α (or the equivalent λ_3) in Eq. (11.1). Through this link, we can generalize the ANOVA-based CSEMs by deriving alternative expressions for $C_{\lambda_l}(\mathbf{\Sigma}_{X_j})$ based on other lower-bound reliability indices from the Guttman λ series and use it in Eq. (11.4).

11.2.2 CSEMs Derived from Internal Consistency Reliability Estimates

CSEMs based on internal consistency estimates – including the ANOVA approach by Feldt et al. – evaluate Eq. (11.4) in subgroups defined by the sum score X . However, if the number of persons in a score group is small, the estimates become unstable, yielding erratic fluctuations in the estimated CSEMs across the score scale. Therefore, to ensure that the score groups have enough observations, adjacent sum-score groups may be merged until the desired minimum number (minsize) is reached. The *minsize* acts as a smoothing (binning) parameter, an idea that is used intensively in nonparametric item response theory (Sijtsma & Molenaar, 2002). However, its choice is critical to provide an adequate balance between sampling variance and bias. That is, the minsize should not be set too small to avoid unstable estimates, but not too large either because then one cannot pick up the local trend in the CSEMs. More insights into the optimal choices of minsize for grouping the sample on X were investigated using simulations, which are discussed below.

Let us assume for the moment that the sample is large enough so that each sum-score group has enough observations. Let $\sigma_{E|x}^2(\lambda_3)$ be the CSEM for score group x based on the variance-covariance summary C underlying reliability index λ_3 . Because within this score group we have $\mathbf{1}^T \Sigma_{X_j} \mathbf{1} = 0$, Eq. (11.6) further simplifies to

$$\sigma_{E|x}^2(\lambda_3) = \text{tr}(\Sigma_{X_j|x}) + \frac{1}{J-1} \bullet \text{tr}(\Sigma_{X_j|x}) = \frac{J}{J-1} \bullet \text{tr}(\Sigma_{X_j|x}), \tag{11.9}$$

which for dichotomous items further simplifies to

$$\sigma_{E|x}^2(\lambda_3) = \frac{J}{J-1} \sum \pi_{j|x} (1 - \pi_{j|x}) \tag{11.10}$$

where $\pi_{j|x}$ is the proportion of correct answers for item j in group $X = x$. Taking into account the degrees of freedom for the variances (i.e., the diagonal elements $\Sigma_{X_j|x}$), the asymptotically unbiased estimate $\hat{\sigma}_{E|x}^2(\lambda_3)$ can be written as a function of the observed item statistics as

$$\hat{\sigma}_{E|x}^2(\lambda_3) = \frac{J}{J-1} \sum_{j=1}^J \hat{\sigma}_{X_j|x}^2,$$

which for dichotomous further simplifies to

$$\hat{\sigma}_{E|x}^2(\lambda_3) = \frac{J}{(J-1)} \bullet \frac{n_x}{(n_x-1)} \sum \hat{\pi}_{j|x} (1 - \hat{\pi}_{j|x}). \tag{11.11}$$

The n_x in Eq. (11.11) is the number of respondents in the sum-score group and $\hat{\pi}_{j|x}$ the observed proportion correct. A similar expression was also derived in Feldt et al. (1985; see Eq. 11.16) within the context of repeated measures ANOVA. The above expression applies to score groups of persons all having the same x -score, thus $\hat{\sigma}_{X|g}^2 = 0$. When the score group (denoted g) covers a range of X scores because some adjacent score groups had to be merged to reach the desired minsize, we have

$$\hat{\sigma}_{E|g}^2(\lambda_3) = \frac{J}{J-1} \sum_{j=1}^J \hat{\sigma}_{X_j}^2 - \frac{\hat{\sigma}_{X|g}^2}{J-1}. \quad (11.12)$$

Hence, when groups are merged, we have this additional term that also takes the inter-item covariances into account. Equation (11.12) represents the general case of the ANOVA method for dichotomous items in case adjacent groups have to be merged.

Next, we consider λ_1 and λ_2 . Although λ_1 is lowest in the ranking of the lower-bound indices, and therefore seemingly the least preferable choice of all Guttman coefficients, it may be interesting to study the properties of the CSEM that is implied by $\hat{\sigma}_{E|x}^2(\lambda_1)$. To get λ_1 using Eq. (11.3), we need to set $C_{\lambda_l}(\Sigma_{X_j}) = 0$. Hence, λ_1 takes the sum of the item-variance as the squared SEM; that is $\sum_{j=1}^J \hat{\sigma}_{X_j}^2$. Using $C_{\lambda_l}(\Sigma_{X_j}) = 0$ in score groups, and assuming dichotomous items, we have

$$\hat{\sigma}_{E|g}^2(\lambda_1) = \frac{n_g}{(n_g - 1)} \sum_{j=1}^J \hat{\pi}_{j|g} (1 - \hat{\pi}_{j|g}). \quad (11.13)$$

Hence, conditional on $X = x$, we see that $\hat{\sigma}_{E|x}^2(\lambda_3)$ is larger than $\hat{\sigma}_{E|x}^2(\lambda_1)$ by a factor $(J-1)/J$. As J grows, differences between $\hat{\sigma}_{E|g}^2(\lambda_1)$ and $\hat{\sigma}_{E|g}^2(\lambda_3)$ become negligible.

In this chapter, we also consider $C_{\lambda_2}(\Sigma_{X_j})$ that defines Guttman's λ_2 . As pointed out by Sijtsma (2009), λ_2 is preferred over λ_3 , although numerical and simulation studies suggest that differences tend to be limited from a practical point of view (Oosterwijk et al., 2016). The CSEM associated with λ_2 equals

$$\hat{\sigma}_{E|x}^2(\lambda_2) = \text{tr}(\Sigma_{X_j|x}) - \sqrt{\frac{J}{J-1} \left[\mathbf{1}' (\Sigma_{X_j|x} \odot \Sigma_{X_j|x}) \mathbf{1} - \text{tr}(\Sigma_{X_j|x} \odot \Sigma_{X_j|x}) \right]} \quad (11.14)$$

(Sijtsma & Van der Ark, 2020). Symbol \odot denotes that the products in the variance-covariance matrices are taken element-by-element (i.e., Hadamard product). Interestingly, because the square root in (11.14) always provides positive number, it follows that $\hat{\sigma}_{E|x}^2(\lambda_2) \leq \hat{\sigma}_{E|x}^2(\lambda_1) \leq \hat{\sigma}_{E|x}^2(\lambda_3)$. That is, the order of the implied CSEMs is not the same as the order for the lower-bound reliabilities. Finally, we may note that the other lambdas λ_4 through λ_6 make use of optimizations,

making them sensitive to chance capitalization (Oosterwijk, 2016) and less suited for the purposes we have in mind. Therefore they need not be considered here.

The derivations for obtaining expressions for $\hat{\sigma}_{E|x}^2(\lambda_1)$, $\hat{\sigma}_{E|x}^2(\lambda_2)$, and $\hat{\sigma}_{E|x}^2(\lambda_3)$ as shown above are straightforward, but there is an important issue that needs further consideration. When conditioning on the *observed* X , negative covariances arise between the item-level errors (Woodruff, 1990). As a result, the covariances in $\Sigma_{X_j|x}$ are no longer governed by true-score variance alone. However, inspection of (11.11) and (11.13) shows that $\hat{\sigma}_{E|x}^2(\lambda_1)$ and $\hat{\sigma}_{E|x}^2(\lambda_3)$ do not involve the covariances but only conditional item-score variances. For $\hat{\sigma}_{E|g}^2(\lambda_2)$ and $\hat{\sigma}_{E|g}^2(\lambda_3)$ based on merged groups, negative covariances come into play. It is unclear how these negative covariances play out for $\hat{\sigma}_{E|g}^2(\lambda_2)$, but for $\hat{\sigma}_{E|g}^2(\lambda_3)$, the overestimation becomes even larger because the group-specific variance $\hat{\sigma}_{X|g}^2$ will be underestimated. Regarding the item variances, Woodruff (1990) proved that the conditional error variance given X will on average be smaller than the error variance conditional on true scores; that is, $E(\sigma_{E|x}^2) < \sigma_{E|T}^2$. This effect depends on the number of items and will be most prominent for scores at the boundaries of the scale. Furthermore, given the aforementioned inequality, even though $\hat{\sigma}_{E|x}^2(\lambda_3)$ is overestimating $\sigma_{E|x}^2$, it might still be considered as a pragmatic but practical estimator of $\sigma_{E|T}^2$ (Woodruff, 1990). The key question is, of course, how accurate these estimates are from a practical perspective.

11.3 Using CSEMs in Practice: A Simulation Study

Data sets were generated using IRT modeling (e.g., Hambleton & Swaminathan, 1985). Let θ be the unidimensional latent variable, and let $P_{jx}(\theta) \equiv P(X_j = x|\theta)$ be the probability of observing response $x \in \{0, \dots, M\}$. The item responses are assumed to be independent conditional on θ (i.e., local independence). Dichotomous item responses (i.e., $M = 1$) were simulated using Birnbaum's (1968) two-parameter logistic model (2-PLM) and polytomous item-response data using the graded response model (GRM; Samejima, 1969). For details on these models and other IRT models, the reader is referred to Van der Linden (2016).

CTT and IRT are closely connected to each other (e.g., Holland & Hoskens, 2003; Lord, 1980). First, CTT's true score τ is linked to θ via $\tau = \varepsilon(X|\theta)$. As pointed out by Holland and Hoskens (2003), τ and θ are equivalent expressions of a latent attribute on different scales (see also Lord, 1980, p. 46). Second, using the $P_{jx}(\theta)$ s, we can construct for every arbitrary value of θ or τ the associated conditional frequency distribution for X . For dichotomous items, this conditional distribution is the compound binomial (Lord, 1980), and for polytomous, we have the compound multinomial distribution (Thissen et al., 1995). Recursive formulae can be found in Lord and Wingersky (1984) and in Kolen and Brennan (1995, pp. 182–182, 219). The variance of the conditional distribution is the

squared conditional CSEM. For dichotomous scores, the variance of the conditional frequency distribution of observed score X conditional on τ specializes to

$$\sigma_{X|\theta}^2 = \sum_{j=1}^J P_j(\theta) [1 - P_j(\theta)], \tag{11.15}$$

(Lord, 1980, p. 45). The IRT-based CSEMs constitute our focal quantity of interest (i.e., the estimand), and we evaluate the accuracy of $\sigma_{E|x}^2(\lambda_1)$, $\sigma_{E|x}^2(\lambda_2)$, and $\sigma_{E|x}^2(\lambda_3)$ as estimators of the estimand under varying conditions of item type (dichotomous versus polytomous), test length, and sample sizes.

To illustrate, Fig. 11.1 shows the population values for the true CSEM and the three proposed estimators of CSEMs, for a hypothetical test of 30 dichotomously scored items under the 2-PLM. The vertical bars show the boundaries for the sum score within which 95% of the population falls. The horizontal dashed line is the SEM. The CSEMs were obtained as follows. For convenience, we assume $\theta \sim N(0, 1)$. First, we find the θ s that satisfy the equality $\sum_j P_{j1}(\theta) = x$ ($x = 1, \dots, J - 1$). Notice that there is no solution for $x = 0$ and $x = J$; therefore we arbitrarily chose $\theta = -3$ and $\theta = 3$ for these x -values, respectively. Then, given θ we computed σ_X^2 using the compound binomial (Eq. 11.15). The result is the sum-score variance given the true score, and its square root is the population value of the CSEM, which will be denoted $\sigma_{E|x}$. The population values for our estimators $\sigma_{E|x}^2(\lambda_j)$ were obtained in a very large sample (i.e., $n = 100,000$). In particular, the

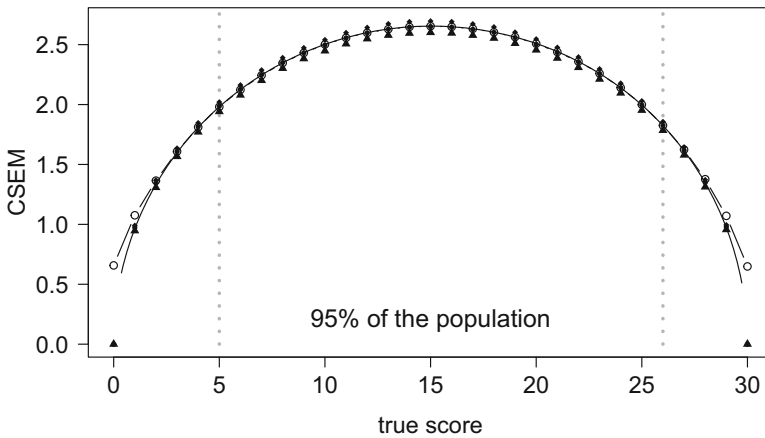


Fig. 11.1 True CEMS (blank dots) and population values for the estimators $\sigma_{E|x}^2(\lambda_1)$ (black dots), $\sigma_{E|x}^2(\lambda_2)$ (triangles) and $\sigma_{E|x}^2(\lambda_3)$ (diamonds), for a 30-item test (dichotomous items). Vertical bars indicate the limits within which 95% of the true scores in the population fall. Population values for the estimators were obtained in a simulated sample of 100,000 observations under the postulated model

black dots represent $\sigma_{E|x}^2(\lambda_1)$, the triangles $\sigma_{E|x}^2(\lambda_2)$, and the diamonds $\sigma_{E|x}^2(\lambda_3)$. Fig. 11.1 shows that $\sigma_{E|x}^2(\lambda_1)$ is least biased, $\sigma_{E|x}^2(\lambda_2)$ has a small negative bias, and $\sigma_{E|x}^2(\lambda_3)$ has a small positive bias, but in general the differences were small, and the estimators are fairly consistent across ranges where most of the observations fall. Considerable bias was only found for $x = 0$ and $x = J$, which was to be expected because the item variances within subgroups defined by $x = 0$ or $x = J$ are zero by definition, and thus $\sigma_{E|0}^2(\lambda_l) = \sigma_{E|J}^2(\lambda_l) = 0$. As a practical solution, we may use the CSEMs from neighboring sum-score levels to come up with a better estimate. However, perfect scores are rare if the overall difficulty of the test fits well with the target population of testees. Moreover, the extreme scores are generally not the cutoff points for important individual decisions. Hence, the bias at the endpoints most likely has little practical implications.

11.3.1 Study Design

Independent factors in the design were (a) test length, using $J = 10, 20,$ or 30 items for dichotomous items and $J = 5, 10, 15$ for polytomous items each with four ordered categories; (b) size of the norming sample ($N = 100, 500,$ or 1000); (c) group-level reliability (moderate or strong); and (d) minsize for sum score groups. Regarding the choices of J , we may add that because polytomous items are informative across a wider range of the trait scale, polytomous item tests usually have fewer items than dichotomous items tests. The factors were fully crossed for the dichotomous items (i.e., 36 conditions in total) and for the polytomous items (i.e., 36 conditions) separately. Data sets were generated using parameters that were obtained as follows. For both dichotomous and polytomous items, the a -parameters were either drawn from the uniform distribution on the interval $[0.5, 1.5]$ (“moderate reliability”) or $[1, 3]$ (“high reliability”). Location parameters (b) for the dichotomous items were randomly drawn from the uniform interval $[-1, 1]$. For polytomous items, the three threshold parameters for each item were randomly sampled from $[-1.5, -0.75], [-0.50, 0.50],$ and $[0.75, 1.5],$ respectively. Then for this set of parameters, we generated 500 data sets under either the 2-PLM or the GRM. Then for each data set, we computed the CSEMs $\hat{\sigma}_{E|x}(\lambda_l), l = 1, \dots, 3,$ for two different minsizes for the score groups g . These minsizes were tailored to the norming sample size (N). In particular, we used as *minsizes* 5 or 10 for $N = 100,$ 10 and 15 for $N = 500,$ and 15 and 30 for $N = 1000.$

Accuracy was operationalized using the following two outcome measures. Let $X_{\max} = J \times M$ be the maximum possible sum score. First, we computed the (weighted) bias as

$$\text{bias} [\hat{\sigma}_{E|x}(\lambda_l)] = \sum_{x=0}^{X_{\max}} \left[\overline{\hat{\sigma}_{E|x}(\lambda_l)} - \sigma_{E|x} \right] \bullet g(x), \tag{11.16}$$

and weighted squared bias as

$$\text{sq - bias} [\hat{\sigma}_{E|x}(\lambda_l)] = \sum_{x=0}^{X_{\max}} \left[\overline{\hat{\sigma}_{E|x}(\lambda_l)} - \sigma_{E|x} \right]^2 \bullet g(x). \quad (11.17)$$

The term $\overline{\hat{\sigma}_{E|x}(\lambda_l)}$ in the formulae above is the mean CSEM across the 500 data sets, and $g(x)$ is the frequency distribution of total score x in the population. Distribution $g(x)$ was obtained by taking the weighted average of the conditional sum-score distributions at 1000 Gaussian quadrature points on the interval -5 to 5 . The values of $\sigma_{E|x}$ ($x = \{0, \dots, X_{\max}\}$) were also obtained from the conditional sum-score distributions. In particular, we grouped the true scores (i.e., $\tau = E(X|\theta)$) into $(X_{\max} + 1)$ bins; that is, we have one bin for each level of x , such that within the bin associated with x , we have $|E(\theta) - x| < 0.5$. The population-level conditional sum-score variance $\sigma_{E|x}^2$ is the variance of the weighted mean of the conditional sum-score distributions within the bin.

Second, we defined the (weighted) precision as

$$\text{precision} [\hat{\sigma}_{E|x}(\lambda_l)] = \sqrt{\sum_{x=0}^{X_{\max}} S^2 [\hat{\sigma}_{E|x}(\lambda_l)] \bullet g(x)}, \quad (11.18)$$

where $S^2 [\hat{\sigma}_{E|x}(\lambda_l)]$ is the variance of the estimates across the 500 data sets. The complete design was replicated 100 times, yielding information about bias and precision for 100 different random tests.

11.3.2 Results

Table 11.1 gives the results for bias for dichotomous items. The reported values are the means across 100 replications of the design; that is, all values are based on 50,000 simulated data sets in total. Statistic $\hat{\sigma}_{E|x}(\lambda_2)$ showed considerable bias. Statistic $\hat{\sigma}_{E|x}(\lambda_l)$ performed best for moderate discrimination, whereas $\hat{\sigma}_{E|x}(\lambda_3)$ for high discrimination, but differences were small. Statistic $\hat{\sigma}_{E|x}(\lambda_1)$ has the tendency to underestimate the CSEMs, whereas $\hat{\sigma}_{E|x}(\lambda_3)$ slightly overestimated the CSEM. This trend was found for all levels of N . Table 11.2 gives the bias for polytomous items. Trends are similar as those for dichotomous items. Both $\hat{\sigma}_{E|x}(\lambda_1)$ and $\hat{\sigma}_{E|x}(\lambda_3)$ seem acceptable candidate estimators with respect to bias.

Tables 11.3 and 11.4 show the results for precision for dichotomous and polytomous items, respectively. The three methods have comparable precision. For polytomous items we see a substantial differences in precision between different levels of reliability. In general, the methods seem to have acceptable results when the norming samples have 500 or more observations and perform well if the size of the norming sample is 1000 or higher.

Table 11.1 Results for bias between true CSEMs and estimated CSEMS, for dichotomous items. Results averaged across sample sizes for the norming sample

Minsize	Item discrimination/reliability							
	Moderate				High			
	$\rho_{XX'}$	$\hat{\sigma}_{E x}(\lambda_1)$	$\hat{\sigma}_{E x}(\lambda_2)$	$\hat{\sigma}_{E x}(\lambda_3)$	$\rho_{XX'}$	$\hat{\sigma}_{E x}(\lambda_1)$	$\hat{\sigma}_{E x}(\lambda_2)$	$\hat{\sigma}_{E x}(\lambda_3)$
	<i>J</i> = 10							
5	0.65	-0.028	-0.173	0.041	0.84	-0.087	-0.220	-0.032
10	0.65	-0.016	-0.151	0.051	0.84	-0.069	-0.188	-0.016
	<i>J</i> = 20							
10	0.79	-0.004	-0.247	0.045	0.91	-0.035	-0.247	0.004
15	0.79	0.002	-0.199	0.048	0.91	-0.016	-0.192	0.020
	<i>J</i> = 30							
15	0.85	0.001	-0.337	0.040	0.94	-0.016	-0.314	0.016
30	0.85	0.007	-0.260	0.043	0.94	0.001	-0.222	0.029

Table 11.2 Results for bias between true CSEMs and estimated CSEMS, for polytomous items. Results averaged across sample sizes

Minsize	Item discrimination/reliability							
	Moderate				High			
	$\rho_{XX'}$	$\hat{\sigma}_{E x}(\lambda_3)$	$\hat{\sigma}_{E x}(\lambda_2)$	$\hat{\sigma}_{E x}(\lambda_3)$	$\rho_{XX'}$	$\hat{\sigma}_{E x}(\lambda_1)$	$\hat{\sigma}_{E x}(\lambda_2)$	$\hat{\sigma}_{E x}(\lambda_3)$
	<i>J</i> = 5							
5	0.55	-0.152	-0.526	0.086	0.81	-0.159	-0.443	0.008
10	0.55	-0.130	-0.469	0.100	0.81	-0.132	-0.389	0.026
	<i>J</i> = 10							
10	0.72	-0.077	-0.549	0.087	0.89	-0.090	-0.447	0.025
15	0.71	-0.061	-0.451	0.097	0.89	-0.068	-0.357	0.036
	<i>J</i> = 15							
15	0.79	-0.056	-0.680	0.077	0.93	-0.066	-0.536	0.026
30	0.79	-0.042	-0.511	0.082	0.93	-0.043	-0.388	0.036

11.4 Using CSEMs in Practice: Individual Change Assessment

Social and emotional well-being of students has gained widespread interest next to educational achievements in the cognitive domain. Good mental health is important for educational progress (e.g., Payton et al., 2008; Wang et al., 1997). Schools may regularly assess their pupil’s well-being using standardized questionnaires, which can be filled out by the pupil’s parents/caretakers, or by the pupils themselves. If there are indications that the students are not functioning or feeling well, targeted interventions can be offered that provide social-emotional support and empowerment. The effects of interventions at the individual level can be evaluated using the difference score, which is simply the score after treatment (X_{post}) minus the pre-treatment score (X_{pre}). However, because the scores have measurement errors,

Table 11.3 Precision of estimated CSEMs, for dichotomous items. Results averaged across sample sizes for the norming sample

Minsize	Item discrimination/reliability							
	Moderate				High			
	$\rho_{XX'}$	$\hat{\sigma}_{E x}(\lambda_3)$	$\hat{\sigma}_{E x}(\lambda_2)$	$\hat{\sigma}_{E x}(\lambda_3)$	$\rho_{XX'}$	$\hat{\sigma}_{E x}(\lambda_1)$	$\hat{\sigma}_{E x}(\lambda_2)$	$\hat{\sigma}_{E x}(\lambda_3)$
<i>J</i> = 10, <i>N</i> = 100								
5	0.65	0.089	0.088	0.095	0.84	0.127	0.122	0.133
10	0.65	0.084	0.077	0.091	0.84	0.143	0.128	0.149
<i>J</i> = 20, <i>N</i> = 100								
10	0.79	0.095	0.111	0.099	0.91	0.143	0.132	0.146
15	0.79	0.089	0.086	0.094	0.91	0.134	0.114	0.138
<i>J</i> = 30, <i>N</i> = 100								
15	0.85	0.102	0.126	0.105	0.94	0.143	0.131	0.145
30	0.85	0.098	0.093	0.103	0.94	0.147	0.126	0.150
<i>J</i> = 10, <i>N</i> = 500								
5	0.65	0.069	0.065	0.072	0.84	0.029	0.032	0.031
10	0.65	0.048	0.045	0.050	0.84	0.055	0.053	0.057
<i>J</i> = 20, <i>N</i> = 500								
10	0.79	0.042	0.042	0.043	0.91	0.065	0.065	0.067
15	0.79	0.051	0.050	0.054	0.91	0.090	0.085	0.092
<i>J</i> = 30, <i>N</i> = 500								
15	0.85	0.050	0.056	0.051	0.94	0.083	0.085	0.084
30	0.85	0.058	0.061	0.060	0.94	0.086	0.079	0.087

we first have to ascertain whether the difference is *reliable* before drawing strong conclusions. One speaks of reliable change if the difference score is significantly larger than the differences expected by chance alone (Jacobson & Truax, 1991).

A popular method for testing the significance of change is the *reliable change index* (RCI; Jacobson & Truax, 1991). Let $d = X_{\text{post}} - X_{\text{pre}}$ denote the difference score. The RCI is defined as

$$RCI = \frac{d}{\sqrt{2 \cdot \hat{\sigma}_E^2}}, \tag{11.19}$$

where $\hat{\sigma}_E^2$ is the error variance. Absolute values of the RCI of 1.96 or higher point at reliable change. In fact what we have is a two-tailed test at a 5% significance level (Sijtsma & Emons, 2011). There is no consensus in the literature as to which error term should be used in the RCI (see, for example, Maassen, 2004), but our experience is that most researchers use the (unconditional) SEM.

The traditional RCI approach ignores the differences in precision at different attribute levels, which causes bias in *all* the RCI tests. This bias works in both ways. In the middle ranges of the score scale, where the SEMs tend to underestimate the precision (Fig. 11.1), the traditional RCI becomes *liberal*. This means that the chance that ineffective interventions are erroneously conceived as very successful

Table 11.4 Precision of estimated CSEMs, for polytomous items. Results averaged across sample sizes for the norming sample

Minsize	Item discrimination/reliability							
	Moderate				High			
	$\rho_{XX'}$	$\hat{\sigma}_{E x}(\lambda_3)$	$\hat{\sigma}_{E x}(\lambda_2)$	$\hat{\sigma}_{E x}(\lambda_3)$	$\rho_{XX'}$	$\hat{\sigma}_{E x}(\lambda_1)$	$\hat{\sigma}_{E x}(\lambda_2)$	$\hat{\sigma}_{E x}(\lambda_3)$
	$J = 10, N = 100$							
Small	0.65	0.230	0.219	0.260	0.84	0.216	0.189	0.242
Medium	0.65	0.201	0.177	0.232	0.84	0.174	0.151	0.199
	$J = 20, N = 100$							
Small	0.79	0.246	0.242	0.263	0.91	0.240	0.215	0.256
Medium	0.79	0.196	0.180	0.215	0.91	0.193	0.172	0.210
	$J = 30, N = 100$							
Small	0.85	0.255	0.251	0.268	0.94	0.257	0.227	0.269
Medium	0.85	0.203	0.185	0.219	0.94	0.208	0.185	0.222
	$J = 10, N = 500$							
Small	0.65	0.101	0.093	0.112	0.84	0.097	0.087	0.108
Medium	0.65	0.110	0.097	0.125	0.84	0.108	0.095	0.120
	$J = 20, N = 500$							
Small	0.79	0.129	0.126	0.137	0.91	0.134	0.126	0.142
Medium	0.79	0.124	0.119	0.133	0.91	0.125	0.116	0.133
	$J = 30, N = 500$							
Small	0.85	0.142	0.147	0.148	0.94	0.151	0.145	0.156
Medium	0.85	0.125	0.123	0.133	0.94	0.133	0.123	0.140

is higher than the chosen nominal level α . In the extremes of the X -score scale, where the SEMs tend to overestimate precision (Fig. 11.1), the RCI test becomes *conservative*. The result is a reduction in the power, thus a higher risk that potentially effective interventions are overlooked. It is for this reason that IRT methods were considered superior to the traditional RCI approach because they make use of the local accuracy (Reise & Haviland, 2005). However, such an approach is also possible with CSEMs discussed in this chapter. In fact, if there is one application where the use of CSEMs may have added value, it would be change assessment using the RCI.

Following Jabrayilov et al. (2016), we can extend the RCI using the CSEMs, denoted by cRCI, as follows:

$$cRCI = \frac{x_{post} - x_{pre}}{\sqrt{\hat{\sigma}_{E|x_{pre}}^2(\lambda_l) + \hat{\sigma}_{E|x_{post}}^2(\lambda_l)}}. \tag{11.20}$$

The generalization of the RCI as presented in Eq. (11.20) is straightforward. Based on the simulations above, λ_1 or λ_3 would be feasible choices.

The question now is how to test the cRCI for significance? A simple approach would be to assume that the cRCI is also a standard normal deviate (Z -score). However, one must take into account that the CSEMs in Eq. (11.20) used in cRCI are

based on the fallible *observed* pretest and posttest scores. This means that not only the numerator in Eq. (11.20) but also the denominator would vary if we would repeatedly pre-posttest the individual under identical conditions. As a consequence, the standard error of the cRCI (denominator in Eq. 11.19) is likely to be greater than one (i.e., the error for a normal deviate). If you were to apply the normal distribution anyway, the test for individual change becomes somewhat liberal; that is, the chance of a Type I error exceeds the nominal level α . Ideally, we would like to have the exact sampling distribution, but for the cRCI, its derivation is not self-evident and beyond the scope of this chapter. Therefore, we take the *practical* approach at this stage, and we use the Z-distribution as a practical approximation for testing significance of the cRCI (see also Jabrayilov et al., 2016). The accuracy of this practical approach, and how it improves the traditional RCI, is addressed in a simulation study below.

11.4.1 Comparing RCI and cRCI: A Simulation Study

Because it is common to assess non-cognitive attributes using Likert items, we only simulated data for polytomous items. In particular, we considered tests of 5, 10, or 15 items. Data were generated as follows. For each level of J , we obtained the item parameters as follows. First, we drew the mean threshold \bar{b} from the uniform distribution on the interval $[-0.25, 0.75]$. Then, second, the individual thresholds were set at $\bar{b}-0.75$, $\bar{b} - 0.25$, $\bar{b} + 0.25$, and $\bar{b} + 0.75$, respectively. The result are tests that are most informative for above-average individuals. This is a typical pattern for non-cognitive (clinical) assessments (e.g., Jabrayilov et al., 2016). Based on the item parameters, we simulated a norming sample of 1000 respondents using randomly drawn θ -values from the standard normal distribution. This sample was used to compute the CSEMs. We chose a relatively large norming sample because for tests that are used on a large scale for important decisions, one may expect (or even require) that the psychometric properties are based on sufficiently large representative samples.

Next we simulated 1000 pairs of pretest and posttest scores at three dedicated ability levels, that is $\theta = 0.5, 1.0$, and 1.5 (i.e., 3000 score pairs in total). For each pair of pretest and posttest scores, we computed the RCI and cRCIs using either $\hat{\sigma}_{E|x_{\text{pre}}}^2(\lambda_1)$ or $\hat{\sigma}_{E|x_{\text{pre}}}^2(\lambda_3)$. These CSEMs were chosen because they performed adequately in the simulations above. The proportion of absolute RCIs or cRCIs that exceed 1.645 constitutes the empirical Type I error rate at a 10% significance level. Ideally, the empirical Type I error rates are close to their nominal level of 0.10.

Table 11.5 shows that for $\theta = 0$ and 0.5, the traditional RCI yields empirical Type I errors rates that are substantially larger than the nominal level of 0.10. This result is to be expected given that in the middle of the ability scale, the cRCI exceeds RCI (see Fig. 11.1). This trend was strongest for high reliability. For $\theta = 1$ the RCI method performed satisfactory, for both moderate and high reliability. Statistic $\hat{\sigma}_{E|x_{\text{pre}}}^2(\lambda_1)$ shows a similar pattern, but the deviations from the nominal level are

Table 11.5 Results for change assessment (results across 100 replications)

J	Test-score reliability							
	Moderate				High			
	$\rho_{XX'}$	RCI	cRCI λ_1	cRCI λ_3	$\rho_{XX'}$	RCI	cRCI λ_1	cRCI λ_3
	$\theta = 0$							
5	0.740	0.163	0.148	0.120	0.830	0.187	0.142	0.106
10	0.838	0.131	0.122	0.116	0.895	0.149	0.124	0.096
15	0.896	0.154	0.114	0.104	0.936	0.161	0.114	0.100
	$\theta = 0.5$							
5	0.740	0.164	0.150	0.122	0.830	0.181	0.145	0.108
10	0.838	0.128	0.124	0.115	0.895	0.149	0.126	0.099
15	0.896	0.154	0.115	0.104	0.936	0.160	0.115	0.102
	$\theta = 1.0$							
5	0.740	0.125	0.153	0.117	0.830	0.122	0.151	0.106
10	0.838	0.093	0.139	0.112	0.895	0.103	0.131	0.116
15	0.896	0.116	0.119	0.107	0.936	0.111	0.118	0.107

Note: J = test length

smaller. Statistic $\hat{\sigma}_{E|x_{pre}}^2(\lambda_3)$ performed adequately at all θ -levels and all levels of test length J.

11.5 Discussion

This chapter was very much inspired by the presidential address that Sijtsma gave to the Psychometric Society in 2012 (Sijtsma, 2012). In his address, Sijtsma emphasized the importance of disseminating psychometric knowledge and identified several research topics on the basis of: “Ask what psychometrics can do for psychology”. One of those topics was individual measurement and individual decision-making, highly relevant for education and clinical psychology. High-stakes assessment decisions can have far-reaching consequences for the individual. It is therefore essential that test users understand the uncertainties with which they make individual decisions. Group-level measures such as the test score reliability and the SEM are helpful to select tests that are used for research purposes, but fall short when tests are selected for use at the individual level. In this chapter, we focused on CSEMS based on lower-bound reliabilities that may take this role. These methods are straightforward to implement, mild in their assumptions, and the simulations suggest that they work well for realistic sample sizes and test lengths.

Nevertheless, there are some important issues to keep in mind. First, it is important to emphasize that the CSEMs still are an average of the precision, but now for a restricted group related to the trait level. However, because we *hypothesize* that inter-individual differences in measurement precision are smaller for persons with

nearby trait levels than across a wide range of the trait levels, we also *hypothesize* that the CSEMs provide a better *predictor* of the local precision than the SEM. Because we cannot truly retest persons, this view remains a working hypothesis. Note that this issue also applies to IRT-based CSEMs. IRT models are essentially cross-sectional models, and the IRT-based CSEMs are the averages in precision across different cross sections defined by θ . This conception of CSEMs represents the random sampling view on IRT models (Holland, 1990). When disaggregating group-level CSEMs to individuals, one is switching to a stochastic subject view, which involves additional (untestable) assumptions regarding local homogeneity (Ellis & Van den Wollenberg, 1993; Holland, 1990). The random sampling view, however, does not exclude deterministic response processes at the individual level. Such deterministic responses would imply perfect reliability at the individual level (see Lumsden, 1978). We contend that assuming some degree of stochasticity within the individual is a defensible position, which then *partly* explains variability at the group level. From that point of view, the CSEMs can be conceived as an *upper* bound for individual measurement precision. Admittedly, it is a conservative approach, but it prevents test users from taking the observed scores too literally.

Second, as shown in the second series of simulations, using the CSEMs may improve the RCI methodology. However, caution must be exercised when interpreting the RCI. A non-significant RCI does *not* imply that the person did *not* change, but only that the evidence of change is not strong enough to draw strong conclusions. Furthermore, determining whether the change is reliable is usually just the first step in a clinical analysis. An equally important question is whether the change is meaningful and clinically relevant (e.g., Jacobson & Truax, 1991), which requires normative information (e.g., Gu et al., 2021). Depending on whether or not the observed change can be regarded as reliable, the clinician can weigh the clinical significance of the change in different ways.

Third, the reported CSEMs are sample estimates and thus have sampling errors themselves. Given the importance CSEMs may have for future testees over a longer period of time, one cannot simply take them at face value. Their use is only justified if it has been demonstrated that the CEMS have been estimated with sufficient precision given their applications envisaged. As an aside, we may note that this requirement applies to all psychometric quantities obtained in samples (Oosterwijk et al., 2019). To accomplish this goal, we need the sampling distributions of the quantities at hand. Having access to the exact sampling distribution would be ideal, but often an asymptotic approximation works well too. Deriving sampling distributions is a subject for further research. The marginal modeling approach proposed by Kuijpers et al. (2013) may provide elegant solutions.

One may also use the nonparametric bootstrap to gauge standard errors in the estimated CSEMs. The procedure is straightforward. One creates K (say 500) replicates of the data set by drawing observations from the sample with replacement. For each replicated data set, one computes the CSEMs. The variance of the CSEMs across the replicated data sets can serve as an indicator of the precision. In addition, one may derive general guidelines for the sample size requirements that test constructors must adhere to. See, for example, the review system of test

quality of the Dutch committee of testing for a similar approach to other important psychometric quantities (COTAN; Evers et al., 2010). Our simulations suggest that samples of 500 or larger and minsizes of 10 persons per score group may already give acceptable results. Future research may focus on more fine-grained guidelines that test developers can use.

A.1 Appendix

A.1.1 Proof of Eq. 11.8

We start from the well-known definition of the standard error of measurement; that is,

$$\sigma_E^2(\lambda_3) = (1 - \alpha) \bullet S_X^2, \quad (11.A1)$$

where α is coefficient alpha and S_X^2 the variance of the total scores across persons. Because $\alpha \equiv ICC(3, J) = \left[1 - \frac{MS_{N \times J}}{MS_s}\right]$ (Eq. 11.7), substituting the definition of $ICC(3, J)$ for α gives

$$\sigma_E^2(\lambda_3) = \frac{MS_{N \times J}}{MS_s} \bullet S_X^2. \quad (11.A2)$$

Furthermore, we have $MS_s = \frac{J \sum_v \bar{X}_v^2 - n J \bar{X}^2}{n-1}$ (e.g., Brennan, 2001, p. 41), where \bar{X}_v^2 is the square average test score for an arbitrary person v . It can be shown – after some tedious algebra – that MS_s is equivalent with $\frac{\sum_v X_v^2 - n \bar{X}^2}{J(n-1)} = \frac{S_X^2}{J}$, showing that MS_s can be conceived as the average variance of subjects across items. Substituting $\frac{S_X^2}{J}$ for MS_s in Eq. (11.A2) gives $\frac{MS_{N \times J}}{\frac{S_X^2}{J}} \bullet S_X^2 = J MS_{N \times J}$, and that completes the proof.

References

- Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in modern psychometrics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511490026>

- Brennan, R. L. (1998). Raw-score conditional standard error of measurement in generalizability theory. *Applied Psychological Measurement*, 22(4), 307–331. <https://doi.org/10.1177/014662169802200401>
- Brennan, R. L. (2001). *Generalizability theory*. Springer. <https://doi.org/10.1007/978-1-4757-3456-0>
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN review system for evaluating test quality*. Nederlands Instituut van Psychologen. <https://www.psynip.nl/wp-content/uploads/2019/05/NIP-Brochure-Cotan-2018-correctie-1.pdf>
- Ellis, J. L., & Van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models: A characterization of the homogeneous monotone IRT model. *Psychometrika*, 58(3), 417–429. <https://doi.org/10.1007/BF02294649>
- Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, 9(4), 351–361. <https://doi.org/10.1177/014662168500900402>
- Gu, Z., Emons, W. H. M., & Sijtsma, K. (2021). Precision and sample size requirements for regression-based norming methods for change scores. *Assessment*, 28(2), 503–517. <https://doi.org/10.1177/1073191120913607>
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–282. <https://doi.org/10.1007/BF02288892>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Academic Publishers.
- Harvill, L. M. (1991). Standard error of measurement. *National Council on Educational Measurement*, 10(2), 33–41. <https://doi.org/10.1111/j.1745-3992.1991.tb00195.x>
- Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, 55(4), 577–601. <https://doi.org/10.1007/BF02294609>
- Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68(1), 123–149. <https://doi.org/10.1007/BF02296657>
- Hopster-Den Otter, D., Muilenburg, S. N., Wools, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2019). Comparing the influence of various measurement error presentations in test score reports on educational decision making. *Assessment in Education: Principles, Policy & Practice*, 26(2), 123–142. <https://doi.org/10.1080/0969594X.2018.1447908>
- Hoyt, C. (1941). Test reliability obtained by analysis of variance. *Psychometrika*, 6(3), 153–160. <https://doi.org/10.1007/BF02289270>
- Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, 40(8), 559–572. <https://doi.org/10.1177/0146621616664046>
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to define meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12–19. <https://doi.org/10.1037//0022-006x.59.1.12>
- Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement*, 10(2), 175–186. <https://doi.org/10.1177/014662168601000209>
- Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. Springer.
- Kuijpers, R. E., Van der Ark, L. A., & Croon, M. (2013). Testing hypotheses involving Cronbach's alpha using marginal models. *British Journal of Mathematical and Statistical Psychology*, 66(3), 503–220. <https://doi.org/10.1111/bmsp.12010>
- Lee, W.-C., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement*, 37(1), 1–20. <https://doi.org/10.1111/j.1745-3984.2000.tb01073.x>
- Lek, K. M., & Van De Schoot, R. (2018). A comparison of the single, conditional and person-specific standard error of measurement: What do they measure and when to use them? *Frontiers in Applied Mathematics and Statistics*, 4(1). <https://doi.org/10.3389/fams.2018.00040>

- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Laurence Erlbaum. <https://doi.org/10.4324/9780203056615>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, 8(4), 453–461. <https://doi.org/10.1177/014662168400800409>
- Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, 31(1), 19–26. <https://doi.org/10.1111/j.2044-8317.1978.tb00568.x>
- Maassen, G. H. (2004). The standard error in the Jacobson and Truax reliable change index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society*, 10(6), 888–893. <https://doi.org/10.1017/s1355617704106097>
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data. A model comparison perspective* (2nd ed.). Lawrence Erlbaum. <https://doi.org/10.4324/9781315642956>
- Mellenbergh, G. J. (1996). Measurement precision in test scores and item response models. *Psychological Methods*, 1(3), 293–299. <https://doi.org/10.1037/1082-989X.1.3.293>
- Nicewander, A. (2019). Conditional precision of measurement for test scores: Are conditional standard errors sufficient. *Educational and Psychological Measurement*, 79(1), 5–18. <https://doi.org/10.1177/0013164418758373>
- Oosterwijk, P. (2016). *Statistical properties and practical use of classical test-score reliability methods [Unpublished doctoral dissertation]*. Tilburg University.
- Oosterwijk, P., Van der Ark, L. A., & Sijtsma, K. (2016). Numerical differences between Guttman's reliability coefficients and the GLB. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wieberg (Eds.), *Quantitative psychology research*. Springer. https://doi.org/10.1007/978-3-319-38759-8_12
- Oosterwijk, P., Van der Ark, L. A., & Sijtsma, K. (2019). Using confidence intervals for assessing reliability of real tests. *Assessment*, 26(7), 1207–1216. <https://doi.org/10.1177/1073191117737375>
- Payton, J., Weissberg, R. P., Durlak, J. A., Dymnicki, A. B., Taylor, R. D., Schellinger, K. B., et al. (2008). *The positive impact of social and emotional learning for kindergarten to eighth-grade students: Findings from three scientific reviews*. Collaborative for Academic, Social, and Emotional Learning. <https://files.eric.ed.gov/fulltext/ED505370.pdf>
- Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement*, 29(3), 213–225. <https://doi.org/10.1111/j.1745-3984.1992.tb00374.x>
- Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, 31(3), 169–180. <https://doi.org/10.1177/0146621606291569>
- Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, 84(3), 228–238. https://doi.org/10.1207/s15327752jpa8403_02
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores I. *ETS Research Bulletin Series*, 1968(1), 1–169.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika*, 77(1), 4–20. <https://doi.org/10.1007/s11336-011-9242-4>
- Sijtsma, K., & Emons, W. H. M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, 70(6), 565–572. <https://doi.org/10.1016/j.jpsychores.2010.11.002>
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage.
- Sijtsma, K., & Van der Ark, L. A. (2020). Measurement models for psychological attributes. *CRC Press*. <https://doi.org/10.1201/9780429112447>

- Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*(1), 39–49. <https://doi.org/10.1177/014662169501900105>
- Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Sage.
- Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement*. Addison-Wesley.
- Van der Linden, W. J. (2016). Handbook of item response theory. *CRC Press*. <https://doi.org/10.1201/9781315374512>
- Wang, M. C., Haertel, G. D., & Walberg, H. J. (1997). Learning influences. In H. J. Walberg & G. D. Haertel (Eds.), *Psychology and educational practice* (pp. 199–211). McCutchan.
- Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search for the greatest lower bound. *Psychometrika, 42*(4), 579–591. <https://doi.org/10.1007/BF02295980>
- Woodruff, D. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement, 27*(3), 191–208. <https://doi.org/10.1111/j.1745-3984.1990.tb00743.x>
- Woodruff, D., Traynor, A., Cui, Z., & Fang, Y. (2013). *A comparison of three methods for computing scale score conditional standard errors of measurement*. ACT. <https://files.eric.ed.gov/fulltext/ED555593.pdf>

Part III
Item Response Theory

Chapter 12

Composition Algorithms for Conditional Distributions



Maarten Marsman, Timo B. Bechger, and Gunter K. J. Maris

Abstract This chapter is about two recently published algorithms that can be used to sample from conditional distributions. We show how the efficiency of the algorithms can be improved when a sample is required from many conditional distributions. Using real-data examples from educational measurement, we show how the algorithms can be used to sample from intractable full-conditional distributions of the person and item parameters in an application of the Gibbs sampler.

12.1 Introduction

Bayesian statistics often requires sampling from conditional, posterior distributions. For example, to estimate Bayesian models using Gibbs sampling (Geman & Geman 1984), we have to repeatedly sample from the full-conditional distributions of model parameters, and to produce plausible values (Mislevy 1991; Mislevy et al. 1993) for secondary analyses of educational surveys, we have to sample from pupils' conditional, posterior ability distributions. This chapter is about two algorithms that were designed for this problem: A rejection algorithm that was mentioned by Rubin (1984) and was applied in the *European Survey on Language Competences* (ESLC; Maris 2012) and the *Single-Variable Exchange* (SVE) algorithm developed by Murray et al. (2012).

Both algorithms are based on the observation that a sample from a conditional distribution can be obtained from samples drawn from the joint distribution. The practical significance of this observation lies in the fact that sampling from the joint distribution is often easier because it can be done in two ways. Specifically, the joint

M. Marsman (✉)
University of Amsterdam, Amsterdam, The Netherlands
e-mail: M.Marsman@uva.nl

T. B. Bechger · Gunter K. J. Maris
Tata Consultancy Services, Amsterdam, The Netherlands
e-mail: timo.bechger@tcs.com; Gunter.maris@tcs.com

density of X and Y can be factored in two ways:

$$f(x | y)f(y) = f(y | x)f(x),$$

and to obtain a sample from the joint distribution, we can use the *method of composition* (Tanner 1993) and sample from $f(y)$ and then from $f(x | y)$ or sample from $f(x)$ and then from $f(y | x)$. Thus, if it is difficult to sample from $f(x | y)$, we can try to sample from $f(y | x)$, or vice versa. For instance, if we encounter a posterior distribution that is highly intractable, we can sample from it by generating data. Thus, the algorithms are extremely useful when it is difficult to sample from the posterior but easy to generate data as is the case for *Item Response Theory (IRT)* models. As both algorithms use composition to sample from the joint distribution, we refer to them as *composition algorithms*. The algorithms differ in the way they select observations from the joint distribution to obtain a sample from the conditional distribution of interest.

Marsman et al. (2017) recently showed that the two composition algorithms could be made more efficient when we need not one but many samples from similar posterior distributions. This occurs, for instance, in educational surveys, where we have to sample from the posterior distribution of each of N individuals to produce plausible values. In this chapter, we use the composition algorithms to sample from conditional distributions of the following form:

$$f_r(\theta | \mathbf{x}_r) \propto f(\mathbf{x}_r | \theta)f_r(\theta) \tag{12.1}$$

where Θ is a random effect that varies across replications $r = 1, \dots, N$. We follow Marsman et al. (2017) and demonstrate how the composition algorithms can be tailored for the situation where N is very large. Over the last decade, large values of N have become increasingly more common as more and more data are being produced. This implies that there is a growing need to analyze large data sets and our algorithms are specifically designed for this purpose, mainly because their efficiency increases with N . The algorithms are not developed for situations where N is small.

The algorithms are useful in many contexts. Marsman et al. (2017) discussed their use for models in the exponential family and illustrated them using the Rasch (1960) model. The main goal of this chapter is to illustrate how the algorithms can be used in educational measurement applications where \mathbf{X} is a vector of discrete item responses,¹ Θ is a latent ability, $P(X | \theta)$ is an IRT model with fixed item parameters, and we use the composition algorithms to sample from the posterior distribution of ability for each of N persons, either for its one right or as part of a Gibbs sampler. Compared to alternative approaches, the main advantage of the composition algorithms is that they become more efficient when the number of persons increases, as explained in Sect. 12.3.

¹ The responses are allowed to be continuous in the SVE algorithm, and we use this to sample from posteriors of the form $f(\theta | \mathbf{x}) \propto f(\mathbf{x} | \theta)f(\theta)$ in the examples section.

The composition algorithms only require that we can generate data which is trivial for common IRT models. A nice feature is that we only need to know $f(\theta)$ and $P(X | \theta)$ up to a constant. This opens the door to new applications which would be difficult to handle with existing algorithms. We will illustrate this with an example involving a random-effects gamma model for response times. The normalizing constant (i.e., the gamma function) is not available in closed form and sometimes difficult to approximate.

To set the stage, we will first introduce the two composition algorithms as they stand. After having introduced the composition algorithms, we explain how they can be made more efficient and illustrate their use with simulated and real-data applications. The chapter ends with a discussion.

12.2 Sampling from a Conditional Distribution

12.2.1 The Rejection Algorithm

The rejection algorithm (see Algorithm 1) works as follows. To sample from a conditional distribution $f(\theta | \mathbf{x})$, we repeatedly sample $\{\theta^*, \mathbf{x}^*\}$ from the joint distribution of θ and \mathbf{x} until we produce a sample for which $\mathbf{x}^* = \mathbf{x}$. This generates an i.i.d. sample from the conditional distribution $f(\theta | \mathbf{x})$. The algorithm requires two things: First, it must be possible to sample from $f(\theta)$ and $P(\mathbf{x} | \theta)$; that is, we should be able to generate data under the model. Second, the random variable \mathbf{X} must be discrete with a finite range so that there is a non-zero probability to generate a value \mathbf{x}^* equal to the observed value \mathbf{x} .

Algorithm 1 A rejection algorithm for $f(\theta | \mathbf{x})$

- 1: **repeat**
 - 2: Generate $\theta^* \sim f(\theta)$
 - 3: Generate $\mathbf{x}^* \sim P(\mathbf{x} | \theta^*)$
 - 4: **until** $\mathbf{x}^* = \mathbf{x}$
 - 5: Set $\theta = \theta^*$
-

It will be clear that the number of trials needed increases with the number of values \mathbf{X} can assume so that the rejection algorithm is only useful when this number is small. In the special case when $P(\mathbf{x} | \theta)$ belongs to the exponential family, the posterior depends on the data only via the sufficient statistic $t(\mathbf{x})$ (Dawid 1979). Since \mathbf{X} is a discrete random variable, $t(\mathbf{X})$ is also a discrete random variable, and this means that we may replace $\mathbf{x}^* = \mathbf{x}$ with $t(\mathbf{x}^*) = t(\mathbf{x})$ in line 4 of Algorithm 1. This version of the rejection algorithm was developed for the ESLC, and it is the focus of the present chapter.

Note that the more realizations of \mathbf{X} lead to the same value on the sufficient statistic, the more efficient the algorithm becomes. The ESLC shows that the algorithm

is efficient enough to be used in large-scale educational surveys using the *Partial Credit Model* (PCM; Masters 1982). The same holds for other exponential family IRT models, such as the *Rasch model* (Rasch 1960), the *One-Parameter Logistic Model* (OPLM; Verhelst & Glas, 1995), and special cases of the *Generalized Partial Credit Model* (GPCM; Muraki 1992) and *Nominal Response Model* (NRM; Bock 1972) where the category parameters are integer.

12.2.2 The Single-Variable Exchange Algorithm

The rejection algorithm rejects all samples for which \mathbf{x}^* does not exactly match \mathbf{x} and thus requires the random variable \mathbf{X} to be *discrete*, preferably assuming a small number of values. To allow \mathbf{X} to be continuous, we adapt the rejection step such that we accept or reject samples with a probability other than 0 or 1. That is, we consider the generated θ^* as a sample from the *proposal distribution* $f(\theta \mid \mathbf{x}^*)$ and accept this value as a realization from the *target distribution* $f(\theta \mid \mathbf{x})$ with a probability $f(\theta^* \mid \mathbf{x}) / (M f(\theta^* \mid \mathbf{x}^*))$, where $M > 0$ is an appropriate bound on $f(\theta^* \mid \mathbf{x}) / f(\theta^* \mid \mathbf{x}^*)$ for all possible values of \mathbf{x} and \mathbf{x}^* . In general, however, it is difficult to find M , and we therefore consider a Metropolis algorithm. That is, we choose the probability to accept such that the accepted values are a sample from a Markov chain whose stationary distribution is $f(\theta \mid \mathbf{x})$. The price to pay is that we now produce a *dependent and identically distributed (d.i.d.)* sample.

To ensure that the Markov chain generated by the Metropolis algorithm has the desired stationary distribution, the following detailed balance condition must hold (Tierney 1994):

$$\begin{aligned} \pi(\theta' \rightarrow \theta^*) & \frac{P(\mathbf{x} \mid \theta') f(\theta')}{P(\mathbf{x})} \frac{P(\mathbf{x}^* \mid \theta^*) f(\theta^*)}{P(\mathbf{x}^*)} \\ & = \pi(\theta^* \rightarrow \theta') \frac{P(\mathbf{x} \mid \theta^*) f(\theta^*)}{P(\mathbf{x})} \frac{P(\mathbf{x}^* \mid \theta') f(\theta')}{P(\mathbf{x}^*)}, \end{aligned}$$

where θ' is the current parameter setting and $\pi(\theta' \rightarrow \theta^*)$ the probability to make a transition of θ' to θ^* . It is easily checked that the detailed balance condition holds when $\pi(\theta' \rightarrow \theta^*) = \min\{1, \omega(\theta' \rightarrow \theta^*)\}$, with

$$\omega(\theta' \rightarrow \theta^*) = \frac{P(\mathbf{x} \mid \theta^*) f(\theta^*) P(\mathbf{x}^* \mid \theta') f(\theta')}{P(\mathbf{x} \mid \theta') f(\theta') P(\mathbf{x}^* \mid \theta^*) f(\theta^*)} = \frac{P(\mathbf{x} \mid \theta^*) P(\mathbf{x}^* \mid \theta')}{P(\mathbf{x} \mid \theta') P(\mathbf{x}^* \mid \theta^*)}, \quad (12.2)$$

and the probability to accept θ^* depends on the relative likelihood to observe \mathbf{x}^* and \mathbf{x} given the parameter settings θ' or θ^* , respectively. Using this probability in the Metropolis algorithm, we arrive at the SVE; see Algorithm 2.

Algorithm 2 The Single-Variable Exchange algorithm

```

1: Draw  $\theta^* \sim f(\theta)$ 
2: Draw  $\mathbf{x}^* \sim P(\mathbf{x} \mid \theta^*)$ 
3: Draw  $u \sim \mathcal{U}(0, 1)$ 
4: if ( $u < \pi(\theta' \rightarrow \theta^*)$ ) then
5:    $\theta' = \theta^*$ 
6: end if

```

To use the SVE algorithm, we must be able to compute $\omega(\theta' \rightarrow \theta^*)$, and the SVE algorithm was designed to make this task as simple as possible. To see this, we write

$$P(\mathbf{x} \mid \theta) = \frac{h(\mathbf{x}; \theta)}{Z(\theta)},$$

where $Z(\theta) = \sum_{\mathbf{x}} h(\mathbf{x}; \theta)$ is a normalizing constant, or partition function, which is often difficult or even impossible to compute.² Since $\omega(\theta' \rightarrow \theta^*)$ in (12.2) is the product of likelihood ratios, it follows that

$$\omega(\theta' \rightarrow \theta^*) = \frac{\frac{h(\mathbf{x}; \theta^*)}{Z(\theta^*)} \frac{h(\mathbf{x}^*; \theta')}{Z(\theta')}}{\frac{h(\mathbf{x}; \theta')}{Z(\theta')} \frac{h(\mathbf{x}^*; \theta^*)}{Z(\theta^*)}} = \frac{h(\mathbf{x}; \theta^*)h(\mathbf{x}^*; \theta')}{h(\mathbf{x}; \theta')h(\mathbf{x}^*; \theta^*)}.$$

Thus, there is no need to compute $Z(\theta)$ (or $P(\mathbf{x})$).

As an illustration, Table 12.1 gives $\ln(\omega(\theta' \rightarrow \theta^*))$ for a selection of IRT models. Note that for many of the models in Table 12.1, $\ln(\omega(\theta' \rightarrow \theta^*))$ is of the form:

$$(\theta^* - \theta')(t(\mathbf{x}) - t(\mathbf{x}^*)).$$

That is, the acceptance probability depends on the product of the difference in parameter settings and the difference between the statistics of the generated and observed data. It also shows that, as the range of $t(\mathbf{X})$ increases, $\omega(\theta' \rightarrow \theta^*)$ tends to become lower, on average.

12.2.3 Limitations

In educational measurement, we often have to sample from the posterior ability distribution of each of N persons, where N is large. In the Programme for International Student Assessment, a large-scale educational survey, plausible values

² When both $Z(\theta)$ and $P(\mathbf{x})$ are difficult or even impossible to compute, the posterior distribution is called doubly intractable. Murray et al. (2012) specifically developed the SVE algorithm for these doubly intractable distributions.

Table 12.1 $\ln(\omega(\theta' \rightarrow \theta^*))$ for a selection of IRT models

IRT model	$\ln(\omega(\theta' \rightarrow \theta^*))$	$t()$
Rasch	$(\theta^* - \theta')(t(\mathbf{x}) - t(\mathbf{x}^*))$	$\sum_j x_j$
2PL	$(\theta^* - \theta')(t(\mathbf{x}, \boldsymbol{\alpha}) - t(\mathbf{x}^*, \boldsymbol{\alpha}))$	$\sum_j \alpha_j x_j$
3PL	$\sum_j (x_j - x_j^*) \ln \left(\frac{c_j + \exp(\alpha_j(\theta^* - \delta_j))}{c_j + \exp(\alpha_j(\theta' - \delta_j))} \right)$	
1PNO	$\sum_j (x_j - x_j^*) \ln \left(\frac{\Phi(\theta^* - b_j)(1 - \Phi(\theta' - b_j))}{\Phi(\theta' - b_j)(1 - \Phi(\theta^* - b_j))} \right)$	
2PNO	$\sum_j (x_j - x_j^*) \ln \left(\frac{\Phi(a_j \theta^* - b_j)(1 - \Phi(a_j \theta' - b_j))}{\Phi(a_j \theta' - b_j)(1 - \Phi(a_j \theta^* - b_j))} \right)$	
3PNO	$\sum_j (x_j - x_j^*) \left[\ln \left(\frac{c_j + (1 - c_j)\Phi(a_j \theta^* - b_j)}{c_j + (1 - c_j)\Phi(a_j \theta' - b_j)} \right) + \ln \left(\frac{1 - \Phi(a_j \theta' - b_j)}{1 - \Phi(a_j \theta^* - b_j)} \right) \right]$	
PCM	$(\theta^* - \theta')(t(\mathbf{x}) - t(\mathbf{x}^*))$	$\sum_j \sum_k x_{jk}$
GPCM	$(\theta^* - \theta')(t(\mathbf{x}, \boldsymbol{\alpha}) - t(\mathbf{x}^*, \boldsymbol{\alpha}))$	$\sum_j \alpha_j \sum_k x_{jk}$
NRM	$(\theta^* - \theta')(t(\mathbf{x}, \boldsymbol{\alpha}) - t(\mathbf{x}^*, \boldsymbol{\alpha}))$	$\sum_j \sum_k \alpha_{jk} x_{jk}$
MD2PL	$(\theta^* - \theta')^T (\mathbf{t}(\mathbf{x}, \boldsymbol{\alpha}) - \mathbf{t}(\mathbf{x}^*, \boldsymbol{\alpha}))$	$\sum_j x_j \alpha_j$

The abbreviations 2PL and 3PL stand for the Two- and Three-Parameter Logistic models; 1PNO, 2PNO, and 3PNO stand for the One-, Two-, and Three-Parameter Normal Ogive models; and MD2PL stands for the Multidimensional Two-Parameter Logistic model. We used $\Phi(x)$ as shorthand for $\int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)dy$

have to be produced for more than half a million pupils. And below, we have to sample from the posterior distribution of ability when we analyze a hierarchical IRT model for the responses from over 150, 000 pupils on a Dutch educational test. To sample from N posterior distributions, the composition algorithms would require about N times the amount of work needed to sample from a single posterior; see below. Thus, the algorithms do not become more efficient when N increases and are inefficient when N is large. The algorithms are also inefficient for applications with many items. Suppose the number of possible response patterns (or sufficient statistics) increases. In that case, the rejection algorithm will need increasingly more trials, and the SVE algorithm will tend to have lower acceptance probabilities so that the correlation between successive draws will tend to be higher.

We illustrate this with a small simulation study, the results of which are shown in Fig. 12.1. We simulate data with N persons answering to each of J dichotomous items, with N varying between 100 and 10,000, and $J \in \{10, 20, 30\}$. We assume a standard normal distribution for ability Θ . For the rejection algorithm, the IRT model is the Rasch model. For the SVE algorithm, we use the *Two-Parameter Logistic* (2PL) model. The item parameters are fixed, with difficulty parameters sampled from a standard normal distribution and discrimination parameters sampled uniformly between 1 and 3. For each combination of N and J , we generated 100

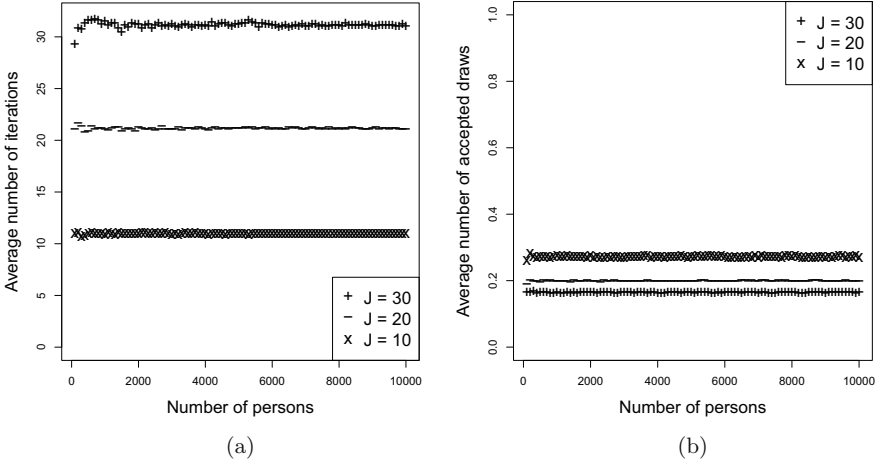


Fig. 12.1 Simulation results. (a) Number of trials for rejection. (b) Acceptance probability for SVE

data sets. With the item parameters fixed, our goal is to sample for each of the N persons an ability from the posterior distribution given his or her observed response pattern.

Results for the rejection algorithm are in Fig. 12.1a, which shows the average number of trials that are required to sample from each of the N posteriors as a function of N and J . It is clear that the average number of trials required quickly stabilizes around the number of possible realizations of $t(\mathbf{X})$, which is $J + 1$ in this simulation.³ Thus, we need approximately $(J + 1) \times N$ iterations to produce a value from each of the N posteriors, and this number grows linear in both N and J . Results for the SVE algorithm are in Fig. 12.1b which shows the average proportion of values accepted in the 100th iteration of the algorithm as a function of N and J . The acceptance probabilities are seen to be low and decreasing with an increase of the number of items. Thus, for both algorithms, it follows that as N and J grow, we need more iterations to obtain a certain amount of independent replicates from each of the N posteriors. We conclude that the algorithms, as they stand, are unsuited for applications with large N (and J).

³ The number of trials $W = w$ required to generate a realization $t(\mathbf{x})$ follows a geometric distribution with parameter $P(t(\mathbf{x}))$, the (marginal) probability to generate $t(\mathbf{x})$ under the model. From this, we see that $\mathbb{E}(W | t(\mathbf{x}))$ equals $P(t(\mathbf{x}))^{-1}$ and

$$\mathbb{E}(W) = \sum_{t(\mathbf{x})} \mathbb{E}(W | t(\mathbf{x}))P(t(\mathbf{x})),$$

where the sum is taken over all possible realizations. It follows that $\mathbb{E}(W)$ equals the number of possible realizations of $t(\mathbf{X})$.

12.3 Large-Scale Composition Sampling

The rejection and SVE algorithm sample from one posterior at the time. Consequently, sampling from N posteriors requires N times the amount of work needed to sample from a single posterior. If the algorithms are to be prepared for applications with an increasing number of posteriors, the amount of work per posterior has to decrease with N . To see how, observe that both algorithms generate samples that are not used efficiently, i.e., samples that are either rejected or accepted with a low probability. Thus, to improve the efficiency of the algorithms for increasing N , we need to make more efficient use of the generated samples. To this aim, we consider the SVE algorithm as an instance of what Tierney (1994, 1998) refers to as a *mixture of transition kernels*. This way of looking at the SVE algorithm suggests two approaches to improve its efficiency. One of these will be seen to apply to the rejection algorithm as well.

12.3.1 A Mixture Representation of the SVE Algorithm

In every realization of the SVE algorithm, we sample one of the possible response patterns (denoted \mathbf{x}^*), together with a random value for ability (denoted θ^*). The sampled ability value is a sample from the posterior distribution $f(\theta \mid \mathbf{x}^*)$ which is the proposal distribution in the SVE algorithm. The probability that we use $f(\theta \mid \mathbf{x}^*)$ as proposal distribution in the SVE algorithm is equal to $P(\mathbf{x}^*)$, which follows from the factorization:

$$P(\mathbf{x}^* \mid \theta^*)f(\theta^*) = f(\theta^* \mid \mathbf{x}^*)P(\mathbf{x}^*).$$

That is, every simulated response pattern corresponds to a unique proposal distribution and, hence, to a unique transition kernel $f(\theta^* \mid \theta, \mathbf{x}^*)$. Each of these transition kernels has the target posterior distribution as its invariant distribution; that is,

$$f(\theta^* \mid \mathbf{x}) = \int_{\mathbb{R}} f(\theta^* \mid \theta, \mathbf{x}^*)f(\theta \mid \mathbf{x}) d\theta.$$

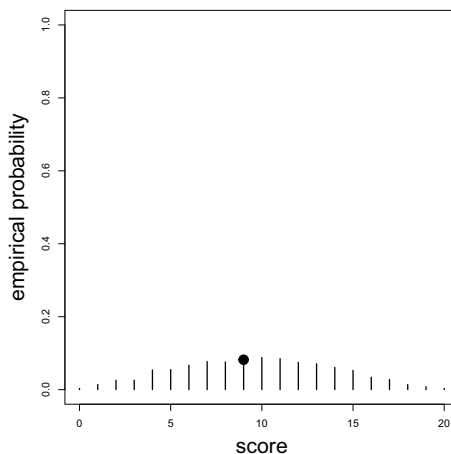
As shown by Tierney (1994), the same is true for their mixture, that is,

$$f(\theta^* \mid \mathbf{x}) = \int_{\mathbb{R}} \sum_{\mathbf{x}^*} f(\theta^* \mid \theta, \mathbf{x}^*)P(\mathbf{x}^*)f(\theta \mid \mathbf{x}) d\theta,$$

where the sum is taken over all possible response patterns, and we now see that the $P(\mathbf{x}^*)$ are the mixture weights.

To make matters concrete, consider the posterior distribution for a Rasch model with J items and a standard normal prior for ability θ . Because the Rasch model

Fig. 12.2 Empirical distribution over transition kernels for the SVE algorithm



is an exponential family model with the test score $t(\mathbf{x})$ as sufficient statistic for ability, we know that posteriors for the different ways to obtain the same test score are all the same (Dawid 1979). That is, the mixture weights are nothing but the distribution of test scores. Moreover, the posterior distributions $f(\theta \mid t(\mathbf{x}))$ are stochastically ordered by the test score, which makes the acceptance probability lower, the larger the difference between the value of $t(\mathbf{x})$ conditioned on in the target and $t(\mathbf{x}^*)$ conditioned on in the proposal distribution; see Table 12.1. Figure 12.2 shows the mixture probabilities $P(t(\mathbf{x}))$ for a test of 20 items. We see in Fig. 12.2 that the SVE algorithm will tend to generate many transition kernels for which the acceptance probability is not very high.

12.3.2 Oversampling

Since the SVE algorithm tends to frequently generate transition kernels for which the acceptance probability is low, we consider changing the mixture probabilities, in such a way that more probability mass is concentrated on transition kernels with high acceptance probability.

Suppose that instead of simulating a single proposal value θ^* , with a corresponding single response pattern \mathbf{x}^* , we simulate a number of i.i.d. proposal values, each with its own response pattern. From those, we choose the one for which the test score is closest to the test score conditioned on in the target distribution, and hence the acceptance rate tends to be the highest.

In Fig. 12.3, we illustrate the effectiveness of this oversampling approach in sampling from a posterior $f(\theta \mid t(\mathbf{x}) = 9)$. Clearly, even with 5 samples, we already improve the probability to generate directly from the target from close to 0.1 to close to 0.4. With 20 samples, this probability even exceeds 0.8. Moreover, if the proposal is not identical to the target, it is increasingly more likely to be close to the target as the number of samples increases.

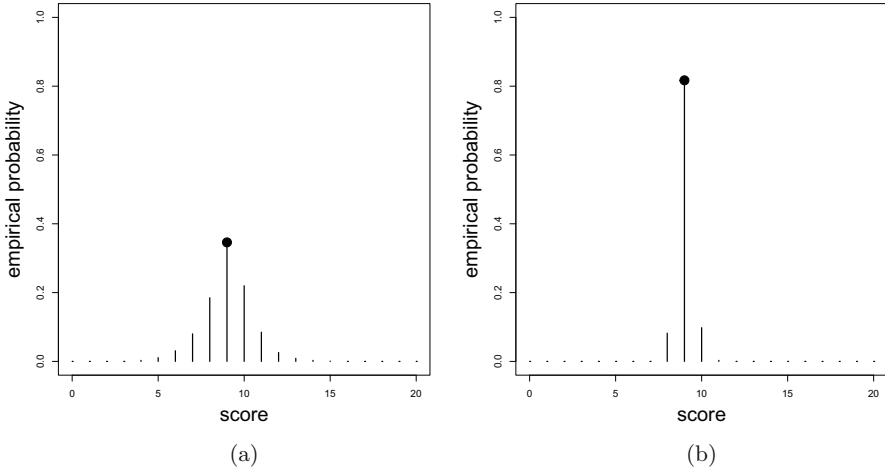


Fig. 12.3 Probability distribution over transition kernels after modulating the mixture probabilities. **(a)** 5 samples. **(b)** 20 samples

Since oversampling can easily be implemented in a parallel implementation, this approach need not lead to a large increase in computer time. This makes the approach computationally attractive.

12.3.3 Matching

Consider the situation where there are many proposal distributions (i.e., N large) and hence many target posterior distributions, each one independent from the others. The SVE algorithm can once again be considered as a mixture of transition kernels for the whole collection of N posteriors:

$$f(\theta^* | \underline{\mathbf{x}}) = \int_{\mathbb{R}^N} \prod_i f(\theta_i^* | \theta_i, \mathbf{x}_i^*) P(\underline{\mathbf{x}}^*) f(\theta | \underline{\mathbf{x}}) d\theta,$$

where $\underline{\mathbf{x}}$ denotes the matrix $\underline{\mathbf{x}} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. Observe that the transition kernel for person i only depends on $\underline{\mathbf{x}}^*$ via the i -th response pattern. Suppose that for a matrix $\underline{\mathbf{x}}^*$, we permute the person indices i , in some fixed way (denoted $\text{perm}(i)$). Then, the transition kernel for person i depends on $\underline{\mathbf{x}}^*$ via one of the response patterns in $\underline{\mathbf{x}}^*$, and every response pattern is used exactly once:

$$f(\theta^* | \underline{\mathbf{x}}) = \int_{\mathbb{R}^N} \prod_i f(\theta_{\text{perm}(i)}^* | \theta_i, \mathbf{x}_{\text{perm}(i)}^*) P(\underline{\mathbf{x}}^*) f(\theta | \underline{\mathbf{x}}) d\theta.$$

Clearly, not all proposal distributions lead to the same acceptance probability, and thus, not all permutations lead to the same overall acceptance rate. Hence, some permutations work better than others. Notice that all permutations lead to a valid transition kernel with the posterior distribution as its invariant distribution, as long as our permutation strategy does not depend on Θ' or Θ^* . Finding, for every matrix \mathbf{x}^* and every observed matrix \mathbf{x} , the best permutation will in general be an NP-complete problem. However, the better the permutation, the more efficient the algorithm.

In Algorithm 3, we consider the general situation where each person may receive its own prior distribution, and we denote the prior of a person i with $f_i(\theta)$. We generate a proposal using the prior $v = 1, \dots, N$ (v now indexes the proposals), and we reorder the index vector $\mathbf{V} = [v_i]$ of the proposals by using a *permutation function* $\text{perm}()$. When we use $\theta_v^* \sim f_v(\theta \mid \mathbf{x}_v^*)$ as a proposal for a posterior $f_i(\theta \mid \mathbf{x}_i)$ (i need not equal v), then we accept θ_v^* with probability $\pi(\theta'_i \rightarrow \theta_v^*) = \min\{1, \omega(\theta'_i \rightarrow \theta_v^*)\}$, and

$$\omega(\theta'_i \rightarrow \theta_v^*) = \frac{f_i(\theta^* \mid \mathbf{x}_i) f_v(\theta' \mid \mathbf{x}^*)}{f_i(\theta' \mid \mathbf{x}_i) f_v(\theta^* \mid \mathbf{x}^*)} = \frac{h(\mathbf{x}_i; \theta^*) h(\mathbf{x}^*; \theta')}{h(\mathbf{x}_i; \theta') h(\mathbf{x}^*; \theta^*)} \times \frac{f_i(\theta^*) f_v(\theta')}{f_i(\theta') f_v(\theta^*)},$$

a product of likelihood ratios times a product of prior ratios, where the normalizing constants $P(\mathbf{x})$ and $Z(\theta)$ cancel as before (as do the normalizing constants of the prior distributions).

Algorithm 3 Single-Variable Exchange algorithm with matching

Require: Index vector $\mathbf{V} = [v_i] = i$, for $i = 1, 2, \dots, N$

Require: A permutation function $\text{perm}()$

```

1: for  $v = 1$  to  $N$  do
2:   Generate  $\theta_v^* \sim f_v(\theta)$ 
3:   Generate  $\mathbf{x}_v^* \sim P(\mathbf{X} \mid \theta_v^*)$ 
4: end for
5: Match proposals to targets by rearranging  $\mathbf{V}$  based on  $\text{perm}()$ .
6: for  $i = 1$  to  $N$  do
7:   Set  $v = v_i$ 
8:   Draw  $u \sim \mathcal{U}(0, 1)$ 
9:   if ( $u < \pi(\theta'_i \rightarrow \theta_v^*)$ ) then
10:    Set  $\theta'_i = \theta_v^*$ 
11:   end if
12: end for

```

Simple permutation functions are often readily available. For instance, the test score is usually correlated with Θ and gives a simple procedure to permute the indices of proposals and targets. When the IRT model is a member of the exponential family, the sufficient statistic $t(\mathbf{x})$ contains all information about Θ from the data and gives another simple procedure for permutation. More general solutions would be the use of maximum likelihood or Bayes' modal estimates, when they are not too expensive to compute. We give some examples of permutation strategies in our applications below.

12.3.4 Recycling in the Rejection Algorithm

The main idea underlying matching is that a proposal need not be associated to one particular posterior. We can use the same idea for the rejection algorithm for the situation with N posteriors using a common prior $f(\theta)$. The idea behind recycling is that if we sample $\{\theta^*, \mathbf{x}^*\}$, θ^* can be assigned to *any* observation i where $t(\mathbf{x}_i) = t(\mathbf{x}^*)$ (or $\mathbf{x}_i = \mathbf{x}^*$). In general, we need to sample from $N = \sum_{u=1}^U n_u$ posteriors $f(\theta \mid t(\mathbf{x}) = t_u)$, where t_u is one of the U unique values the statistic $t(\mathbf{X})$ can take, N_u is the number of observations of response patterns \mathbf{x}_i for which $t(\mathbf{x}_i) = t_u$, and it is arbitrary how the values of $t(\mathbf{X})$ are indexed. As seen in Algorithm 4, we sample from the joint distribution of Θ and \mathbf{X} until we have n_u values for each u . In Algorithm 4, we store generated values in a vector \mathbf{R} and the index corresponding to the generated statistic in a vector \mathbf{S} . If necessary, we can use \mathbf{S} to assign the drawn parameters to observations. Note that this version of the rejection algorithm has been implemented in the R-package `dexter` (Maris et al. n.d.).

Algorithm 4 A rejection algorithm with recycling

Require: n_u for $u = 1, 2, \dots, U$.

Require: A counter c and vectors $\mathbf{R} = [r_i]$ and $\mathbf{S} = [s_i]$, $i = 1, 2, \dots, N$.

```

1:  $c = 0$ .
2: repeat
3:   Generate  $\theta^* \sim f(\theta)$ 
4:   Generate  $\mathbf{x}^* \sim P(\mathbf{X} \mid \theta^*)$ 
5:   Determine  $u$ , such that  $t(\mathbf{x}^*) = t_u$ 
6:   if  $n_u \geq 1$  then
7:      $n_u = n_u - 1$ 
8:      $c = c + 1$ 
9:      $[r_c] = \theta^*$ 
10:     $[s_c] = u$ 
11:   end if
12: until  $n_u = 0$  for  $u = 1, \dots, U$ 

```

In the context of IRT, the situation with N posteriors using a common prior describes the situation of N persons sampled from the same population. In practice, however, we often encounter situations where the persons are sampled from different groups, e.g., boys and girls. In this situation, posteriors are of the form

$$f(\theta \mid \mathbf{x}_i) \propto P(\mathbf{x}_i \mid \theta) f_m(\theta),$$

i.e., persons are grouped into marginals m , where $f_m(\theta)$ denotes the prior distribution in marginal m , and recycling applies to each marginal separately. It will be clear that in this situation, the algorithm becomes efficient only when there are many persons in each marginal. When the prior distributions are person specific, and each person has its own marginal distribution, *recycling* reduces to the standard rejection algorithm.

12.3.5 Has the Efficiency of the Algorithms Improved?

We considered *recycling* and *matching* as ways to improve the rejection and SVE algorithm when samples are required from many posteriors. To illustrate that this works, we compare the efficiency of the rejection algorithm with and without recycling and the SVE algorithm with and without matching under the conditions of our previous simulation.

Results for the rejection algorithm with recycling are in Fig. 12.4a, which shows the average number of trials required to sample from the N posteriors as a function of N and J . If we compare the results in Fig. 12.4a with the results in Fig. 12.1a, we see that recycling requires relatively few iterations per posterior. Note that the required number of iterations decreases as N increases and increases when J increases. It is clear from Fig. 12.4a that as both N and J increase, recycling makes the rejection algorithm more efficient when N increases faster than J . For fixed J , Fig. 12.4a confirms that as N becomes large, the number of iterations per posterior tends to 1.

To illustrate that the *matching* procedure improves the efficiency of the SVE algorithm, we consider the following simple strategy. We order target distributions using the statistic $t(\mathbf{x}_i, \boldsymbol{\alpha}) = \sum_{j=1}^J x_{ij} \alpha_j$ (see Table 12.1), such that the values of the statistic are ordered from small to large, and we do the same for the proposal distributions using the $t(\mathbf{x}^*, \boldsymbol{\alpha})$. This simple permutation strategy ensures that if the Markov chain is stationary, the first proposal is likely to be a good proposal for the first target (since the difference between $t(\mathbf{x}, \boldsymbol{\alpha})$ and $t(\mathbf{x}^*, \boldsymbol{\alpha})$ is likely to be small), and the same holds for the second, the third, and so on. Results for the SVE algorithm using this procedure are given in Fig. 12.4b, which shows the average acceptance rate in the 100th iteration of the algorithm as a function of N and J .

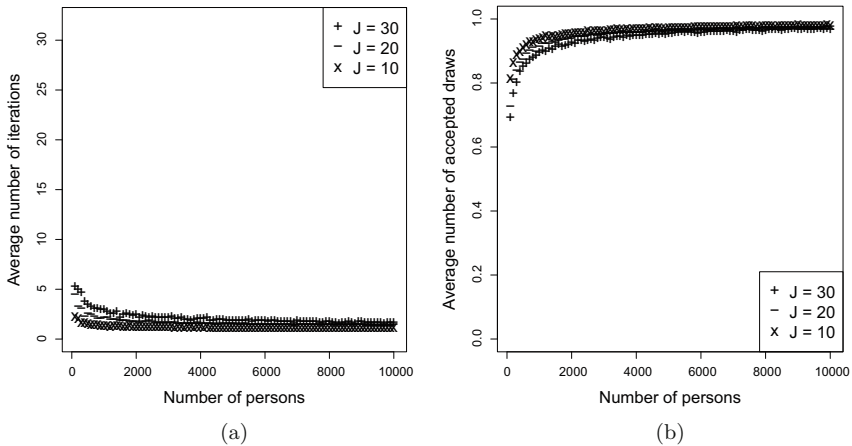


Fig. 12.4 Simulation results. (a) Number of trials with recycling. (b) Proportion accepted with matching

If we compare the results in Fig. 12.4b with the results in Fig. 12.1b, we see that matching results in much higher acceptance rates. Note that, similar to the results for the recycling, the proportion of accepted values increase as N increases and decrease as J increases and matching makes the SVE algorithm more efficient when N increases faster than J . For fixed J , Fig. 12.4b confirms that as N becomes large, the average acceptance rate tends to 1.

We conclude that recycling and matching make sampling from a large number of posteriors entirely feasible. Most appealing is that the efficiency improves as a function of N . As N tends to infinity, this means that we need to generate the data only once to obtain a draw from each of N posteriors and both algorithms generate i.i.d. from each of the N posteriors. For moderate N , we can already see that the number of trials needed for the rejection algorithm approaches 1 and that the acceptance rate of the SVE algorithm approaches 1. This shows that, even for moderate N , both algorithms require little more than one generated data set and that the SVE algorithm is close to sampling i.i.d.

To illustrate that matching makes the autocorrelation in the SVE algorithm a decreasing function of N , we perform a small simulation. We run 5000 Markov chains for 500 iterations each. We use the 5000 Markov chains to estimate the autocorrelation by correlating the 5000 draws in some iteration i and iteration $i + 1$, $i + 2$, \dots . Figure 12.5 shows the autocorrelation spectra for the SVE algorithm with matching. In Fig. 12.5, we see that the autocorrelations are a decreasing function of N , meaning that as N becomes sufficiently large, we sample approximately i.i.d.

12.3.6 How Do Our Algorithms Compare to Existing Algorithms?

When it is difficult to sample from $f(\theta \mid \mathbf{x})$ directly, it is sometimes easier to sample from a more complex (augmented) posterior distribution $f(\theta, \mathbf{y} \mid \mathbf{x})$ using the Gibbs sampler. In the context of educational measurement, this approach has been advocated by Albert (1992) for Normal Ogive models and by Jiang and Templin (2018, 2019) for logistic IRT models. Due to the use of conditioning in the Gibbs sampler, the *data augmentation* procedure of Albert (1992) introduces a constant amount of autocorrelation to the Markov chain (Liu et al. 1994). As a result, the number of iterations that are required to obtain a fixed amount of independent replicates from each of the N posteriors is linear in N . In this sense, our algorithms scale better, since the amount of autocorrelation reduces as a function of N .

A more general approach to sampling from $f(\theta \mid \mathbf{x})$ is to sample a proposal value θ^* from a conditional distribution $f(\theta^* \mid \theta')$ and use the Metropolis-Hastings algorithm to either move to the proposed value θ^* or stay at the current state θ' . This approach has been advocated by Patz and Junker (1999), who suggest to use $f(\theta^* \mid \theta') = \mathcal{N}(\theta', \sigma^2)$ as proposal distribution (i.e., a *random walk*). Setting the value of σ^2 in the proposal distribution requires some effort from the user (Rosenthal

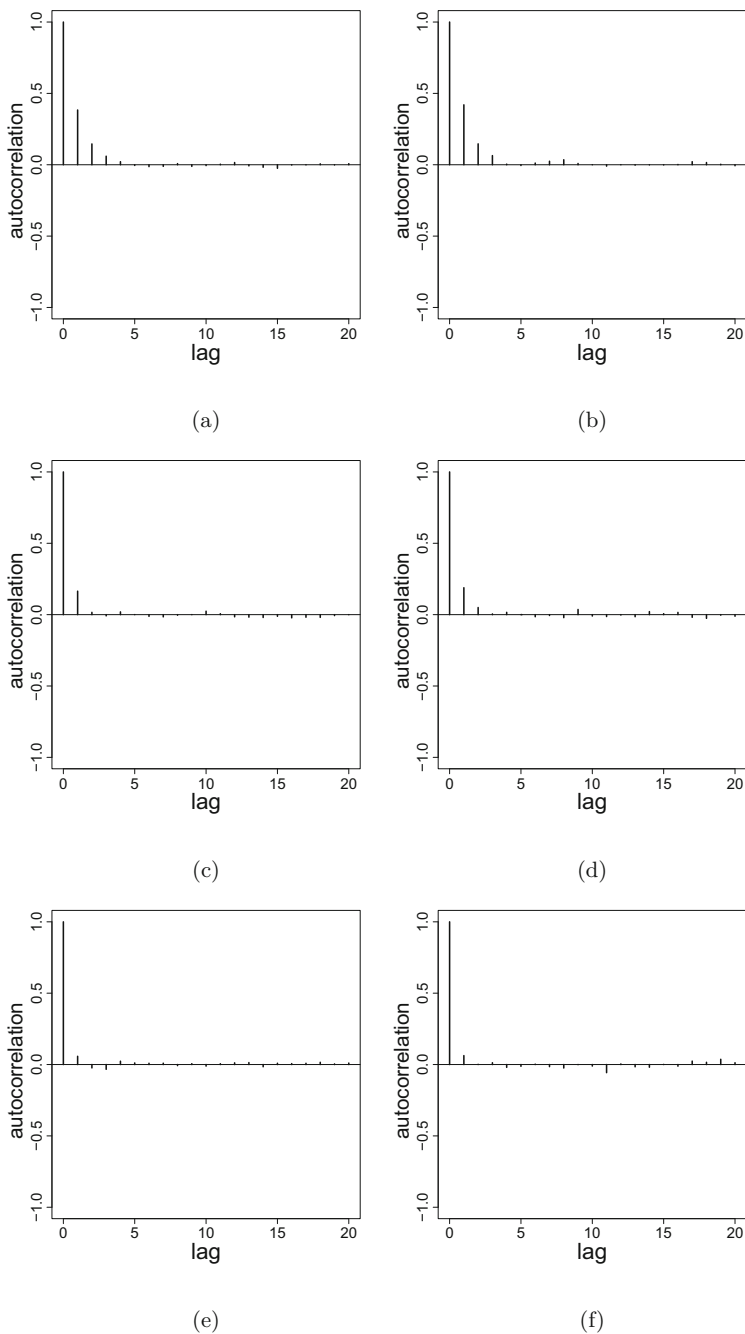


Fig. 12.5 Estimated autocorrelation spectra using $J = 30$ items. (a) $N = 100$ persons, $\theta = 0$. (b) $N = 100$ persons, $\theta = 0.5$. (c) $N = 1000$ persons, $\theta = 0$. (d) $N = 1000$ persons, $\theta = 0.5$. (e) $N = 10,000$ persons, $\theta = 0$. (f) $N = 10,000$ persons, $\theta = 0.5$

2011): when σ^2 is too large, most samples are rejected, but when σ^2 is too small, only small steps are taken, and the chain does not mix properly. To overcome this problem altogether, one could use an unconditional proposal distribution $g(\theta)$ (i.e., an *independence chain*). This is the approach we took in this chapter. Whenever the proposal distribution $g(\theta)$ closely resembles the target distribution, the Metropolis-Hastings algorithm is very efficient. In general, it can be difficult to find good proposal distributions, but the matching procedure automatically finds proposal distributions $g(\theta | \mathbf{x}^*)$ that closely resemble the target $f(\theta | \mathbf{x})$, and as N increases, this procedure becomes more likely to generate good proposal distributions.

12.4 Simulated and Real-Data Examples

In this section, we discuss three examples illustrating the practical use of the SVE algorithm for Bayesian estimation using the Gibbs sampler. The Gibbs sampler is an abstract divide-and-conquer algorithm that generates a dependent sample from a multivariate posterior distribution. In each iteration, the algorithm generates a sample from the distribution of each variable in turn, conditional on the current values of the other variables. These are called the *full-conditional distributions*. It can be shown that the sequence of samples constitutes a Markov chain and the stationary distribution of that Markov chain is the joint posterior distribution of interest.

In each of our examples, there will be one or more full-conditional distributions that are not easily sampled from, and we use the SVE algorithms developed in this chapter to sample from these full-conditional distributions. All analyses have been performed using a Dell OptiPlex 980 PC with an Intel Core 5 CPU and clock speed 3.20 Ghz and 4Gb of memory running on Windows 7 Enterprise(32 bit) with a single core.

12.4.1 Gamma Regression

The random-effects gamma model is a model for responses times proposed by Fox (2013) as an alternative to the log-normal model that is commonly used (van der Linden 2007; Klein Entink et al. 2009). The model is difficult to estimate, because the normalizing constant of the gamma distribution (i.e., the gamma function $\Gamma(\cdot)$) is not available in closed form and can produce overflow errors in its computation. We develop a Gibbs sampler for this model to illustrate how the SVE algorithm can be used to avoid the calculation of the gamma function.

Let X_{ij} denote the time needed by person i to respond to item j ; $i = 1, \dots, N$, and $j = 1, \dots, J$. The X_{ij} are assumed to be independent, gamma distributed random variables with

$$f(\mathbf{x} | \boldsymbol{\lambda}, \boldsymbol{\eta}) = \prod_{i=1}^N \prod_{j=1}^J \frac{\lambda_{ij}^{\eta_{ij}}}{\Gamma(\eta_{ij})} x_{ij}^{\eta_{ij}-1} \exp\{-x_{ij} \lambda_{ij}\}. \quad (12.3)$$

In the model of Fox (2013), a relatively simple regression structure was used, namely, $\lambda_{ij} = \nu/(2\theta_i)$ and $\eta_{ij} = \nu/2$. We will use a slight alteration in this simulation, with $\lambda_{ij} = \nu/(\theta_i \delta_j)$ and $\eta_{ij} = \nu$, such that $\mathbb{E}[X_{ij}] = \theta_i \delta_j$, and $\text{Var}(X_{ij}) = \mathbb{E}[X_{ij}]^2/\nu$. The person parameter $\theta_i > 0$ represents the speed of person j , the item parameter $\delta_j > 0$ the time intensity of item j , and ν a common rate parameter. We further assume that $\theta_i \sim \text{ln}\mathcal{N}(\mu_\theta, \sigma_\theta^2)$, and $\delta_j \sim \text{ln}\mathcal{N}(\mu_\delta, \sigma_\delta^2)$, where $\text{ln}\mathcal{N}(\mu, \sigma^2)$ denotes the log-normal distribution with mean μ and variance σ^2 . The location and scale parameters of the person and item parameters are unknown and are to be estimated. To complete the specification of the model, we use the following priors: $\nu \sim \Gamma(a, b)$, $f(\mu_\theta, \sigma_\theta^2) \propto \sigma_\theta^{-2}$, and $f(\mu_\delta, \sigma_\delta^2) \propto \sigma_\delta^{-2}$.

Given the person and item parameters, the location and scale parameters are easily sampled from their full-conditional distributions (Gelman et al. 2004):

$$\begin{aligned} f(\mu_\theta | \boldsymbol{\theta}, \sigma_\theta^2) &\propto \mathcal{N}\left(\frac{1}{N} \sum_{i=1}^N \ln(\theta_i), \sigma_\theta^2/N\right) \\ f(\sigma_\theta^2 | \boldsymbol{\theta}) &\propto \text{Inv-}\chi^2\left(N-1, \frac{1}{N-1} \sum_{i=1}^N \left(\ln(\theta_i) - \frac{1}{N} \sum_{i=1}^N \ln(\theta_i)\right)^2\right) \\ f(\mu_\delta | \boldsymbol{\delta}, \sigma_\delta^2) &\propto \mathcal{N}\left(\frac{1}{J} \sum_{j=1}^J \ln(\delta_j), \sigma_\delta^2/J\right) \\ f(\sigma_\delta^2 | \boldsymbol{\delta}) &\propto \text{Inv-}\chi^2\left(J-1, \frac{1}{J-1} \sum_{j=1}^J \left(\ln(\delta_j) - \frac{1}{J} \sum_{j=1}^J \ln(\delta_j)\right)^2\right). \end{aligned}$$

The full-conditional distribution of ν , the person, and the item parameters, however, are not easily sampled from, and for these, we will use the SVE algorithms developed in this chapter.

To sample from the full-conditional distribution of ν , we generate ν^* from the prior $f(\nu | a, b)$ and generate a data matrix \mathbf{x}^* from $f(\mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\delta}, \nu^*)$. The probability $\pi(\nu' \rightarrow \nu^*)$ to make a transition from ν^* to ν' using this set-up is then equal to $\min\{1, \omega(\nu' \rightarrow \nu^*)\}$, with

$$\ln \omega(\nu' \rightarrow \nu^*) = (\nu^* - \nu')(t(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\delta}) - t(\mathbf{x}^*, \boldsymbol{\theta}, \boldsymbol{\delta})),$$

where

$$t(\underline{\mathbf{x}}, \boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{i=1}^N \sum_{j=1}^J \left(\ln(x_{ij}) - \frac{x_{ij}}{\theta_i \delta_j} \right).$$

Note that we do not need to evaluate the $\Gamma()$ function at v' or v^* , making $\ln \omega$ a relatively simple function to compute.

We have seen earlier that in this set-up, the SVE algorithm is likely to generate transition kernels for which the acceptance probability is low. We therefore use the *oversampling* procedure. That is, we generate a number of i.i.d. proposal values v^* , each with its own data matrix $\underline{\mathbf{x}}^*$. From these, we choose the one for which the statistic $t(\underline{\mathbf{x}}^*, \boldsymbol{\theta}, \boldsymbol{\delta})$ is closest to $t(\underline{\mathbf{x}}, \boldsymbol{\theta}, \boldsymbol{\delta})$. We use 100 proposals in this example. The R-code that we used for this full-conditional is given in Appendix A.

To sample from the full-conditional distributions of the person and the item parameters, we use the matching procedure. Since we use the same matching procedure for the person and the item parameters, we only describe the procedure for the person parameters. We generate θ_v^* , $v = 1, \dots, N$, from $f(\theta \mid \mu_\theta, \sigma_\theta^2)$ and use it to generate a vector of response times \mathbf{x}_v^* from $f(\mathbf{x} \mid \theta_v^*, \boldsymbol{\delta}, v)$. Say that we use $f(\theta \mid \mathbf{x}_v^*, v, \mu_\theta, \sigma_\theta)$ as proposal for a target i (i need not equal v), the probability $\pi(\theta'_i \rightarrow \theta_v^*)$ to make a transition from θ'_i to θ_v^* is then equal to $\min\{1, \omega(\theta'_i \rightarrow \theta_v^*)\}$, with

$$\ln \omega(\theta'_i \rightarrow \theta_v^*) = v \left(\frac{1}{\theta_v^*} - \frac{1}{\theta'_i} \right) (t(\mathbf{x}_v^*, \boldsymbol{\delta}) - t(\mathbf{x}_i, \boldsymbol{\delta})), \quad (12.4)$$

where

$$t(\mathbf{x}_i, \boldsymbol{\delta}) = \sum_{j=1}^J \frac{x_{ij}}{\delta_j}.$$

Note again that we do not need to evaluate the $\Gamma()$ function in $\ln \omega$ and the acceptance probabilities are simple to compute.

From (12.4), we see that it is opportune to use $t(\mathbf{x}_i, \boldsymbol{\delta})$ to permute proposals and targets. To this aim, we compute $t(\mathbf{x}, \boldsymbol{\delta})$ for each person in the sample and for each proposal. Then, we order the targets using the $t(\mathbf{x}_i, \boldsymbol{\delta})$, such that the corresponding statistics are ordered from small to large, and do the same for the proposals using the $t(\mathbf{x}_v^*, \boldsymbol{\delta})$. This simple permutation strategy ensures that if the Markov chain is stationary, the first proposal is likely to be a good proposal for the first target (since the difference between $t(\mathbf{x}, \boldsymbol{\delta})$ and $t(\mathbf{x}^*, \boldsymbol{\delta})$ is likely to be small) and the same holds for the second, the third, and so on. The R-code that we used for this full-conditional is given in Appendix B.

To see how it works, we simulated data for $N = 10,000$ persons on a test consisting of $J = 40$ items. We set the mean and variance of the person and the item parameters equal to 10 and 1, respectively, from which we can solve for the

location and scale parameters in the log-normal model. Using these location and scale parameters, we sample the person and item parameters from the log-normal model. The parameter ν was set equal to 40.

Note that the gamma model that we use is not identified, since multiplying the person parameters with a constant and dividing the item parameters with the same constant give the same model. Since we know the true values of the parameters in this simulation, we simply set the estimated parameter of the first item equal to its true value.

We ran the Gibbs sampler for 2000 iterations, which took approximately 2.5 h (about 4.7 s per iteration). The main computational cost of this Gibbs sampler resides in sampling the entire $N \times J$ data matrix $m + 2$ times in each iteration, of which $m = 100$ times for sampling from the full-conditional of ν . Since the cost per iteration is the same in each iteration, we see that we need approximately 0.1 s to sample the person and the item parameters in each iteration and approximately 4.6 s to sample ν . This means that we can reduce the computational time by reducing m . Note, however, that this would also reduce the acceptance rate in sampling ν .

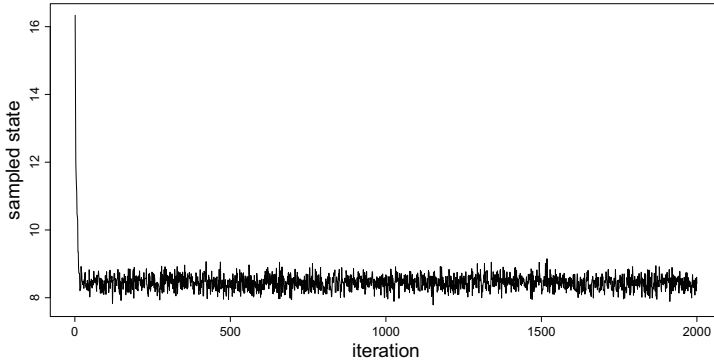
The results are in Figs. 12.6 and 12.7. As expected, our use of the SVE algorithm does not lead to high acceptance rates for the item parameters; the average acceptance rate was 0.05. The main reason is that we only generate 40 proposals to assign to 40 targets, with a large variation on the conditioning statistic $t(\mathbf{x}, \theta)$ due to the large number of observations. In the next example, we show that the oversampling procedure can be used to remedy this. We did obtain high efficiency for the person parameters, with an average acceptance rate of 0.96. In Fig. 12.6, we show the trace plot for a person and an item parameter. It is clear that both converge quickly to the stationary distribution. In Fig. 12.7, we show scatterplots of the true person and item parameters against the parameter states in iteration 2000, which illustrates that we are able to recover the parameters of the generating model. Finally, the proportion of accepted values for the ν parameter equalled 0.30, which is certainly reasonable for such a complex full-conditional distribution. In Fig. 12.6c, we show the trace plot of ν , from which we see that once the person and item parameters converge, ν also quickly converges to its stationary distribution.

12.4.2 The Amsterdam Chess Test Data

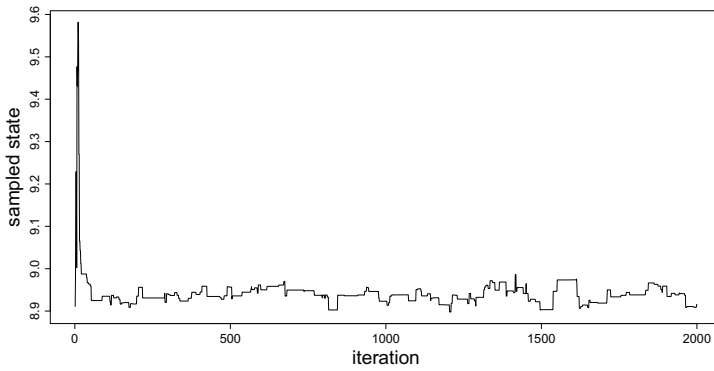
The *Signed Residual Time (SRT) model* is an exponential family IRT model for item response accuracy and response times and is derived by Maris and van der Maas (2012) from the following scoring rule:

$$(2X_{ij} - 1)(d - S_{ij}),$$

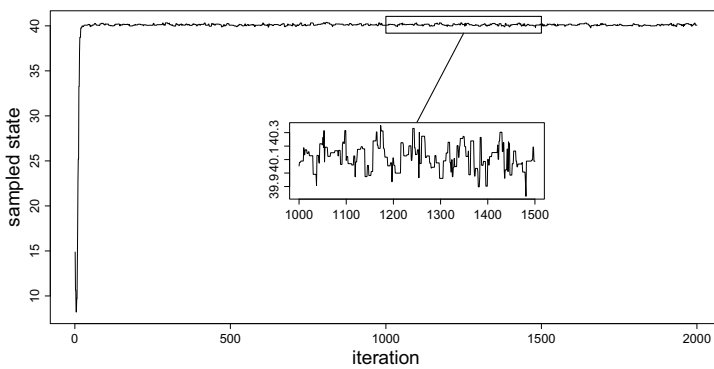
for an item response X_{ij} , which equals 1 if the response is correct and 0 if incorrect, after S_{ij} time units when the time limit for responding is d . This scoring rule assigns



(a)



(b)



(c)

Fig. 12.6 Trace plot of ν , a person, and an item parameter in the gamma mixture example. (a) Trace plot of a person parameter. (b) Trace plot of an item parameter. (c) Trace plot of ν

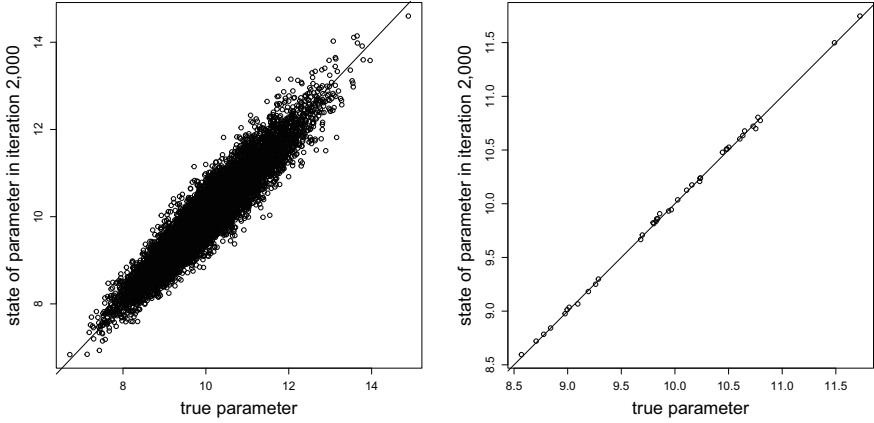


Fig. 12.7 Scatterplot of the true person (item) parameters at the states of the person (item) parameters in iteration 2000 of the Gibbs sampler for the gamma mixture example. (a) Scatterplot of the person parameters. (b) Scatterplot of the item parameters

the residual time as the score for a correct response and minus the residual time for an incorrect response. Thus, subjects need to be both fast and accurate to obtain a high score and, thereby, a high estimated ability. The SRT model is

$$f(X_{ij} = x_{ij}, S_{ij} = s_{ij} | \theta_i, \delta_j, d) = (\theta_i - \delta_j) \frac{\exp[(2x_{ij} - 1)(d - s_{ij})(\theta_i - \delta_j)]}{\exp[d(\theta_i - \delta_j)] - \exp[-d(\theta_i - \delta_j)]},$$

for $0 \leq s \leq d$. The statistics

$$t(\mathbf{x}_i, \mathbf{s}_i) = \sum_{j=1}^J (2x_{ij} - 1)(d - s_{ij}) \tag{12.5}$$

$$t(\mathbf{x}_j, \mathbf{s}_j) = - \sum_{i=1}^N (2x_{ij} - 1)(d - s_{ij})$$

are sufficient for the ability θ_i of a person i and the difficulty δ_j of an item j , respectively. We assume that $\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$ and $\delta_j \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2)$, and to complete specification of the model used the following priors: $f(\mu_\theta, \sigma_\theta^2) \propto \sigma_\theta^{-2}$ and $f(\mu_\delta, \sigma_\delta^2) \propto \sigma_\delta^{-2}$.

Given the person and item parameters, the location and scale parameters are easily sampled from their full-conditional distributions (Gelman et al. 2004):

$$\begin{aligned}
 f(\mu_\theta \mid \boldsymbol{\theta}, \sigma_\theta^2) &\propto \mathcal{N}\left(\frac{1}{N} \sum_{i=1}^N \theta_i, \sigma_\theta^2/N\right) \\
 f(\sigma_\theta^2 \mid \boldsymbol{\theta}) &\propto \text{Inv-}\chi^2\left(N-1, \frac{1}{N-1} \sum_{i=1}^N \left(\theta_i - \frac{1}{N} \sum_{i=1}^N \theta_i\right)^2\right) \\
 f(\mu_\delta \mid \boldsymbol{\delta}, \sigma_\delta^2) &\propto \mathcal{N}\left(\frac{1}{J} \sum_{j=1}^J \delta_j, \sigma_\delta^2/J\right) \\
 f(\sigma_\delta^2 \mid \boldsymbol{\delta}) &\propto \text{Inv-}\chi^2\left(J-1, \frac{1}{J-1} \sum_{j=1}^J \left(\delta_j - \frac{1}{J} \sum_{j=1}^J \delta_j\right)^2\right).
 \end{aligned}$$

The full-conditional distributions of the person and item parameters are not easily sampled from, and we will use an SVE algorithm to sample from these full-conditional distributions. To save space, we will only describe the procedure for the person parameters.

We generate θ_v^* , $v = 1, \dots, N$, from $f(\theta \mid \mu_\theta, \sigma_\theta^2)$ and use it to generate a vector of item responses \mathbf{x}_v^* and response times \mathbf{s}_v from $f(\mathbf{x}, \mathbf{s} \mid \theta_v^*, \boldsymbol{\delta})$ (see Appendix C). Say that we use $f(\theta \mid \mathbf{x}_v^*, \mathbf{s}_v^*, \mu_\theta, \sigma_\theta)$ as proposal for a target i (i need not equal v), the probability $\pi(\theta'_i \rightarrow \theta_v^*)$ to make a transition from θ'_i to θ_v^* is then equal to $\min\{1, \omega(\theta'_i \rightarrow \theta_v^*)\}$, with

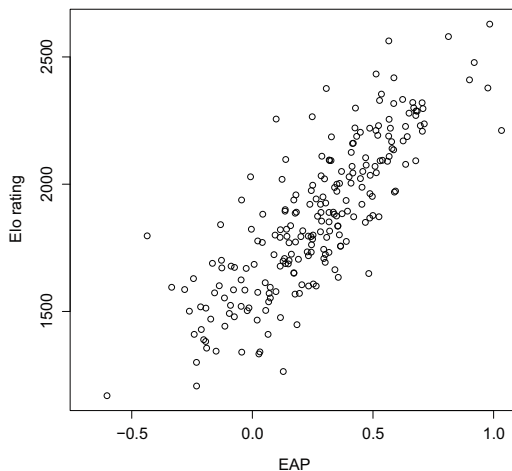
$$\ln \omega(\theta'_i \rightarrow \theta_v^*) = (\theta_v^* - \theta'_i) (t(\mathbf{x}_v^*, \mathbf{s}_v^*) - t(\mathbf{x}_i, \mathbf{s}_i)),$$

with $t(\mathbf{x}_i, \mathbf{s}_i)$ defined in (12.5).

Although the sufficient statistics (12.5) can be used to permute the indices of targets and proposals, we only have a few person and item parameters in this example. To obtain some efficiency of the SVE algorithm in this application, we use a variant of the *oversampling* strategy. In each iteration, we generate a number of i.i.d. proposals and for each target distribution choose the proposal for which the statistic $t(\mathbf{x}^*, \mathbf{s}^*)$ is closest to the observed statistic $t(\mathbf{x}, \mathbf{s})$ while ensuring that each proposal is used only once.

Van der Maas and Wagenmakers (2005) describe data from the Amsterdam Chess Test (ACT), collected during the 1998 open Dutch championship in Dieren, the Netherlands. The data we consider consists of the accuracy and response times of $N = 259$ subjects on $J = 80$ choose-a-move items administered with a time limit of 30 s. We started the mean and variance of the person and item parameters at 0 and 1, respectively. Using these values, we sampled the person and item parameters from the prior. In each iteration, we generated $2 \times N = 498$ proposals for the

Fig. 12.8 Scatterplot of EAP versus Elo rating in the ACT example



persons and $5 \times J = 400$ proposals for the items. We ran the Gibbs sampler for 10,000 iterations, which took approximately 12 min (about 0.07 s per iteration). The average acceptance rate was 0.98 for the persons and 0.93 for the items.

An important advantage of this illustrative application is that for chess expertise, an established external criterion is available in the form of the Elo ratings of chess players, which has high predictive power for game results. For those 225 participants for whom a reliable Elo rating was available, we correlated the *expected a posteriori* (EAP) estimates with their Elo ratings. The results are given in Fig. 12.8. The correlation between EAP estimates and Elo ratings is equal to 0.822.

12.4.3 The 2012 Eindtoets Data

In educational measurement, population models are commonly used to describe structure in the distribution of the latent abilities. For example, in equating two exams, one can characterize the two exam groups by using a normal distribution with a group-specific mean and variance; in the analyses of tests consisting of different scales, a multivariate normal distribution can be used to characterize the latent correlations; and in educational surveys, a normal regression model can be used to study the effects of covariates on the ability distribution. Whenever the abilities are observed, inference is relatively straightforward in each of these situations. Our focus in this section is to show how the SVE algorithm can be used to sample from the full-conditional distribution of the latent abilities, allowing the analyses of structural IRT models using the Gibbs sampler, even for large data sets.

We use response data of $N = 158,637$ Dutch end of primary school pupils on the 2012 Cito Eindtoets to illustrate our approach using a multidimensional IRT model. In specific, we used data from the non-verb spelling (10 items), verb spelling

(10 items), reading comprehension (30 items), basic arithmetic (14 items), fractions (20 items), and geometry (15 items) scales. That is, we have six unidimensional IRT models (a between multidimensional IRT model) and use a multivariate normal distribution to infer about the latent correlations between the six scales. To keep our focus on sampling the latent abilities, we assume that an IRT model is given (i.e., the parameters characterizing the items in the IRT model are known). For simplicity, we use the Rasch model for each of the scales in our example and fix the item parameters at the conditional maximum likelihood (CML) estimates.

We use a multivariate normal distribution with an unknown $Q \times 1$ vector of means $\boldsymbol{\mu}$ and $Q \times Q$ covariance matrix $\boldsymbol{\Sigma}$ to describe the latent correlations between the $Q = 6$ dimensions. To complete the model, we use the multivariate Jeffreys prior for the mean vector and the covariance matrix:

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{Q+1}{2}}.$$

The Gibbs sampler is used to sample from the joint posterior distribution $f(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \underline{\mathbf{x}})$. For this model, the full-conditional distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are easily sampled from (Gelman et al. 2004):

$$\begin{aligned} f(\boldsymbol{\mu} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) &\propto \mathcal{N}_Q(\bar{\boldsymbol{\theta}}, \boldsymbol{\Sigma}/N) \\ f(\boldsymbol{\Sigma} \mid \boldsymbol{\theta}) &\propto \text{Inverse-Wishart}_{N-1}(\mathbf{S}^{-1}) \end{aligned}$$

where $\bar{\boldsymbol{\theta}} = \frac{1}{N} \sum_{i=1}^N \boldsymbol{\theta}_i$ is the mean ability vector and $\mathbf{S} = \sum_{i=1}^N (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T$ the sums of squares matrix around the mean ability vector. The full-conditional distributions $f(\boldsymbol{\theta}_i \mid \underline{\mathbf{x}}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are intractable, however, and for this, we use the SVE algorithm.

Instead of sampling from $f(\boldsymbol{\theta}_i \mid \underline{\mathbf{x}}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ directly, we sample pupil abilities in a dimension q given the $Q - 1$ other dimensions, for $q = 1, \dots, Q$. The full-conditional distribution for the ability of a pupil i in a dimension q is proportional to

$$f(\theta_{iq} \mid \underline{\mathbf{x}}_{iq}, \boldsymbol{\theta}^{(q)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \prod_{i=1}^{J_q} \frac{\exp\{x_{ijq}(\theta_{iq} - \delta_{jq})\}}{1 + \exp\{\theta_{iq} - \delta_{jq}\}} \exp\left\{-\frac{(\theta_{iq} - \lambda_{iq})^2}{2\eta_q^2}\right\},$$

where δ_{jq} is the difficulty of the j -th out of J_q items in dimension q , $\boldsymbol{\theta}_i^{(q)}$ is the ability vector of pupil i excluding entry q , and λ_{iq} and η_q^2 are the conditional mean and variance of θ_{iq} given $\boldsymbol{\theta}_i^{(q)}$ in the population model, respectively, given by

$$\begin{aligned} \lambda_{iq} &= \mu_q + \boldsymbol{\sigma}_q^{(q)} \left(\boldsymbol{\Sigma}^{(q,q)}\right)^{-1} \left(\boldsymbol{\theta}_i^{(q)} - \boldsymbol{\mu}^{(q)}\right) \\ \eta_q^2 &= \sigma_{qq} - \boldsymbol{\sigma}_q^{(q)} \left(\boldsymbol{\Sigma}^{(q,q)}\right)^{-1} \left(\boldsymbol{\sigma}_q^{(q)}\right)^T, \end{aligned}$$

where $\sigma_q^{(q)}$ contains the off-diagonal elements of the q -th row in Σ , i.e., $\sigma_2^{(2)} = [\sigma_{21}, \sigma_{23}, \dots, \sigma_{26}]$.

We sample from the full-conditionals $f(\theta_{iq} \mid \mathbf{x}_{iq}, \boldsymbol{\theta}^{(q)}, \boldsymbol{\mu}, \Sigma)$, as follows. First, we compute λ_{iq} for $i = 1, \dots, N$ (note that these depend on the abilities from the remaining $Q - 1$ dimensions). Then, we sample θ_{vq}^* from $\mathcal{N}(\lambda_{vq}, \eta_q^2)$ and use these to generate an item response vector \mathbf{x}_{vq}^* from $P(\mathbf{X}_q \mid \theta_{vq}^*, \boldsymbol{\delta}_q)$, for $v = 1, \dots, N$. Say that we use $f(\theta_{vq} \mid \mathbf{x}_{vq}^*, \boldsymbol{\theta}_v^{(q)}, \boldsymbol{\mu}, \Sigma)$ as proposal for a target i (i need not equal v), then the probability $\pi(\theta'_{iq} \rightarrow \theta_{vq}^*)$ to make a transition of θ'_{iq} to θ_{vq}^* is equal to $\min\{1, \omega(\theta'_{iq} \rightarrow \theta_{vq}^*)\}$, with

$$\ln \omega(\theta'_{iq} \rightarrow \theta_{vq}^*) = (\theta'_{iq} - \theta_{vq}^*)(t(\mathbf{x}_{vq}^*, \lambda_{vq}, \eta_q) - t(\mathbf{x}_{iq}, \lambda_{iq}, \eta_q)),$$

where

$$t(\mathbf{x}_{iq}, \lambda_{iq}, \eta_q) = \sum_{i=1}^{J_q} x_{ijq} + \lambda_{iq}/\eta_q^2.$$

Note that $t(\mathbf{x}_{iq}, \lambda_{iq}, \eta_q)$ combines information from the likelihood with information from the population model.

To match proposals to targets (full-conditionals), it is opportune to use $t(\mathbf{x}_{iq}, \lambda_{iq}, \eta_q)$, since if $t(\mathbf{x}_{vq}^*, \lambda_{vq}, \eta_q)$ is close to $t(\mathbf{x}_{iq}, \lambda_{iq}, \eta_q)$, the acceptance probability tends to be high. In matching the N proposals to the N targets, we start with computing $t(\mathbf{x}_{iq}, \lambda_{iq}, \eta_q)$ for each target and computing $t(\mathbf{x}_{vq}^*, \lambda_{vq}, \eta_q)$ for each proposal. Then, we order the targets using the $t(\mathbf{x}_{iq}, \lambda_{iq}, \eta_q)$, such that the corresponding statistics are ordered from small to large and do the same for the proposals using the $t(\mathbf{x}_{vq}^*, \lambda_{vq}, \eta_q)$. If the Markov chain is stationary, the first proposal is likely to be a good proposal for the first target (since the difference between $t(\mathbf{x}, \lambda, \eta)$ and $t(\mathbf{x}^*, \lambda, \eta)$ will be small), and the same holds for the second, the third, and so on.

We start our analyses by setting $\boldsymbol{\mu}$ equal to $\mathbf{0}$ and Σ equal to the $Q \times Q$ identity matrix. To get reasonable starting values for the latent ability vectors, we performed a single run of the SVE algorithm where we accepted all proposals. We ran the Gibbs sampler for 2000 iterations, which took approximately 80 min (about 2.5 s per iteration). The acceptance rates of the SVE algorithm were high in this example, averaging to 0.98, 1.00, 0.97, 0.99, 0.99, and 1.00 for dimensions 1 to 6, respectively. This means that we sample approximately i.i.d. from the full-conditional distributions of the abilities, and thus, using the SVE algorithm in this example does not introduce additional autocorrelation to the Markov chain.

Despite the observation that we sample the abilities approximately i.i.d. in this example, the amount of autocorrelation in the chain is high. To illustrate, we show the trace plot for three parameters: an ability, a mean, and a variance. Note the wave-like patterns that emerge, which indicate a strong relation between subsequent states in the Markov chain (i.e., high amount of autocorrelation). The reason for this high amount of autocorrelation is due to the high correlations that we obtain between

Table 12.2 Estimated correlations between scales in the 2012 Cito Eindtoets

Dimension	Correlations					
Non-verb spelling	1.00					
Verb spelling	0.93	1.00				
Reading comprehension	0.64	0.71	1.00			
Basic arithmetic	0.60	0.61	0.71	1.00		
Fractions	0.63	0.63	0.71	0.99	1.00	
Geometry	0.61	0.62	0.69	0.97	0.98	1.00

some of the dimensions (see Table 12.2) and the fact that we sampled from each dimension conditional upon the others. The high correlations between dimensions then provide a strong relation between draws in subsequent iterations, inducing a high amount of autocorrelation (Fig. 12.9).

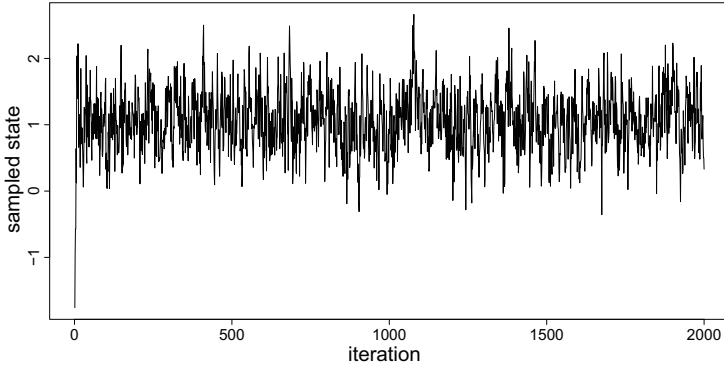
The estimated correlation matrix is shown in Table 12.2. From Table 12.2, it is seen that the two spelling scales are closely related, as are the three mathematics scales. The remaining correlations are only moderately large, yet they are all positively correlated. The correlations in Table 12.2 suggest that there are three distinct dimensions in this problem: spelling, reading comprehension, and mathematics.

12.5 Discussion

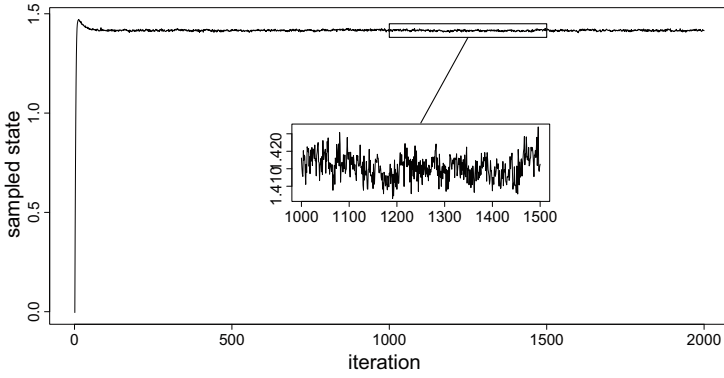
In this chapter, we have described two composition algorithms that can be used to sample from conditional distributions and discussed how their efficiency can be improved to handle large data sets where one needs to sample from many similar distributions.

We have illustrated how the algorithms can be used in a variety of educational measurement applications. We used the composition algorithms for a simulated latent regression example using the random-effects gamma model proposed by Fox (2013), analyzed Amsterdam Chess Test data using the signed residual time model (Maris & van der Maas 2012; Deonovic et al. 2020), and analyzed one big-data example—the Cito Eindtoets—using a multidimensional 2PL model (Reckase 2009). These examples allowed us to illustrate the feasibility of using composition algorithms for simulating from random-effects distributions assessed by complex measurement models. It also allowed us to illustrate that while their efficiency is guaranteed if the algorithms are used in high-dimensional settings (i.e., when there are many instances of a random effect), they are less efficient in low-dimensional settings (e.g., to simulate from the posteriors of the item parameters).

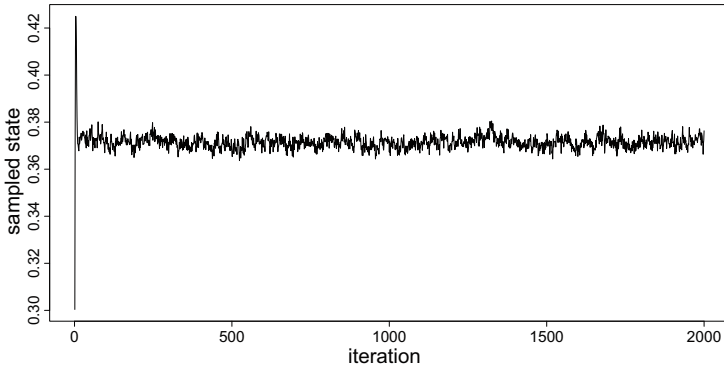
Finally, we note that we used GNU-R to perform the analyses, which was entirely feasible, even for the large applications. Computational time can be decreased by implementing (parts of) the code in a compiled language (e.g., Fortran, C, Delphi). Furthermore, most computer systems run on multiple cores, and computational time



(a)



(b)



(c)

Fig. 12.9 Trace plots of an ability, a mean, and a variance in the Eindtoets example. **(a)** The ability of person $i=59, 137$ in dimension 1. **(b)** The mean of dimension 6. **(c)** The variance of dimension 3

could be decreased further by making use of the additional cores in implementations. For instance, proposals can be generated in batches, with each batch running on a single core.

Appendix A: The Use of Oversampling in the Gamma Example

The Gnu-R (R Core Team 2010) code that was used in the gamma example to sample from the full-conditional distribution of ν is given below.

```
#Compute t(x):
tx = rep(0,N)
for(j in 1:J) tx = tx - X[,j]/(theta * delta[j]) + log
(X[,j])
tx = sum(tx)
#Generate M = 100 proposals:
anu = rgamma(n = M, shape = shape.nu, rate = rate.nu)
#Generate statistics t(x*):
atx = rep(0,j)
for(j in 1:J)
{
  for(m in 1:M)
  {
    tmp = rgamma(n = N,
                 shape = anu[m],
                 rate = anu[m] / (theta * delta[j]))
    atx[m] = atx[m] +
            sum(log(tmp) - tmp / (theta * delta[j]))
  }
}
#Select proposal:
m = which(abs(tx - atx) == min(abs(tx - atx)))[1]
anu = anu[m]
atx = atx[m]
#Calculate log acceptance probability:
ln.omega = (anu - nu) * (tx - atx)
#Metropolis-Hastings step:
if(log(runif(1)) < ln.omega) nu = anu
```

Appendix B: The Use of Matching in the Gamma Example

The Gnu-R (R Core Team 2010) code that was used in the gamma example to sample from the full-conditional distribution of the person parameters is given below.

```
#Generate proposals:
atheta = rlnorm(n = N, #proposals from prior
```

```

        mean = theta.mu,
        sd = theta.sd)
#Compute statistics:
tx = atx = rep(0,n)
for(j in 1:J)
{
    #Compute t(x*):
    atx = atx + rgamma(n = N,
        shape = nu,
        rate = nu / (atheta * delta[j])) / delta
        [j]
    #Compute t(x):
    tx = tx + (X[,j] / delta[j])
}
#Permute proposals:
O = order(order(tx))
o = order(atx)
atheta = atheta[o[O]]
atx = atx[o[O]]
#Calculate the log acceptance probability:
ln.omega = nu * (atx - tx) * (1 / atheta - 1 /
theta)
#Metropolis-Hastings step:
u = log(runif(N))
theta[u < ln.omega] = atheta[u < ln.omega]

```

Appendix C: Sampling Data from the SRT Model

In order to apply the SVE algorithm to sample from the full-conditionals of the person and item parameters, we need to be able to generate data from the model. Since we apply the same procedure for the person as for the item parameters, we only describe the strategy for the person parameters here. We use the factorization $f(\mathbf{X}, \mathbf{S} \mid \theta, \delta, d) = P(\mathbf{X} \mid \theta, \delta, d) f(\mathbf{S} \mid \mathbf{X}, \delta, \theta, d)$ and use composition. Maris and van der Maas (2012) showed that $P(X = x \mid \theta, \delta, d)$ derived from the SRT model is a Rasch model with slope equal to the time limit d and $f(S_{ij} = s_{ij} \mid X_{ij} = x_{ij}, \theta_i, \delta_j, d)$ is

$$f(S_{ij} = s_{ij} \mid X_{ij} = x_{ij}, \delta_j, \theta_i, d) = \frac{(\theta_i - \delta_j) \exp((2x_{ij} - 1)(d - s_{ij})(\theta_i - \delta_j))}{(2x_{ij} - 1) [\exp((2x_{ij} - 1)d(\theta_i - \delta_j)) - 1]}.$$

An interesting feature of this distribution is that the following set of equalities holds (let ϕ denote $\theta - \delta$ in the equalities):

$$(S \mid X = 1, \phi) \underset{st}{=} (d - S \mid X = 0, \phi) \underset{st}{=} (S \mid X = 0, -\phi) \underset{st}{=} (d - S \mid X = 1, -\phi).$$

This indicates that we can introduce a new variable \hat{S} :

$$\hat{S} = \begin{cases} S & \text{if } X = 1 \\ d - S & \text{if } X = 0 \end{cases} \sim (S \mid X = 1, \theta, \delta, d),$$

which Maris and van der Maas (2012) call *pseudo time* and is independent of accuracy ($X \perp\!\!\!\perp \hat{S} \mid \Theta$). Thus, to generate data from the SRT model, we generate X from a Rasch model with slope d , which is a trivial exercise, and to generate S we generate \hat{S} via inversion and solve for S using

$$S = \begin{cases} \hat{S} & \text{if } X = 1 \\ d - \hat{S} & \text{if } X = 0 \end{cases}.$$

That is, draw $u \sim \mathcal{U}(0, 1)$, and set \hat{S}_{ij} equal to

$$\frac{1}{\delta_j - \theta_i} \ln [1 - u(1 - \exp(d(\delta_j - \theta_i)))].$$

References

- Albert, J. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17(3), 251–269. <https://doi.org/10.2307/1165149>.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29–51. <https://doi.org/10.1007/BF02291411>.
- Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society*, 41(1), 1–31. Retrieved from <http://www.jstor.org/stable/2984718>.
- Deonovic, B., Bolsinova, M., Bechger, T., & Maris, G. (2020). A Rasch model and rating system for continuous responses collected in large-scale learning systems. *Frontiers in psychology*, 11. <https://doi.org/10.3389/fpsyg.2020.500039>.
- Fox, J. (2013). Multivariate zero-inflated modeling with latent predictors: Modeling feedback behavior. *Computational Statistics and Data analysis*, 68, 361–374. <https://doi.org/10.1016/j.csda.2013.07.003>.
- Gelman, A., Carlin, B., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). Chapman & Hall/CRC.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741. <https://doi.org/10.1109/TPAMI.1984.4767596>.
- Jiang, Z., & Templin, J. (2018). Constructing Gibbs samplers for Bayesian logistic item response models. *Multivariate Behavioral Research*, 53(1), 132–133. <https://doi.org/10.1080/00273171.2017.1404897>.
- Jiang, Z., & Templin, J. (2019). Gibbs samplers for logistic item response models via the Pólya-Gamma distribution: A computationally efficient data-augmentation strategy. *Psychometrika*, 84(2), 358–374. <https://doi.org/10.1007/s11336-018-9641-x>.

- Klein Entink, R., Fox, J., & van der Linden, W. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74(1), 21–48. <https://doi.org/10.1007/s11336-008-9075-y>.
- Liu, J., Wong, W., & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1), 27–40. <https://doi.org/10.2307/2337047>.
- Maris, G. (2012). Analyses. In N. Jones et al. (Eds.), *First European Survey on Language Competences* (pp. 298–331). European Commission. Retrieved from <https://crell.jrc.ec.europa.eu/?q=article/eslc-database>.
- Maris, G., Bechger, T., Koops, J., & Partchev, I. (n.d.). dexter: Data management and analysis of tests [Computer software manual]. Retrieved from <https://dexter-psychometrics.github.io/dexter/> (R package version 1.1.4).
- Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, 77(4), 615–633. <https://doi.org/10.1007/s11336-012-9288-y>.
- Marsman, M., Maris, G. K. J., Bechger, T. M., & Glas, C. A. W. (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>.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>.
- Mislevy, R. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196. <https://doi.org/10.1007/BF02294457>.
- Mislevy, R., Beaton, A., Kaplan, B., & Sheehan, K. (1993). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133–161. <https://doi.org/10.1111/j.1745-3984.1992.tb00371.x>.
- 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>.
- Murray, I., Ghahramani, Z., & MacKay, D. (2012, August). MCMC for doubly-intractable distributions. *ArXiv e-prints*. Retrieved from <http://arxiv.org/abs/1206.6848>.
- Patz, R., & Junker, B. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2), 146–178. <https://doi.org/10.2307/1165199>.
- R Core Team. (2010). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago, The University of Chicago Press)
- Reckase, M. (2009). *Multidimensional item response theory*. Springer. https://doi.org/10.1007/978-0-387-89976-3_4.
- Rosenthal, J. (2011). Handbook of Markov chain Monte Carlo. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), (p. 93–112). Chapman & Hall. Retrieved from <https://doi.org/10.1201/b10905>.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12(4), 1151–1172. Retrieved from <http://www.jstor.org/stable/2240995>.
- Tanner, M. (1993). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (second ed.). Springer-Verlag.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4), 1701–1762. <https://doi.org/10.1214/aos/1176325750>.
- Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, 8(1), 1–9. <https://doi.org/10.1214/aoap/1027961031>.
- van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287–308. <https://doi.org/10.1007/s11336-006-1478-z>.

- Van der Maas H., & Wagenmakers, E. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology*, *118*(1), 29–60. <https://doi.org/10.2307/30039042>.
- Verhelst, N., & Glas, C. (1995). The one parameter logistic model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 215–238). New York: Springer-Verlag. https://doi.org/10.1007/978-1-4612-4230-7_12.

Chapter 13

To a or not to a : On the Use of the Total Score



Bas T. Hemker

Abstract For the sake of transparency, the use of the unweighted total score is demanded by society in many cases, especially in high-stakes situations such as exams. In the Rasch model, the total score is the sufficient statistic: all relevant information of the measurement is captured by the unweighted sum of the item scores. For this reason, many practitioners want to use the Rasch model. However, in many practical applications, the Rasch model does not fit, and the data is better described by a model that also uses a slope parameter. Although in these types of models, the total score is not the sufficient statistic; the unweighted item sum score can be used to compare candidates' results on different equated tests. In a reevaluation of the true-score equating procedure, we show how the benefits of using the better fitting model can be combined with the application of the total score in the context of equating cut-off scores. The advantages of the total scores are presented, and how the total score can be used also in case the Rasch model does not hold. An example is given to describe how the procedure works in practice. Finally, some reflections are given on the practical implications, meaning, and usefulness of the slope parameter, also known as the a -parameter.

13.1 The Use of the Total Score in Practice

In the behavioral and social sciences, test and questionnaires are often used to measure the position of respondents on a latent trait θ (Hemker et al., 1996, 1997). In this paper we focus on dichotomous items; however some results are extended to polytomous items as well. Let a test consist of J dichotomously scored items. The score on item j is denoted X_j : $X_j = 1$ for responses indicative of the trait, like a correct response to an item measuring ability, and $X_j = 0$ otherwise. In both classical test theory (CTT) and item response theory (IRT), the unweighted sum of J item

B. T. Hemker (✉)
CITO – Research and Development, Arnhem, The Netherlands
e-mail: bas.hemker@cito.nl

scores, also called the unweighted total score or the raw score and denoted by X , is often used for the measurement. In IRT the ordering is used as the most likely order by increasing ability of candidates on a single test. In the case candidates take different but equated tests, the total score can be converted to an equated scale score via an IRT model to position candidates on the latent trait scale.

The reason for the focus on the total score is the ubiquitous use of the total score in real-life situations, often due to societal demands that require transparency. This holds true, especially in the case of summative assessment, where important decisions for an individual candidate are based on the results of a test. These can be entrance tests that in part, or sometimes completely, determine whether a prospect student is allowed to get into a specific course, education, or university. Exams are another example where the decision to pass the course is at least in part dependent on the test result. In those cases the unweighted total score X is a measure for ability that is considered fair and is accepted by the general public.

It is considered to be fair because if there is no additional information given on weights and scoring during the exam, a candidate probably experiences each and every item as equally important and of equal worth. If after the collection of the data it is decided that one item is (much) more important than another item, the candidate cannot decide to put more effort in the items with the higher weights. Also it is expected, and therefore accepted, by the general public that if two candidates that take the same test have the same number of items correct, the pass-fail decision is also the same. It is generally not accepted that, when two candidates make the same test and there is no additional information on item weights, the candidate with the lower number items correct passes the test while the candidate with more items correct fails it. In case weighted scores are used, this very well may happen. Say the pass-fail decision based on the test score is described by $f(X)$, with $f(X) = 1$ if a candidate passes, and $f(X) = 0$ if a candidate fails, the general public expects that, for $0 \leq C < K \leq J$,

$$f(X = C) \leq f(X = K). \quad (13.1)$$

If the number of correct is communicated to the candidate, which is often required, and no information on weights are given beforehand, in practical situations (13.1) needs to hold.

13.1.1 The Unweighted Total Score in Test Theory

In test theory much emphasis is placed on the total scores. In CTT the total score would be the true-score T , on which ideally the actual decision should take place, if it wasn't for that pesky measurement error E , through the relation:

$$X = T + E. \quad (13.2)$$

Because it assumed that $\mu_E = 0$ and $\rho_{TE} = 0$, all deviations of X with respect to T are random, and so generally the decisions on the total score are considered to be as correct as possible.

In CTT (13.2) can be considered a definition, but in IRT, the relation between X and θ follows from assumptions that can be tested. Most IRT models assume unidimensionality and local independence. If in addition for all J items in the test the item response function (IRF) $P(X_j = 1|\theta)$, or $P_j(\theta)$, is nondecreasing in θ , Grayson (1988) and Huynh (1994) have shown that X has monotone likelihood ratio (MLR) in θ . Note that in these proofs, it is also assumed that $0 < P_j(\theta) < 1$. As MLR implies stochastic ordering of θ by X , it means that for $0 \leq C < K \leq J$, and any value s

$$P(\theta > s | X = C) \leq P(\theta > s | X = K). \quad (13.3)$$

This property denotes by (13.3) implies that persons with a higher sum score have on average a higher value of the person parameter than persons with a lower sum score; this property is particularly useful for comparisons between groups of persons. It may not be satisfactory to make ordinal conclusions about individuals without the additional condition of ordinal sufficiency (OS; Zwitser & Maris, 2016). Thus, under these mild assumptions that can be tested, we know that we can use the total score to have a reasonable ordering of (groups of) candidates on the latent trait. The IRT model that is defined by these assumptions that are sufficient for MLR is the Mokken (1971) model of monotone homogeneity (MH model; Mokken & Lewis, 1982; Mokken, 1997).

The CTT and the MH model are useful if we want to compare results on only one test; however in case of educational measurement, often more than one version of the test is required. This can be for multiple reasons. Different versions can be necessary to make cheating more difficult, especially if there is more than one occasion when test takers are allowed to take the test. More than one version can be offered if not all candidates can take the test at the same time, but are also used as a resit. In some cases, different versions with varying degrees of difficulty are offered to cater to specific groups that differ in ability, but still need to be compared on the same scale. The results on these tests need to be comparable over tests, and also the decisions made on the different test need to be fair. To make these tests comparable, in other words to equate the tests, or at least to equate one (pass-fail) or more cut-off scores, parametric IRT models are very helpful.

Here we focus on two types of parametric, logistic IRT models that differ in the number of parameters used to describe each dichotomous item, which is either one parameter or two parameters per item. The first of these two is the one-parameter model (1pl model), or the Rasch (1960, 1968) model. In the 1pl model the IRF is given by

$$P_j(\theta) = \frac{\exp(\theta - \delta_j)}{1 + \exp(\theta - \delta_j)}. \quad (13.4)$$

In this equation the item parameter δ_j describes the location of the item on the latent trait scale where the candidate has a probability of 0.50 to obtain a score of $X_j = 1$. In models where for each item this δ_j is the only parameter, it can be interpreted as the difficulty of the item: independent of the ability of the candidate, for two items j and k with $\delta_j > \delta_k$, the probability to give a correct response is smaller for item j than k .

Note that Eq. (13.4) can also be written as

$$P_j(\theta) = \frac{\exp[\alpha(\theta - \delta_j)]}{1 + \exp[\alpha(\theta - \delta_j)]}, \quad (13.5)$$

with α being a constant that is the same for all items. The value of α does not need to be equal to 1, but the scale can always be transformed in such a way that $\alpha = 1$, which again yields (13.4).

The models with two parameters per item allow the slopes of the IRFs to vary over items, and thus a subscript j is added to α to denote this possible variation over items, which yields α_j . In this study we look at two different models with two item parameters. The most commonly known is the 2pl-model (Birnbaum, 1968; pp. 399–402), in which the IRF is defined by

$$P_j(\theta) = \frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}, \quad (13.6)$$

with α_j being the slope parameter. Here the parameter estimates are allowed to take any number. In an alternative model with a slope parameter, the parameters are limited to integers ranging from 1 to 15. In this model, the slope parameters are imputed, which means that these are determined independently of the estimations of the δ_j -parameters. Confusingly this model was dubbed, the One Parameter Logistic Model (OPLM; Glas & Verhelst, 1989; Verhelst & Glas, 1993), but usually only the abbreviation is used to avoid the mix-up with the 1pl model. To distinguish the OPLM from the 2plm, α_j in (13.6) is given as a_j in this model:

$$P_j(\theta) = \frac{\exp[a_j(\theta - \delta_j)]}{1 + \exp[a_j(\theta - \delta_j)]}. \quad (13.7)$$

The OPLM can be considered a hybrid between the 1pl and 2pl model, because the model allows the slopes to vary, but only δ_j is estimated directly. The variation of a is limited with only 15 possible values in theory. In practice there is even less variation: in most cases only four to seven different values are being used. The advantage of this model compared to the 2pl model is that some fundamental statistical issue with regard to the estimation of parameter estimations are solved in the OPLM (Van den Brink & Mellenbergh, 1998; pp. 215–218). This model has been dominant in the Netherlands for educational measurement in primary

and secondary education for over 25 years. It is also contained in the rules and regulations set by the government (e.g., College voor Toetsing en Examens, 2015).

The value of item parameter δ_j in (13.7) still equals the required ability for a candidate to have a probability of exactly 0.50 to obtain a score of $X_j = 1$. Many psychometricians still refer to this as the “difficulty,” but this often leads to miscommunications with test practitioners. In their review of a test for one population, and comparing the difficulty of two items, they find the item with the lower p-value the more difficult one. It is surprisingly common that this is not the item with the “difficulty parameter” with the highest value. To avoid this confusion, in practice while working with 2pl type of models, it might preferred to refer to δ_j as the “location” of the item.

Both the 1pl model and the 2pl type of models can be considered special cases of the MH model, both for dichotomous and for polytomous item scores (Hemker et al., 1997). Note however that only for the dichotomous models MLR also holds and that for dichotomous models, the stochastic ordering of the candidates can be obtained through X .

13.1.2 Choosing Between the Models with One or Two Parameters

The Rasch model has a specific advantage over all other IRT models, with regard to the relation between the total score and θ , because in the Rasch model, X is the sufficient statistic (Fisher & Russell, 1922) to estimate a persons' θ (e.g. Eggen & Verhelst, 2011). This means that the individual response pattern holds no additional information on the persons' ability θ rather than is given by the unweighted total score: it doesn't matter on what items what scores are obtained. As a result the total scores are easily translated to θ and vice versa. This is a reason why in many applications where tests need to be equated, practitioners prefer the Rasch model. This can be the case of large-scale national or international assessment such as PISA (e.g., OECD, 2000) who used the model until the 2012 cycle, but also in the practice of test publishers who equate test versions.

The property of X as the sufficient statistic is true under the Rasch model. However, if the Rasch model does not hold, X is not the sufficient statistic anymore. Only just estimating the δ_j -parameters, or only deciding beforehand that the total score suffices, it does not mean that the Rasch model holds. The assumptions of the model need to be checked before the nice properties resulting from the model can be applied. Because if you apply the incorrect model, the conclusions you draw from it may be off-target. The model fit needs to be checked (e.g., Molenaar, 1983). This importance of model fit – or at least an investigation of the robustness of the model – was also one of the important issues Kreiner and Christensen (2014) raised in their paper on the use of the Rasch model in PISA. For the 2015 cycle, PISA switched

the 2pl model for dichotomously scored responses and the generalized partial credit model (g-PCM; Muraki, 1992) for polytomous items (OECD, 2017; p. 142).

So instead of applying a non-fitting model, it makes sense to reduce the restrictions and allow variation in the slope of the items, by adding a second parameter, a_j or α_j , for each item. This yields the 2pl model or the OPLM. In most cases the model fit of these models is much better than that of the Rasch model, because of fewer restrictions. If this model holds, we know that the weighted sum score ($\sum a_i X_i$) is the sufficient statistic. However, we can only estimate δ_j -parameters apply the Rasch model and use X as the sufficient statistics. So now, we have two sufficient statistics: one obtained by the model that fits the data and one that doesn't. It is obvious which one should be used to get the maximum information on θ from the data.

If only psychometric criteria would apply, the choice between the models would be easy. However, as was also put forward previously, there are also relevant societal criteria that impose that scale scores have to be explainable to the testees and the public, its computation has to be transparent, and it must be a fair score not interfering with the assessment itself.

If we want to apply scores based on the data, but an important practical question here is: do we inform the candidates in advance about the weight for each item? In case of an exam, or another high-stakes decision, it would be unfair not to inform them because then the candidates cannot anticipate. Without additional information on scoring, for a candidate each item has an equal weight, but in the decision, through scoring, it does not. So applying the statistical weights afterward might be an issue. However, obtaining these statistical weights beforehand in a pretest may not always be the solution. If the weights that are obtained in a pretest are communicated to the candidates as the weights of the items, it may change the test behavior of the candidates compared to that of the pretest candidates who did not have this information. They may spend more time on high weight items and skip the low weight items more often than the pretest candidates and thus altering the type of response patterns. As a result also the a-parameters may change and the weights that were given no longer reflect the model, and the weighted scores with the old weights is no longer the sufficient statistic.

In the end, the unweighted sum score seem to be a good compromise as it fulfills basic measurement desiderata for measuring. It may not always be as efficient (reliable) as weighted sum scores (in case slopes vary), but the loss in efficiency seems negligible from a practical point of view. In most cases the weighted and unweighted scores have a high correlation: the ordering of candidates using raw scores or weighted scores may not be identical – it usually is pretty similar. In practice, it is very rare to find correlations between the weighted and unweighted scores below .95. Usually they are over 0.975. For 2pl type of models with a homogeneous set of slope, it was shown that even ordinal sufficiency may hold (Zwitser & Maris, 2016).

So, we know the a-parameters vary, but we don't want to use weights. We may accept that the stochastic ordering is obtained, but when we have an incomplete design with different test versions, we also want to compare results from different

test versions. In the next section, we advocate a procedure that applies the best fitting model, but still allows to compare the results on the unweighted total score scale.

13.1.3 Using the Total Score to Scale Candidates Without the Rasch Model

The starting point for the approach is the IRF for dichotomous items and its relation to the expected unweighted total score, $E(X|\theta)$. For each value of theta, the expected value X can be determined when you know the IRF for each item, through the following equation:

$$E(X|\theta) = \sum_{j=1}^J P_j(\theta). \quad (13.8)$$

In case we apply the models with a - and b -parameters, like the 2pl model or OPLM, it means that

$$E(X|\theta) = \sum_{j=1}^J \frac{\exp[a_j(\theta - \delta_j)]}{1 + \exp[a_j(\theta - \delta_j)]}. \quad (13.9)$$

Thus, as a result, once the slope and location-parameters are known, for each value of θ , the expected value for X is known. Because each of these IRFs are increasing functions, as long as all slope parameters have a value larger than 0, the function $E(X|\theta)$ is increasing in θ . As a result, for each value of $E(X|\theta)$, there is also only one value of θ that can yield that value. Thus, for each discrete value of (the expected) X , the corresponding θ is known.

Equations (13.8) and (13.9) can easily be extended to polytomous items as well. First, it can be recognized that for dichotomous items $P(X_j=1|\theta) = E(X_j|\theta)$ and (8) can be written as

$$E(X|\theta) = \sum_{j=1}^I E(X_j|\theta). \quad (13.10)$$

For polytomous items a model can be selected, which can be a type of graded response model, a sequential model, or a divide-by-total model (e.g., Hemker et al., 1997). After a type of model is selected, within each type of model, the parametrization can be selected that fits the data the best. For example, within the divide-by-total models, a model with both slope and item-step-parameters (δ_{js}) can be selected, resulting in the g-PCM(with α_j) or the polytomous-OPLM(with a_j). Each items would have only one slope parameter a_i and with item scores ranging from 0 to m on the ($x = 0, \dots, m$) would have m item step parameters δ_{js} , with

$s = 1, \dots, m$. That would yield the following equation:

$$E(X_j|\theta) = \sum_{x=1}^m x P(X_j = x|\theta) = \frac{\sum_{x=1}^m x \sum_{k=1}^x \exp a_j (k\theta - \sum_{s=1}^k \delta_{js})}{1 + \sum_{x=1}^m \sum_{k=1}^x \exp a_j (k\theta - \sum_{s=1}^k \delta_{js})}. \quad (13.11)$$

Applying both Eqs. (13.10) and (13.11) yields

$$E(X|\theta) = \sum_{j=1}^J \frac{\sum_{x=1}^m x \sum_{k=1}^x \exp a_j (k\theta - \sum_{j=1}^k \delta_{js})}{1 + \sum_{x=1}^m \sum_{k=1}^x \exp a_j (k\theta - \sum_{j=1}^k \delta_{js})}, \quad (13.12)$$

which is the polytomous equivalent of Eq. (13.9).

This procedure is an equating method known as true-score equating (e.g., see Kolen & Brennan, 2014, Chap. 6, pp. 176–181; Lord, 1980; pp. 199–202). It has also been generally acknowledged that the true-score equating procedure does not impose any restrictions on the IRT model to be used; that is, it is perfectly fine to use a 2PLM for equating simple sum scores. Although the true-score is a well-known and regularly used procedure, in many applications practitioners seem to think that because they want to use the total score, the equating procedure also needs to be with a Rasch model. In the next section it is discussed how the application of the Rasch model and a OPLM result in different outcomes that may impact the persons who take the test.

13.2 An Example on How to Use the Total Score in Equating Test Versions

The results of true-score equating can differ depending on the model that is applied. In the example we compare the results with equating the same test versions on the same data using the Rasch model and the OPLM. The IRFs are described by Eqs. (13.5) and (13.7), respectively. The first step is to build an item bank based on real data. This step was also taken in the project in which data were collected. In the second step, a reference test with a cut-off score was identified, and this cut-off score was equated to four different test versions. This step is taken here for illustrative purposes: the selection of the items in the reference test and alternative tests and the cut-off score on the reference test are not used in practice

Table 13.1 Test design: distribution of 78 items over 8 test versions

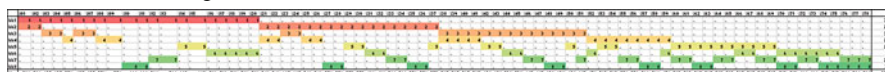


Table 13.2 Characteristics of the distribution of observations (N) per item

	mean	sd	min	Percentiles						max	
				5	10	25	50	75	90		95
N	570	41	479	498	510	546	579	596	621	629	635

13.2.1 Data

The data in this example was collected for a national assessment. In total 78 dichotomously scored items measuring knowledge citizenship in grade 8 in primary education in The Netherlands, which in the Dutch educational system is the year before most students start secondary education. The items were distributed over 8 test versions, with 4 test versions containing 19 items and the other 4 containing 20 items. Each item can be found in two test versions. With 8 test versions, 28 combinations of two test versions can be made. All test combinations have at least two items in common. This is an incomplete design with overlap, where each student is given about 25% of the items in the item bank. In Table 13.1, the design is given for illustrative purposes.

Data was collected from 2275 eight grade students in the regular primary education¹. The number of candidates per test version ranged from 211 to 327, with an average of 284, with more student taking a 20-item test version (304 students on average per version) than a 19-item test version (265 students on average per version). The number of observations per item ranged from 479 to 635, with a median of 579 and a mean number of 570 observations per item. Characteristics of the distribution of observations per item are given in Table 13.2.

13.2.2 Model Fit

Both a Rasch model and an OPLM are used to describe the data. In the OPLM the a-parameters are integers with a minimum value of 1. This means that, with varying a-parameters, the mean a-parameter is by definition larger than 1. We applied an OPLM model where the mean of the a-parameters was arbitrarily set to 3 to identify the scale and to allow for enough variability in the slopes. To make the OPLM and Rasch model more comparable, the a-parameter in the Rasch model as defined in

¹ A specific national assessment on this topic for special education was also performed, but in order not to make the example for complicated, these data are not included here.

Table 13.3 The number of items with a particular a -parameter (in both models total $J = 78$ and mean $a = 3$)

Number of items	Values of the a -parameters					
	1	2	3	4	5	6
Model:						
Rasch	0	0	78	0	0	0
OPLM	9	24	18	13	13	1

Table 13.4 Values of the fit statistic of the Rasch model and the OPLM in relation to the data

Statistic	Likelihood statistics			R1c statistics			Item misfit (% of items)		
	log-likelihood	$-.2*\log$ -likelihood	N pars	R1c	df	p	$p < .01$	$p < .05$	$p < .10$
Rasch	-17574.9	35149.7	77	866.51	515	0.000	17%	26%	31%
OPLM	-14967.2	29934.3	77	581.78	515	0.021	0%	4%	6%

Eq. (13.5) was also set to 3 for all items. Thus, in the Rasch model all items had an a -parameter equal to 3 (Table 13.3).

In the program OPLM that can also be used to analyze the Rasch model, a number of statistics are given to determine model fit (Verhelst & Verstralen, 1994). In Table 13.4 an overview of the values of the fit statistics is given for the two models with regard to the data.






The log-likelihood results are better for the OPLM. Note that in the OPLM, the a -parameters are considered to be imputed and not estimated, which is why the number of estimated parameters is identical. However, if the a -parameters would be considered as estimates, the likelihood is higher, and thus the fit is better for the OPLM. The R1c-statistic shows that the misfit of the OPLM is not significant ($p = 0.021$; $\alpha = 0.01$; one-tailed test $p = .01$). At the level of item fit, we see in the Rasch model that 17% of the items (13 of the 78 items) show a significant misfit at the 1% significance level, while this is the case for none of the items in the OPLM. When we also look at the item misfit at 5% of 10% significance levels, we see that 26% and 31% of the items in the Rasch model show misfit, whereas in the OPLM, this is 4% and 6%, which could be considered as no misfit to the model. The conclusion is that OPLM has a (much) better fit to the data than the Rasch model. Thus, we consider the OPLM as a better description of the item characteristics than the Rasch model.

All items in the item bank can be used to estimate the population distributions. The population distribution shows a mean of 0.397 ($SE = .006$) and a standard deviation of .201 ($SE = .006$) on the Rasch scale. The population distribution on the OPLM has a mean of .406 ($SE = .007$) and a standard deviation of .233 ($SE = .006$).

13.2.3 Test Versions: Reference Test and Alternative Versions

Next, the important question is whether the choice of the model has any impact on the equating results and consequently has any impact for the candidates' pass-fail decisions. Again, note that the selections given next here are only for illustrative

Table 13.5 Characteristics of the reference and alternative test

Test	Number of items	RASCH scale ^a		OPLM scale		Line patterns ^b
		Mean δ_j	Mean δ_j	Mean a_j		
REFERENCE	10	0.10	0.10	3.0		Solid 
b1	10	0.25	-0.02	1.1		Lines & dots 
b2	10	-0.21	-0.01	5.0		Short lines 
b3	10	-0.49	-0.35	3.8		Dotted 
b4	10	0.49	0.57	1.9		Long lines 

^aMean *a* in Rasch scale is always 3

^bThe line patterns colors refer to the line in the Figures in the next section

purposes and do not relate to choices and procedures in the national assessment. For example, the number of items selected for the reference test and the alternative tests are somewhat arbitrary. Somewhat more realistic numbers could have been chosen; however for illustrative purposes, these numbers suffice as the principles behind the example do not change for different numbers of items.

We start with the selection of a set of ten items that form a reference test. These were ten items that have a mean *a*-parameter of 3, also in the OPLM. On this 10-item reference test, the cut-off score that denotes the required ability is set at 6. Usually, a cut-off score is determined by some sort of standard setting procedure (e.g., Cizek & Bunch, 2007). These procedures can considerably vary in complexity.

The four alternative tests versions, b1 through b4, that are selected are also 10-item tests. However, the characteristics of these tests differ over the Rasch and OPLM model. In Table 13.5 the characteristics in terms of parameters from the item bank as a Rasch-scale and as an OPLM scale are given. It is obvious from the mean parameters that the characterization of the alternative tests may vary in case the parameters from the item bank are from the Rasch model or based on the OPLM.

13.2.4 Equating Results

The equating procedure uses the continuous latent trait scale to compare the totals score result. If the cut-off score equals 6 correct on the reference set, we know that the lowest score that is indicative for the required level is 6 items correct out of 10. This also means that the highest score that yields the conclusion that this level is not obtained is 5 out of 10. If no more precise information on the cut-off score is available, for the transfer to the continuous latent trait scale, the actual cut score can be set at 5.5. Thus, we find the value for θ on the scale that yields $E(X_{\text{reference test}}|\theta) = 5.5$ and then relate this point to the expected score of each of the alternative test. We can do this both on the Rasch scale (left side graph in Fig. 13.1) and on the OPLM scale (right side graph in Fig. 13.1). These functions $E(X|\theta)$ can be considered test characteristic curves (TCCs).

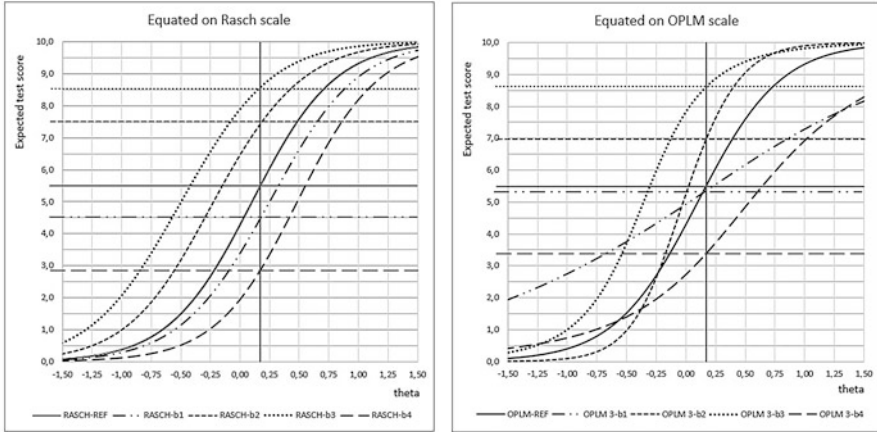


Fig. 13.1 Cut-off score equating on the Rasch scale (left) and OPLM scale (right)

We see that that most TCCs on the left side of Fig. 13.1 are clearly different from those on the right side. Whereas on the left side, the TCCs do not seem to cross², and the tests only seem to differ in difficulty; this is not the case on the right side.

The true-score equating procedure (Kolen & Brennan, 2014) can be shown as follows. We start with the solid horizontal line at the expected score of 5.5 and evaluate where on the latent scale this line crosses the TCC of the reference test. That is where the expected score of the reference test is 5.5. Next, for all the alternative test versions, the expected test score at that position is determined. Now for each test, the expected score is determined that is equated to the expected score of 5.5 on the reference test. In practice the cut-off score is obtained by rounding up this expected score to the next integer. That is the lowest possible number correct that is indicative of the required level on that test.

In practice this is not done by a visual inspection, and these expected values can easily be obtained by applying Eq. (13.7) for the reference test and the alternative test. First the values for θ is found for which $E(X_{\text{reference test}}|\theta) = 5.5$. Next, this value of θ is imputed in Eq. (13.7) for each of the alternative tests. Because the TCCs differ for the Rasch model and the OPLM, the results may differ depending on the model that is fitted to the data. In Table 13.6 the results are given for the four alternative tests.

For three of the four tests in this example, the cut-off score differs depending on the model and thus the scale that is used for equating. In all instances the difference is only one point. If we consider the difference in the percentage of candidates that would pass applying the Rasch model rather than the OPLM, we see that the

² Note that also under the Rasch model, TCCs may cross, depending on the distribution of the item difficulties in the tests. It is a not a property of the Rasch model that the TCCs do not cross. However, in practice it is often found that they do not.

Table 13.6 Equated cut-off scores on the Rasch scale and on the OPLM scale

Scores Test	Equated on Rasch scale		Equated on OPLM scale		Difference
	Exact	Cut-off score	Exact	Cut-off score	Cut-off score
b1 (lines & dots) — . .	4.47	5	5.35	6	1
b2 (short lines) - - - -	7.42	8	6.98	7	-1
b3 (dotted)	8.57	9	8.61	9	0
B4 (long lines) — —	2.85	3	3.38	4	1

difference on average is 4%. This may seem like a relatively small percentage, but is meaningful for the candidates.

In the example the different versions differ in characteristics. This was done for illustrative purposes, but also because in practice different versions can be developed for different groups of testees, for example, to make the test suitable to for different ability levels. This is useful if for one group another ability level is expected than for another group, for example for measurement in special education compared with regular education. Theoretically this should be possible as long as the scale is a unidimensional scale and the model fits the data. The main disadvantage of the short versions is that these are less reliable, and the small difference in cut-off score has a higher impact in terms of difference in pass-fail percentages. If we would equate from the 10-item reference test to the maximum size test, namely, the whole item bank excluding the reference test, we would still have a 3% difference in student passing, again with more passing under the Rasch model. Note, however, that it is not the case that one model necessarily results in stricter cut-off scores than another. One may argue that these small differences could also be found because of measurement error, but note that the differences in equating do not replace measurement error. Measurement error will be an additional nuisance.

Results like these can be relevant to various stakeholders. For example, confronted with two different percentages, policymaker want to know which of the two is the “true percentage”, especially if larger differences are found between the two models. Also after explaining issues like measurement error, they want to know which one to use. Obviously this matters to the testees as well: they only get one result. We would like to argue that this is not an arbitrary choice, and it is not that it is simply the case that models just differ, so one is not better than the other. When we find that one model has a much better fit to the data, we would advise to use that model to equate tests. The predictions made with that model that can be tested, for example, the estimation of CTT item and test characteristics of the eight test versions with the OPLM resembling the observed characteristics better than the estimations with the Rasch model. We would advise to put faith in the model with the best predictions.

13.2.5 Advantage of the Approach

Irrespective of the scales being used, the cut-off scores are only based on (unweighted) total scores. Alternatively, psychometricians often opt for a cut-off score on a theta scale. This can have its advantages, as it is precise and all results from every possible test that are made from the item bank that uses this scale can be “translated” to that scale. However, there are some disadvantages as well. First of all, a value on a theta scale in itself is somewhat meaningless to practitioners, whereas they can understand a number of items correct on a (reference) test. Another disadvantage is that in this procedure, usually the theta scale is fixed, because the cut-off score is defined as a value on this theta scale. Consequently, psychometricians often opt to work with all item parameters as fixed values, even after new data are collected and the parameters can and sometimes should be updated. Updating the parameters is especially relevant when the number of observations exceeds considerably the number of observations on with the original parameters are obtained. Finally, if a 2pl type of model is used to determine the theta scale, usually the person estimates are based on score patterns and items are weighted differently to determine the estimate, while in practice it is often required to use the (unweighted) total number correct responses.

In the true-score equating procedure, these issues can be avoided. It uses the theta scale to compare total scores from one test to another, but no real meaning is given to the actual values on the scale. Any model, also with more than two item parameters, could be used, and it can be updated over time, to give the best possible cut-score in terms of a unweighted total score.

13.3 Tales of Caution on the Application of the a -Parameter

The previous paragraphs expanded on the use of the a -parameter. The fact that we may not need the Rasch model every time we want to use the unweighted total score adds to appeal of the use of the slope parameter and the use of the better fitting model. However, there are a number of reasons to use the slope parameter without giving it too much thought. These cautions are especially relevant for psychometricians and practitioners who want to use the slopes as weights for the weighted score anyway, as it should be a better indication than the unweighted score.

13.3.1 Number of Observations per Item

The advantage of the Rasch model is that it is robust. Also in smaller samples, the estimates for the b -parameters are relatively stable. The COTAN review system for evaluating test quality (Evers et al., 2010) refers to a 1998 paper by Parshall,

Davey, Spray, and Kalohn³ as the basis for a table that gives minimum number of observations to have sufficiently stable results for various models. For the Rasch model, the minimum of 200 was mentioned, while for the two-parameter logistic model, at least 400 observations are necessary⁴. In case the slope parameters are based on lower number of students, the results are not stable, and the better fit to the model is actually a matter of overfitting. With a new sample, the results may change so much that the Rasch results are better. Note that this caution is also relevant for researcher who want to apply true-score equating.

13.3.2 *The Interpretation of the Cause for Varying a -Parameters*

The variation of the slope parameters can simply be used at face value as a characteristics of the IRF: the higher the slope parameter the more information the item has at the location given by the parameter δ_j of that item. However, the evaluation of the estimated slope parameters should also be with regard to the content of the items and the test (e.g., Roelofs et al., 2021).

High values for the slope parameters can, and probably should be found, for items that refer to crucial knowledge rather than items that refer to somewhat less relevant details. Items pertaining to crucial knowledge often yields higher correlations with other items than items relating to details and therefore usually result in higher slope estimates. In this case one may choose to use weighted scores to reflect the importance of these questions. However, it is only fair to the candidates to indicate in the test that this item has a higher weight. Another option might be to delete the items referring to details, if the construct being measured allows it, especially if the estimates of the slope parameters are very low.

Variation in slope parameters may also be a reflection of variation in item quality. Some items may better measure the ability or trait that is measured with the test due to inter- and intra-item writer variability of quality of items. Preferably a review of the items beforehand makes sure that the candidates are not bothered with badly written items, in which it is unclear for the candidates what is actually being asked from them. However, especially inexperienced item writers find it hard to really get it right. Many gruesome examples can be given. For this reason, pretesting is very useful. Badly written items, with (almost) flat slopes for a , should not be in the final test. In some cases rewriting them is possible, but note that pretesting of the new version of the item is highly recommended. As a result, if all items are equally

³ No reference to the actual is given in the COTAN review system and was not found; only a paper from 2002 by these four authors was found online.

⁴ In an unpublished pilot study by Remco Feskens and Bas Hemker in 2020, similar numbers are found for OPLM. The estimated with the Rasch model seemed to be better and more robust in case there are less than 400 observations per item.

relevant, and are of good quality, in the final test the variation of slope parameters should not be very high.

Another reason why slope parameters may vary is because the test pertains to a multidimensional measurement. Note that a variation in slope parameters is not a proof for multidimensionality, neither is the lack of variation a proof for unidimensionality. However, when a large variation in slopes is found, it may reflect that more than one trait is being measured. This could be by design, for example, in a math test where two third of the course work was about algebra and one third was about statistics. It could be wise to simply make two different tests, one for each subject being measured, but often both subjects are in the same test resulting in one test score, especially in case of summative measurement such as exams. In those cases, often the mean a -parameter of the items relating to the dominating topic often is larger than that of the items regarding the other topic, or topics. Then, if weighted scores are used, the dominance of the majority items exceeds the percentages of items allocated the topics: For example, in case of a 30 items test with 20 items (67%) on a majority topic with a mean a -parameter of 3.5 and 10 items (33%) on a minority topic with mean a -parameter of 2 (total mean $a = 3$), the weighted score ranges from 0 to 90. Only $(10 \cdot 2 =) 20$ of these points (22%) then relate to the minority topic, and $(20 \cdot 3.5 =) 70$ points (78%) relate to the majority topic which changes the intended ratio of the content being measured and therefore the intended content validity.

Also it may happen that there is no intended dominating topic, but one type of items that have the same form or refer to the same topic and have a higher inter item correlation than other items. As a results, for these items the mean slope parameter is higher than for other items. Thus, these items start to dominate the score range. A real-life example was once found in pretest of a vocabulary test. Whereas most items related to the meaning of the items, in 20% of the items the respondents were asked to select the opposite meaning. A relatively large part of the students had missed that not a similarity was requested and gave an incorrect response on all these items. This inflated the inter-item correlation on these items. As a result, in the OPLM the mean a -parameter of these items was more than twice as high than the mean a -parameter for the other items. Instead of the 20% of the intended score range, reflected by the number of this type of items, 36% of the weighted scores was related to these items. For this reason, especially if the test contains well-defined topics, or types of items, it is important to evaluate whether the mean slope parameters on one set of items is not much larger than that for the other sets.

A related issue is that local stochastic dependence can be masked as items with high slope parameters. Say, there are two items, and the second can only be responded correctly in case of a correct response to the first, these items are dependent. They will have a relatively high inter-item correlation, and both items will get a relatively high slope parameter. These high slope parameters do not mean that these items measure the topic very well, or reflect the trait the best. However, in case of weighted scores, they will have an huge impact simply because of the interdependence of the items. Thus, if in a calibration two items that are next to each other and that both have very high slope parameters, this may indicate that

the assumption of local stochastic independence has been violated. It is advised to check this before scoring. The solution, if the items turn out to be interdependent, is easy: these items can be scored as one polytomous item. All parameters then should be estimated again. Usually, the slope of the resulting polytomous item is similar to that of the other items.

The main message from this paragraph is that the variability in slope parameters can indicate important issues in the test. The assumption that the slope parameter simply reflects the quality of the item, as items with high *a*-parameters often regarded as good items, may not always be true. A considerable variation in *a*-parameters should trigger a further investigation of the content of the items on whether this optimistic assumption is correct.

13.4 Discussion

As psychometricians who apply IRT, our question is not so much Hamlet's "to be or not to be" (Shakespeare, 1600), but "to *a* or not to *a*." A first question might be: does it matter? As was mentioned, for estimating candidates' abilities, the impact seems relatively mild. Then, if only the ordering of the candidates is important, it could be argued that it hardly matters, because usually high correlations are found between the weighted and unweighted scores. A very important reason to stick to the unweighted sum score is the transparency to the testees and other societal requirements.

If these requirements do not apply, the psychometric perspective optimal weighted score can be used. However, a number of cautions are given, because the weights based on the slope parameters may not always reflect the relevance of the item and may change the validity of the measurement. A wide variation of values of the slope parameters may hide model violations, such as a lack of unidimensionality, or the lack of local independence. Note that varying slopes are definitely not proofs of such violations, but it may be wise to return to the actual items to see why some extreme *a*-parameter estimates occur. In the end psychometrics should always go hand in hand with the expert view of the persons who constructed the items.

In equating tests, where not only the ordering of candidates is important but also whether a specific cut-off point is reached, it can matter what model is used to equate. Whether these differences are considerable or not can be debated. However, with true-score equating, the optimal, best fitting model can be used while still reporting on the basis of the unweighted total score. It will get better results, because in most cases the models with slope parameters give a better description of the data. Therefore, the translation of the cut-off score over test versions is also better when the right model is applied. The need to use the total score does not imply that the Rasch model is necessary in this equating procedure.

An advantage of the true-score equating is the generality of the procedure. In this paper we focused on the Rasch model and the 2plm type of models, but it could

easily be extended to other IRT models as well that throw more item parameters in the mix. This extends to polytomous models as well, so equating of test with other response formats (partial scoring, Likert-scales) is also possible. The caution with regard to the necessary number of observations per item in case a slope parameter is used extends to the addition of more parameters as well: the number of observations necessary to get stable estimate of each parameter increases with the number of parameters per item.

Note that in the example we equated only one cut-off score. The procedure can also be applied for multiple cut-off points, such as a cut-off for the required fundamental level of ability and for a more ambitious target level. In the extreme case, each item score on the reference test can be considered as a cut-off point. In our example that would yield 10 cut-off points. However, some additional challenges occur in those cases that are worthy of their own study.

A final remark is that if you read between the lines, this study also advocates the communication of results in terms of observed, unweighted total scores, and other easy-to-grasp statistics. We can use IRT as a tool, but societal demands require that we need as transparent as possible, or as least have explainable results. To communicate in terms of ability scales with values that in itself have no real meaning, and can be transformed into all kinds of different scales as well, may be confusing, especially, in educational measurement where over time the dimensionality of the scale changes due to changes in the curriculum. The translation to make what we do statistically understandable for, in our case teachers, students and their parents, policy makers, and the general public, is of utmost importance. The use of easy-to-understand statistics is to explain our more difficult models is key here.

References

- Birnbaum, A. (1968). Some latent class models and their use in inferring examinee ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 396–479). Addison-Wesley. <https://ci.nii.ac.jp/naid/10011544105/>
- Cizek, G. J., & Bunch, M. B. (2007). The nedelsky method. In G. J. Cizek & M. B. Bunch (Eds.), *Standard setting* (pp. 68–74). SAGE. <https://doi.org/10.4135/9781412985918>
- College voor Toetsing en Examen. (2015). *Regeling omzetting scores in cijfers bij centrale examinering mbo* [Rules for transforming scores into grades for the central exams in vocational school]. CvTE-15.01457. <https://wetten.overheid.nl/BWBR0036876/2017-08-01>
- EGgen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicologica*, 32(1), 107–132. <https://eric.ed.gov/?id=EJ925442>
- Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results. *International Journal of Testing*, 10, 295–317. <https://doi.org/10.1080/15305058.2010.518325>
- Evers, A., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2015). *COTAN review system for evaluating test quality* (COTAN review system for evaluating test quality) (p. 41). NIP, Utrecht. <https://www.psynip.nl/wp-content/uploads/2019/05/NIP-Brochure-Cotan-2018-correctie-1.pdf>
- Fisher, R. A., & Russell, E. J. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers*

- of a Mathematical or Physical Character, 222(594–604), 309–368. <https://doi.org/10.1098/rsta.1922.0009>
- Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika*, 54(4), 635–659. <https://doi.org/10.1007/BF02296401>
- Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, 53(3), 383–392. <https://doi.org/10.1007/BF02294219>
- Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, 61(4), 679–693. <https://doi.org/10.1007/BF02294042>
- Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, 62(3), 331–347. <https://doi.org/10.1007/BF02294555>
- Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika*, 59(1), 77–79. <https://doi.org/10.1007/BF02294266>
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer. <https://doi.org/10.1007/978-1-4939-0317-7>
- Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, 79(2), 210–231. <https://doi.org/10.1007/s11336-013-9347-z>
- Lord, F. M. (1980). Applications of item-response theory to practical testing problems. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague/De Gruyter.
- Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). Springer. https://doi.org/10.1007/978-1-4757-2691-6_20
- Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, 6(4), 417–430. <https://doi.org/10.1177/014662168200600404>
- Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika*, 48(1), 49–72. <https://doi.org/10.1007/BF02314676>
- 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>
- Organisation for Economic Co-operation and Development. (2000). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy*. OECD Publishing. <https://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-9-Scaling-PISA-Data.pdf>
- Organisation for Economic Co-operation and Development. (2017). *PISA 2015 technical report*. OECD. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut.
- Rasch, G. (1968). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Reading in mathematical social science* (pp. 89–108). Science Research Associates. <https://www.rasch.org/memo19662.pdf>
- Roelofs, E. C., Emons, W. H. M., & Verschoor, A. J. (2021). Exploring task features that predict psychometric quality of test items: The case for the Dutch driving theory exam. *International Journal of Testing*, 21(2), 80–104. <https://doi.org/10.1080/15305058.2021.1916506>
- Shakespeare, W. (1600). *Hamlet*. England.
- Van den Brink, W. P., & Mellenbergh, G. J. (1998). *Testleer en testconstructie* [Test theory and test construction]. Boom Koninklijke Uitgevers.
- Verhelst, N. D., & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika*, 58(3), 395–415. <https://doi.org/10.1007/BF02294648>
- Verhelst, N. D. & Verstralen, H. H. F. M. (1994). *The one parameter logistic model: Computer program and manual*. Universiteit Twente, CITO. <https://research.utwente.nl/en/publications/>

[the-one-parameter-logistic-model%2D%2Dcomputer-program-and-manual\(5966c6a4-dd6d-4cac-8a38-b2527c8de44c\).html](https://doi.org/10.1007/s11336-015-9481-x)

Zwitser, R. J., & Maris, G. (2016). Ordering individuals with sum scores: The introduction of the nonparametric Rasch model. *Psychometrika*, *81*(1), 39–59. <https://doi.org/10.1007/s11336-015-9481-x>

Chapter 14

A Bayesian Test for the Association of Binary Response Distributions



Rudy Ligtoet

Abstract Item response theory models impose testable restrictions on the observed distribution of the response variables. In this chapter, the inequality restrictions are investigated imposed by Mokken’s model of monotone homogeneity (MH) for binary item response variables. A Bayesian test for the observable property of variables being associated is proposed for the trivariate distributions of all triplets of items. This test applies to a wide range of item response theory models that extend beyond the MH model assumptions.

List of Abbreviations

MH	Mokken’s (1971) model of <i>monotone homogeneity</i> .
SPOD	<i>Strongly positive orthant dependence</i> (Definition 1).
A	Variables being <i>associated</i> (Definition 2).
CI	Assumption of (local or) <i>conditional independence</i> .
UD	Assumption of <i>unidimensionality</i> .
M	Assumption of (latent) <i>monotonicity</i> .
LND	Assumption of <i>local non-negative dependence</i> .
DINA	The <i>deterministic inputs, noisy, AND gate</i> model.

14.1 Introduction

Mokken scale analysis consists of a collection of diagnostics to assess the assumption underlying the model of *monotone homogeneity* (MH; Mokken, 1971; Mokken & Lewis, 1982; Molenaar & Sijtsma, 2000; Sijtsma, 1988; Van der Ark, 2007; see List of Abbreviations section). The MH model allows for ordinal inferences about

R. Ligtoet (✉)
University of Cologne, Köln, Germany
e-mail: rligtoe@uni-koeln.de; <https://sites.google.com/site/rligtv/>

a latent variable based on the observed responses to the item variables that make up a test. For example, for binary item variables, the MH model implies a stochastic ordering on the latent variable by the sum of the item scores (Grayson 1988; Ghurye & Wallace 1959; Huynh 1994; Ünlü 2008). The assumptions that constitute the MH model are useful in applications when ordinal inferences are required and no further (parametric) assumptions are deemed either necessary or appropriate.

The diagnostic statistics most prominently used in Mokken scale analysis are the scalability coefficients, which are calculated from the bivariate distribution of pairs of item variables (Loevinger 1948; Molenaar 1991; Warrens 2008). The MH model requires that these scalability coefficients are all non-negative, whereas any negative coefficient discredits or invalidates the model assumptions. However, non-negative scalability coefficients are not sufficient for the MH model. A more restrictive demand of the model is the requirement of non-negative partial correlations, for all triplets of items (Ellis 2014). The restrictions imposed on the trivariate item distributions by this latter requirement are part of a wider class of properties of multivariate positive dependence. Examples of these properties are *multivariate total positivity* (Ellis 2015; Karlin & Rinott 1980) and *conditional association* (Holland & Rosenbaum 1986; Rosenbaum 1984), which are also implied (necessary but not sufficient) by the MH model.

A set of assumptions, other than those that define the MH model, was proposed by Holland (1981), with the aim of testing a wider range of item response theory model. These assumptions pertain to perfect scores (all zeros or ones) on subsets of item variables and hold, if and only if any subset of item variables satisfies the property of *strongly positive orthant dependence* (SPOD; Joag-Dev, 1983). Because conditional association implies SPOD (Definition 1 below), it follows that SPOD includes the MH model as a special case (Holland & Rosenbaum 1986).

In this chapter, a Bayesian test is proposed for a wide range of models for binary response data. This test is based on the observable property of *variables being associated* (A; Esary et al., 1967; Walkup, 1968), but applied to all trivariate distributions of triplets of item variables (cf. non-negative partial correlations). In the following section, I introduce some properties of latent variable models and the observable properties that are implied by these models. In Sect. 14.3, the restrictions imposed on the data distribution by the testable properties are expressed in terms of inequality constraints on the log-odds ratios related to the multinomial parameters. This allows for a convenient way of expressing the Bayes factor in favor of property A. The analysis of response data using the Bayes factor is illustrated in Sect. 14.4, followed by a short discussion in Sect. 14.5.

14.2 Preliminaries and Notation

Let $\mathbf{X} = (X_1, \dots, X_J)$ denote the random vector containing the J binary item response variables. The (real) latent vector $\boldsymbol{\theta}$ is introduced through the law of total probability, $P(\mathbf{X} = \mathbf{x}) = \int P(\mathbf{X} = \mathbf{x} | \boldsymbol{\theta}) dG(\boldsymbol{\theta})$, with its distribution function

G further left unspecified. To define a model, several conditions are considered. The first of these conditions is that of local or *conditional independence* (CI): X_1, \dots, X_J are conditionally independent, given θ . The latent vector is said to be *unidimensional* (UD), whenever $\theta = \theta$ (scalar-valued). The conditions of CI and UD alone, however, are not enough to impose testable restrictions on the distribution of \mathbf{X} (Suppes & Zanotti 1981). The additional condition of *monotonicity* (M) states that the item response functions $P(X_j = 1|\theta)$ are (element-wise) non-decreasing in θ , for all $j = 1, \dots, J$. Holland and Rosenbaum (1986) referred to a model that assumes CI and M as a *monotone latent variable model*, whereby the additional assumption UD defines Mokken's (1971) MH model for (\mathbf{X}, θ) .

14.2.1 Strongly Positive Orthant Dependence

Holland (1981) proposed a set of conditions, with the purpose of providing a test of a wide range of item response theory models. He showed that these conditions hold, if and only if \mathbf{X} satisfies the observable property of SPOD (Joag-Dev 1983), for any selection of variables from \mathbf{X} .

Definition 1 *The vector \mathbf{X} is said to satisfy the property of SPOD (\mathbf{X} is SPOD), if for any partition $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$, the following three inequalities hold:*

$$P(\mathbf{Y} = \mathbf{1})P(\mathbf{Z} = \mathbf{1}) \leq P(\mathbf{X} = \mathbf{1}), \quad P(\mathbf{Y} = \mathbf{0})P(\mathbf{Z} = \mathbf{0}) \leq P(\mathbf{X} = \mathbf{0}),$$

and $P(\mathbf{Y} = \mathbf{1})P(\mathbf{Z} = \mathbf{0}) \geq P(\mathbf{Y} = \mathbf{1}, \mathbf{Z} = \mathbf{0}).$ (14.1)

For the special case $J = 3$, there are only three distinct (non-empty) partitions of \mathbf{X} to consider, with $Y = X_i$ and $Z = (X_j, X_k)$, for $i = 1, 2, 3$ and $j, k \neq i$. This is because, for $J = 3$, the first two inequalities in Definition 1 imply (if and only if) the last inequality for both $Y = X_i$ and $Y = (X_j, X_k)$. For example, for $Y = X_1$ and $p(\mathbf{x}) = P(\mathbf{X} = \mathbf{x})$, with $\mathbf{u} = (0, 1, 1)$, the first inequality implies that

$$P(X_1 = 1)(p(\mathbf{u}) + p(\mathbf{1})) \leq p(\mathbf{1}) \Leftrightarrow p(\mathbf{u})/p(\mathbf{1}) \leq 1/P(X_1 = 1) - 1$$

$$\Leftrightarrow p(\mathbf{1})/p(\mathbf{u}) \geq 1/P(X_1 = 0) - 1 \Leftrightarrow P(X_1 = 0)(p(\mathbf{1}) + p(\mathbf{u})) \geq p(\mathbf{u}),$$

(14.2)

where the last expression corresponds to the third inequality in Definition 1, for $Y = (X_2, X_3)$.

For the general case (any J), the condition of *local non-negative dependence* (LND) is obtained from Definition 1, by conditioning each term on θ . The following result by Holland (1981, Theorem 2) shows that SPOD provides a characterization of a wide class of latent variable models for binary response variables.

Theorem 1 (Holland, 1981) *The binary random vector \mathbf{X} is SPOD, if and only if UD and LND hold, and for any partition $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$, both*

$$\begin{aligned} P(\mathbf{Y} = \mathbf{1}|\theta) &\text{ is non-decreasing in } \theta \text{ and} \\ P(\mathbf{Y} = \mathbf{0}|\theta) &\text{ is non-increasing in } \theta. \end{aligned} \tag{14.3}$$

The set of conditions listed in Theorem 1 contain the MH model as a special case (Holland & Rosenbaum 1986), where Holland (1981) referred to Eq. 14.3 as *monotonicity of the subtest characteristic curves*.

14.2.2 Variables Being Associated

Holland (1981) generalized the conditions that define the MH model (i.e., UD, CI, and M) by relaxing the CI condition and replacing M by Eq. 14.3. Alternatively, one may replace the UD restriction by less restrictive constraints on the multidimensional vector θ , to obtain a multivariate version of the MH model for (\mathbf{X}, θ) . Here, one such relaxation is considered, namely, that θ is A.

Definition 2 (Esary et al., 1967) *The random vector θ is said to be associated with (θ is A), whenever the covariance between $\phi(\theta)$ and $\varphi(\theta)$ is non-negative, for any (element-wise) non-decreasing functions ϕ and φ , for which the involved expected values are defined.*

If θ is A, then any selection of variables from θ is also A, which follows by taking ϕ and φ to pertain only to the selected variables. Assuming CI and M, the following testable result is obtained (Holland & Rosenbaum, 1986, Theorem 8, referring for the proof to Jogdeo, 1978).

Theorem 2 (Jogdeo, 1978) *If θ is A, CI and M imply that \mathbf{X} is also A.*

Proof Together, M and CI imply that $E[\phi(\mathbf{x})|\theta]$ is non-decreasing in θ for any non-decreasing function ϕ (e.g., Holland & Rosenbaum, 1986, Lemma 2). Also, $\mathbf{X}|\theta$ is A, because of CI (Esary et al. 1967, Theorem 2.1). Then, by the Theorem in Jogdeo (1978, p. 234), (\mathbf{X}, θ) is A, and \mathbf{X} is also A, because any subset of associated random variables satisfies A. \square

The MH model is a special case of the conditions in Theorem 2, which in turn are a special case of the conditions in Theorem 1, as property A implies property SPOD (e.g., Holland & Rosenbaum, 1986, p. 1536).

Another example of a model that satisfies the conditions in Theorem 2 can be obtained by considering $\alpha = (\alpha_1, \dots, \alpha_K)$ to be a binary random vector of latent attributes. The DINA model (Doignon & Falmagne 2012; Tatsuoka 1995) is a response model for cognitive assessment, with a successful outcome expected on an item, if all the relevant attributes are possessed. The relevance of the attributes for item j is determined by the binary vector (q_{j1}, \dots, q_{jK}) , which is usually fixed in advance for all items. For the response functions, let $P(X_j = 1|\alpha) = P(X_j =$

$1|\xi_j = 0)P(X_j = 1|\xi_j = 1)$, with $\ln(\xi_j) = q_{j1} \ln(\alpha_1) + \dots + q_{jK} \ln(\alpha_K)$. Then, the DINA model implies condition M, if and only if $P(X_j = 1|\xi_j = 0) \leq P(X_j = 1|\xi_j = 1)$ (Junker & Sijtsma 2001).

Proposition 1 *Assuming CI and M, the DINA model implies that \mathbf{X} is A, if (α, η) satisfies the MH model, with η denoting a second-order latent variable.*

Proof The MH model for (α, η) implies that α is A, so that (\mathbf{X}, α) satisfies the conditions of Theorem 2 (replacing θ by α). □

The second-order latent variable in Proposition 1 can be thought of the cognitive growth that stimulates the development of the attributes, with η positively related to the total number of attributes under the MH model. The purpose of Proposition 1 is, however, not to propose another cognitive diagnostic model, but rather to illustrate the generality of the conditions in Theorem 2.

14.3 Restrictions on the Log-Odds Ratios

Both the property of SPOD and A impose a number of inequality restrictions on the distribution of \mathbf{X} . In order to test these restrictions, it is convenient to denote by \mathbf{p} the vector containing the elements $p(\mathbf{x}) = P(\mathbf{X} = \mathbf{x})$, arranged in lexicographical order of \mathbf{x} (with elements on the right running faster from 0 to 1). Also assume that $\mathbf{p} > \mathbf{0}$. Then, the restrictions imposed by either of the properties can be concisely expressed in terms of inequality restrictions on the log-odds ratios, as

$$\mathbf{K} \ln(\mathbf{M}\mathbf{p}) \geq \mathbf{0}, \tag{14.4}$$

(cf. Bartolucci & Forcina, 2005) with $\mathbf{K} = \mathbf{I}_v \otimes (1, -1, -1, 1)$ and \mathbf{I}_v is the identity matrix of dimensions v equal to the number of restriction. The matrix \mathbf{M} is a binary design matrix which can be adapted to pertain to the restrictions of either A or SPOD. For example, for $J = 2$, take $v = 1$ and $\mathbf{M} = \mathbf{I}_4$, so that Eq. 14.4 yields $\ln p(0, 0) - \ln p(0, 1) - \ln p(1, 0) + \ln p(1, 1) \geq 0$, which corresponds to $\text{Cov}(X_1, X_2) \geq 0$.

Walkup (1968) listed the set of all pairs of binary non-decreasing functions that characterize property A, for up to four items. For $J = 3$, there are nine such pairs of function. One example of such a pair corresponds to $\text{Cov}(X_2, X_3) \geq 0$. It can be verified that this restriction is obtained from Eq. 14.4 using $\mathbf{M} = (1, 1) \otimes \mathbf{I}_4$. Let p_k denote the k th element of \mathbf{p} , with $p_1 = p(\mathbf{0})$, $p_2 = p(0, 0, 1), \dots, p_8 = p(\mathbf{1})$. Another example imposes the restriction $\text{Cov}(1 - (1 - X_1)(1 - X_2), X_3) \geq 0$, which corresponds to the restriction $\ln p_1 - \ln(p_3 + p_5 + p_7) - \ln p_2 + \ln(p_4 + p_6 + p_8) \geq 0$, obtained from Eq. 14.4, using “ \otimes ” for the Kronecker product, as $\mathbf{M} = (\mathbf{I}_2 \otimes (1, 0)', \mathbf{I}_2 \otimes (0, 1)', (1, 1) \otimes \mathbf{I}_2 \otimes (0, 1)')$. By going though all $v = 9$ pairs of functions listed by Walkup (1968), and stacking on top of one another all the corresponding design matrices, we find that property A holds for $J = 3$, if Eq. 14.4

holds, with design matrix

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \\ \mathbf{M}_3 \\ \mathbf{M}_4 \\ \mathbf{M}_5 \\ \mathbf{M}_6 \\ \mathbf{M}_7 \\ \mathbf{M}_8 \\ \mathbf{M}_9 \end{bmatrix} = \begin{bmatrix} (1, 1) \otimes \mathbf{I}_4 \\ \mathbf{I}_2 \otimes (1, 1) \otimes \mathbf{I}_2 \\ \mathbf{I}_4 \otimes (1, 1) \\ \mathbf{I}_2 \otimes ((1, 0)' \otimes (1, 1)) \otimes \mathbf{I}_2 \\ \mathbf{I}_2 \otimes (\mathbf{I}_2, (0, 1)' \otimes (1, 1)) \\ (\mathbf{I}_2 \otimes (1, 0)' \otimes (1, 1), \mathbf{I}_4) \\ (\mathbf{I}_4, \mathbf{I}_2 \otimes (0, 1)' \otimes (1, 1)) \\ ((1, 1) \otimes \mathbf{I}_2 \otimes (1, 0)', \mathbf{I}_2 \otimes (1, 0)', \mathbf{I}_2 \otimes (0, 1)') \\ (\mathbf{I}_2 \otimes (1, 0)', \mathbf{I}_2 \otimes (0, 1)', (1, 1) \otimes \mathbf{I}_2 \otimes (0, 1)') \end{bmatrix}. \tag{14.5}$$

The matrix \mathbf{M} in Eq. 14.5 consists of $v = 9$ stacked matrices $\mathbf{M}_1, \dots, \mathbf{M}_9$, each of dimensions 4×8 .

The following result shows that SPOD and A coincide in case $J = 3$.

Theorem 3 For $J = 3$ binary variables and $\mathbf{p} > \mathbf{0}$, property A is satisfied if and only if property SPOD is satisfied for all subsets of variables.

Proof For any subset of two variables from $\mathbf{X} = (X_1, X_2, X_3)$, SPOD implies that the covariance between the two variables are non-negative. This corresponds to $\mathbf{M}_1, \mathbf{M}_2$, and \mathbf{M}_3 in Eq. 14.5 for the three distinct subsets (X_2, X_3) , (X_1, X_3) , and (X_1, X_2) , respectively. The remainder of the proof consists of going through the process of exhaustively listing all restrictions imposed by SPOD for $J = 3$ and expressing these in terms of the log-odds ratios. It can then be shown that $\mathbf{M}_4, \dots, \mathbf{M}_9$ of the design matrix \mathbf{M} in Eq. 14.5 match one-to-one with those obtained for property SPOD. As an example, consider the inequality $P(\mathbf{Y} = \mathbf{1})P(\mathbf{Z} = \mathbf{0}) \geq P(\mathbf{Y} = \mathbf{1}, \mathbf{Z} = \mathbf{0})$ from Definition 1, which reduces for $\mathbf{Y} = (X_1, X_2)$ and $\mathbf{Z} = X_3$ to $(p_7 + p_8)(p_1 + p_3 + p_5 + p_7) \geq p_7$ and yields $\ln p_8 - \ln(p_2 + p_4 + p_6) - \ln p_7 + \ln(p_1 + p_3 + p_5) \geq 0$. The last inequality is obtained from Eq. 14.4 using \mathbf{M}_8 in Eq. 14.5. The remaining five inequalities can be obtained similarly. \square

One problem when testing either the properties A or SPOD using Eq. 14.4 is that the number of constraints grows fast as J increases to a more realistic size. For example, for $J = 4$, Walkup (1968) listed 99 restrictions imposed by property A. Furthermore, many of the restrictions pertain to outcomes for which observations may be sparse as these restrictions involve ever higher-order interactions between the variables in \mathbf{X} . A solution to both these problems is to consider testing the property A for all triplets of item response variables from \mathbf{X} only. By considering the trivariate distributions, the hope is to have a test that is more powerful than a test that involves only the bivariate distribution while at the same time being broad enough to target a wide range of response models.

14.3.1 Trivariate Associated Distributions

Considering all the triplets of item variables of a test of length $J \geq 3$. With property A imposing $J(J-1)/2$ inequality restrictions on the bivariate distributions, and six restrictions involving three items for each trivariate distribution, the total number of restrictions $v = J(J-1)(J-3/2)$.

The design matrix \mathbf{M} for assessing the v inequality restrictions is obtained as follows. First, let

$$\mathbf{B}_{jk} = \mathbf{B}_{jk1} \otimes \cdots \otimes \mathbf{B}_{jkJ}, \text{ with } \mathbf{B}_{jkl} = \begin{cases} \mathbf{I}_2 & \text{if either } j = l \text{ or } k = l \\ (1, 1) & \text{otherwise,} \end{cases} \quad (14.6)$$

and let the matrix \mathbf{B} be obtained by stacking on top of one another all matrices \mathbf{B}_{ij} , which contains all the restrictions imposed on the bivariate distributions. Second, for the trivariate distributions, let \mathbf{R} be the 8 by 3 matrix with in its rows all binary vectors of length 3 , in lexicographical order. Likewise, let \mathbf{S} be the 2^J by J matrix with in its rows all binary vectors of length J , and let \mathbf{T} denote the matrix in Eq. 14.5, but without \mathbf{M}_1 , \mathbf{M}_2 , and \mathbf{M}_3 . Third, let \mathbf{C}_{jkl} be a matrix of dimensions 24 by 2^J . Matrix \mathbf{C}_{jkl} is assigned to its a th column the same values as \mathbf{T} has in its b th column, whenever $(s_{aj}, s_{ak}, s_{al}) = (r_{b1}, r_{b2}, r_{b3})$, for $1 \leq a \leq 2^J$ and $1 \leq b \leq 8$. Finally, matrix \mathbf{C} is obtained by stacking all matrices \mathbf{C}_{jkl} on top of each other.

The goal is to test the hypothesis H of trivariate A, for all triplets of response variables, with H obtained as the set of vectors \mathbf{p} , which satisfy $\mathbf{p} > \mathbf{0}$ and $\mathbf{1}'\mathbf{p} = 1$ (multinomial model), and Eq. 14.4, with the matrix \mathbf{M} obtained from stacking matrix \mathbf{B} on top of \mathbf{C} . A maximum likelihood procedure for testing inequality restrictions requires the estimation of \mathbf{p} and produces test statistics which asymptotic sampling distributions are difficult to obtain (e.g., Bartolucci & Forcina, 2000, 2005; Vermunt, 1999). Here, a Bayesian approach is considered instead, which requires a prior density to be assigned to \mathbf{p} , but has the advantage that it allows for hypothesis H to be tested against its complement of at least one violation of Eq. 14.4. As a prior for \mathbf{p} , a flat (uniform) Dirichlet distribution is chosen, where the influence of this particular choice is expected to be small as long as all observations of $\mathbf{X} = \mathbf{x}$ have enough support.

14.3.2 Bayes Factor for Trivariate Associated Distribution

The Bayes factor in support of hypothesis H is expressed in terms of the prior and posterior probabilities that the restrictions imposed by H are satisfied (Klugkist & Hoijtink 2007; Tilmstra et al. 2015). The prior probability of H is estimated by sampling a large number of vectors \mathbf{p} from the flat Dirichlet distribution and calculating the proportion c that satisfies Eq. 14.4. Let \mathbf{n} denote the vector containing the observed frequencies of $\mathbf{X} = \mathbf{x}$, arranged as \mathbf{p} . Also, let d denote the proportion

of samples that satisfy Eq. 14.4, with the samples obtained from a Dirichlet distribution, with the hyper-parameters equal to $\mathbf{n} + \mathbf{1}$. The ratio d/c provides an estimate of the Bayes factor in favor of H over the multinomial model. The Bayes factor for the evidence in favor of H over its complement (at least one violation) then becomes

$$L = (1/c - 1)/(1/d - 1), \quad (14.7)$$

where a value $L > 1$ expresses support in favor of H , whereas a value $L < 1$ expresses support for the hypothesis that there is at least one violation of H (Lavine & Schervish 1999; Kass & Raftery 1995).

The sampling procedure for obtaining the proportions c and d can be made more efficient by “activating” the restrictions one by one. Let c_k denote the conditional proportion of samples that satisfy the k th restrictions, given that all previous $k - 1$ restrictions are satisfied. Then, $c = c_1 \cdots c_v$ and similarly for d (Mulder et al. 2009; Tijmstra & Bolsinova 2019), where a sample of vectors \mathbf{p} under the first $k - 1$ restrictions can be obtained using a Gibbs sampler similar to Hoijsink and Molenaar (1997; Ligtoet & Vermunt, 2012).

For illustration of the Gibbs sampler, let $J = 3$, and consider sampling p_3 from the prior distribution constrained by Eq. 14.4 using for \mathbf{M} in Eq. 14.5 only \mathbf{M}_1 , \mathbf{M}_2 , and \mathbf{M}_3 . Let

$$a = \min((\tilde{p}_1 + \tilde{p}_5)(\tilde{p}_4 + \tilde{p}_8)/(\tilde{p}_2 + \tilde{p}_6) - \tilde{p}_7, (\tilde{p}_1 + \tilde{p}_2)(\tilde{p}_7 + \tilde{p}_8)/(\tilde{p}_5 + \tilde{p}_6) - \tilde{p}_4) \\ \text{and } b = (\tilde{p}_2 + \tilde{p}_4)(\tilde{p}_5 + \tilde{p}_7)/(\tilde{p}_6 + \tilde{p}_8) - \tilde{p}_1, \quad (14.8)$$

with \tilde{p}_k denoting the values sampled at the previous iteration. The newly sampled value \tilde{p}_3 obtained from the gamma distribution truncated between a and b then yields $\tilde{\mathbf{p}}/1'\tilde{\mathbf{p}}$ as a single sample from the prior distribution constrained by the restrictions imposed by matrices \mathbf{M}_1 , \mathbf{M}_2 , and \mathbf{M}_3 . This sampling procedure is repeated (for both the prior and posterior) for each p_k many times over, gradually adding the restrictions by extending \mathbf{M}_1 , \mathbf{M}_2 , \dots , up to \mathbf{M}_v . An R program for implementing this algorithm and calculating the Bayes factor is available from the author’s website.

14.4 Application

As a small application, consider the transitive reasoning data (Verweij et al. 1996, for details), which are available from the **mokken** package in R (Van der Ark 2007). These data consist of the binary responses of $N = 425$ children to transitive reasoning tasks, where we limit the analyses to those tasks that relate to the task property Length ($J = 4$) and Weight ($J = 5$).

For the $J = 4$ items related to the task property Length, hypothesis H imposes $v = 30$ inequality restrictions. Considering only the restrictions on the bivariate distributions, we get a Bayes factor of $L = 1.2936$, indicating no clear evidence in favor or against property A. Including the remaining 24 restrictions imposed by hypothesis H yields $1/L = 4.6555$, which indicates substantial evidence against the variable of the task property Length being A. The result thus discredits any model that is a special case of the general conditions listed in Theorem 2, including the MH model. This result, however, was not obvious when only the information contained in the bivariate distributions was considered.

For the $J = 5$ items related to the task property Weight, $L = 13.1766$, indicating strong evidence in favor of hypothesis H .

14.5 Discussion

The Bayes factor for the hypothesis that the observable property A holds for all trivariate distributions of triplets of item variables (hypothesis H) provides a convenient way of summarizing evidence in favor of the many restrictions the property imposes on the observed binary data distribution. The application illustrated that the restrictions imposed by the property on the trivariate distributions, in addition to the restrictions on the bivariate distributions, cannot generally be ignored. A test of property A for all J item variables becomes practically infeasible due to the large numbers of restrictions. Hence, the proposed test strikes a balance between the power of the test and what is practically feasibility. However, the procedure for computing the Bayes factor is still computationally very intensive and is no longer feasible for more than seven items. The procedure would thus benefit from alternative ways of estimating or approximating the prior and posterior probabilities for the Bayes factor.

One easy way of alleviating the computational burden when assessing trivariate A is to consider calculating the Bayes factors separately for each of the triplets of item variables, rather than combining these same restrictions into a single test. Note that the number of restrictions in both cases is the same, but the Gibbs sampler runs faster many times on smaller problems than on a single run across the entire 2^J multinomial outcome space. However, the challenge then is to combine the $J(J - 1)(J - 2)/6$ Bayes factors to come to a judgment about the validity of the assumptions being tested for subsets of response variables. The use of a single global test has a clear advantage here.

Theorem 2 shows that property A for all trivariate distributions is implied by any model for binary response data that assumes the conditions CI and M to hold and additionally assumes that the random vector of (multidimensional) latent variables satisfies property A. These conditions include those that were proposed by Holland (1981) for trivariate distributions, Mokken (1971)s MH model, and a special multilevel version of the DINA model (Proposition 1). Whereas specific tests can be designed for each of the special instances of the conditions listed in

Theorem 2, the test proposed here is aimed at assessing whether the pursuit of any of such models is worth the effort at all.

References

- Bartolucci, F., & Forcina, A. (2000). A likelihood ratio test for MTP2 within binary variables. *The Annals of Statistics*, 28(4), 1206–1218. <https://www.jstor.org/stable/2673960>.
- Bartolucci, F., & Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika*, 70(1), 31–43. <https://doi.org/10.1007/s11336-001-0934-z>.
- Doignon, J. P., & Falmagne, J. C. (2012). *Knowledge spaces*. Springer Science & Business Media.
- Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika*, 79(2), 303–316. <https://doi.org/10.1007/s11336-013-9341-5>.
- Ellis, J. L. (2015). MTP2 and partial correlations in monotone higher-order factor models. In R. E. Millsap, D. M. Bolt, L. A. Van der Ark, & W. C. Wang (Eds.), *Quantitative psychology research* (pp. 261–272). Springer.
- Esary, J. D., Proschan, F., & Walkup, D. W. (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics*, 38(5), 1466–1474. <https://doi.org/10.1214/aoms/1177698701>.
- Ghurye, S. G., & Wallace, D. L. (1959). A convolutive class of monotone likelihood ratio families. *The Annals of Mathematical Statistics*, 30(4), 1158–1164. <https://doi.org/10.1214/aoms/1177706101>.
- Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, 53(3), 383–392. <https://doi.org/10.1007/BF02294219>.
- Hojitink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, 62(2), 171–189. <https://doi.org/10.1007/BF02295273>.
- Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika*, 46(1), 79–92. <https://doi.org/10.1007/BF02293920>.
- Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, 14(4), 1523–1543. <https://doi.org/10.1214/aos/1176350174>.
- Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika*, 59(1), 77–79. <https://doi.org/10.1007/BF02294266>.
- Joag-Dev, K. (1983). Independence via uncorrelatedness under certain dependence structures. *The Annals of Probability*, 11(4), 1037–1041. <https://doi.org/10.1214/aop/1176993452>.
- Jogdeo, K. (1978). On a probability bound of Marshall and Olkin. *The Annals of Statistics*, 6(1), 232–234. <https://doi.org/10.1214/aos/1176344082>.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272. <https://doi.org/10.1177/01466210122032064>.
- Karlin, S., & Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4), 467–498. [https://doi.org/10.1016/0047-259X\(80\)90065-2](https://doi.org/10.1016/0047-259X(80)90065-2).
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>.
- Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, 51(12), 6367–6379. <https://doi.org/10.1016/j.csda.2007.01.024>.

- Lavine, M., & Schervish, T. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, 53(2), 119–122. <https://doi.org/10.2307/2685729>.
- Ligtvoet, R., & Vermunt, J. K. (2012). Latent class models for testing monotonicity and invariant item ordering for polytomous items. *British Journal of Mathematical and Statistical Psychology*, 65(2), 237–250. <https://doi.org/10.1111/j.2044-8317.2011.02019.x>.
- Loevinger, J. A. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, 45(6), 507–530.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. De Gruyter.
- Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous responses. *Applied Psychological Measurement*, 6(4), 417–430. <https://doi.org/10.1177/014662168200600404>.
- Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden*, 12(37), 97–117.
- Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows*. [Computer software], IEC ProGAMMA.
- Mulder, J., Klugkist, I., Van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, 53(6), 530–546. <https://doi.org/10.1016/j.jmp.2009.09.003>
- Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49(3), 425–435. <https://doi.org/10.1007/BF02306030>.
- Sijtsma, K. (1988). *Contributions to Mokken's nonparametric item response theory*. Free University Press.
- Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, 48(2), 191–199. <https://doi.org/10.1007/BF01063886>.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Erlbaum.
- Tijmstra, J., & Bolsinova, M. (2019). Bayes factors for evaluating latent monotonicity in polytomous item response theory models. *Psychometrika*, 84(3), 846–869. <https://doi.org/10.1007/s11336-019-09661-w>.
- Tijmstra, J., Hoijtink, H., & Sijtsma, K. (2015). Evaluating manifest monotonicity using Bayes factors. *Psychometrika*, 80(4), 880–896. <https://doi.org/10.1007/s11336-015-9475-8>.
- Ünlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 61(1), 179–187. <https://doi.org/10.1348/000711007X173391>.
- Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software* 20(11), 1–19. <https://doi.org/10.18637/jss.v020.i11>.
- Vermunt, J. K. (1999). A general class of nonparametric models for ordinal categorical data. *Sociological Methodology*, 29(1), 187–223. <https://doi.org/10.1111/0081-1750.00064>.
- Verweij, A. C., Sijtsma, K., & Koops, W. (1996). A Mokken scale for transitive reasoning suited for longitudinal research. *International Journal of Behavioral Development*, 19(1), 219–238. <https://doi.org/10.1177/016502549601900115>.
- Walkup, D. W. (1968). Minimal conditions for association of binary variables. *SIAM Journal on Applied Mathematics*, 16(6), 1394–1403. <https://doi.org/10.1137/0116115>.
- Warrens, M. J. (2008). On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, 73(4), 777–789. <https://doi.org/10.1007/s11336-008-9070-3>.

Chapter 15

Efficiency and Effectiveness of Teacher-Informed Targeting Testing from Different Perspectives



J. Hendrik Straat, Renske E. Kuijpers, Kimberley Lek,
and Wilco H. M. Emons

Abstract This chapter explores targeting testing in applications where the main interest is in classifying the test takers into three (or more) ordered proficiency levels. A targeted test consists of several fixed booklets, balanced in content but varying in the overall difficulty. The booklets are assigned to test takers using background information about the ability. Usually it is the teacher who assigns the booklets. Targeting testing can be conceived as a modest form of adaptivity that balances psychometric, substantive, instructional, and practical requirements. Using simulations, we studied the consistency and accuracy of targeted tests for polytomous classifications. A distinction is made between the use of targeting testing for making decisions about individuals and its use for interpreting group-level results. Results are obtained for various number of items per booklet, different booklet compositions, and different optimal or less optimal strategies for assigning booklets to candidates.

15.1 Efficiency and Effectiveness of Teacher-Informed Targeting Testing from Different Perspectives

Educational assessments are oftentimes used for multiple purposes, even if obtained with a fixed set of instruments. The results are of interest to different groups of stakeholders. For example, teachers use the assessments to monitor student learning, to provide feedback to students (Hattie & Timperley, 2007), to communicate about learning achievements with student's parents, or to make informed instructional

J. H. Straat · R. E. Kuijpers (✉) · K. Lek
Stichting Cito – National Institute for Educational Measurement, Arnhem, the Netherlands
e-mail: Hendrik.Straat@cito.nl; Renske.Kuijpers@cito.nl; kimberley.lek@cito.nl

W. H. M. Emons
Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands
e-mail: W.H.M.Emons@tilburguniversity.edu

decisions either at the individual level or at the class or course level (i.e., formative assessment; e.g., Wiliam, 2011). School councils use the results to evaluate their performance and for accountability purposes. Policy-makers use the results of educational tests in evaluation studies on school effectiveness, to monitor nation-level student achievements over the years and to evaluate the impact of specific educational policies.

Different testing purposes generally require different test specifications. If the test is *only* used for pass-fail decisions – like with mastery testing and credentialing (Jodoin et al., 2006) – then the test items must be composed in such a way that pass-fail decisions can be made as reliably as possible. To accomplish this goal, the test has to consist of items that are most informative for ability levels close to the cutoff, which amounts to items having about the same difficulty. However, if the test is used to monitor student achievement over a longer period of time, either at the individual level or the group level, then the test should be informative about the entire ability range. This means that the items should have a greater variation in difficulty. In other words, from a psychometric perspective, what works well for one purpose may be less effective for another.

To achieve the intended educational goals, it is important that the tests meet the basic psychometric requirements, but designing assessment programs involves many other considerations as well. What different testing goals often have in common is that they require assessments across different content domains and, perhaps more importantly, that measurements are taken periodically. Research has suggested that periodic monitoring student achievement is a key driver to successful learning and advancing school effectiveness (e.g, Fuchs et al., 1984; Ysseldyke et al., 2003). It enables teachers to detect individual learning delays at an early stage and intervene timely if there is a reason to do so. It also gives teachers, schools, and policy-makers the necessary tools to make informed decisions, particularly in the event of unforeseen complications such as the school closures during COVID-19 pandemic in 2020. For many weeks, students had to follow the lessons online, and students were dependent on homeschooling. The impact that may have had on student learning created a sudden urgency to use all available test information to study the immediate effects on educational progress and possible learning loss (Engzell et al., 2021; see also Lek et al., 2020).

Extensive periodic testing, however, may have a downside because it may consume valuable teaching time. Therefore, it is imperative that the allocated assessment time is used as efficient as possible. It is well-known that tests are most informative if the item difficulties match the ability level of the test taker (Lord, 1980). Items that are way too easy or way too difficult provide little information about the student's ability and should better be avoided. Adaptive testing (e.g., Weiss, 1982) refers to assessment methodologies that take the ability level of the candidate into account in the selection of the items to be administered. Computerized adaptive assessments (CAT; Lord, 1980; Eggen & Straetmans, 2000; Van der Linden & Glas, 2010; Wainer, 2000; Weiss, 1982) select the items in real time, each item using the information about the ability which is available so far. In multistage testing (Berger et al., 2019; Lord, 1971; Yan et al., 2014), ability-

matching clusters of items are selected sequentially throughout the test. These clusters are equally balanced in content. Research has shown that well-designed CATs may reduce test length considerably (e.g., Van der Linden & Glas, 2010).

It goes without saying that CATs have high potential, but building CATs can be difficult and costly to realize in practice. It requires a well-calibrated item bank, which in turn requires extensive item pretesting in large samples. In addition to the high development costs, there are also educational concerns regarding their feasibility in practical settings. For individual learning and feedback, but also to evaluate learning in relation to the curriculum at aggregated levels, it is important that the tests adequately cover the relevant educational objectives, where each objective is tested at the desired cognitive level. When different test forms would vary too much with respect to the tested subjects – for example, one math test contains more items about addition and another math test contains more items about division – the test results may not be generalizable to the content domain. This need to adequately control the content has led to further technical innovations in CAT such as Van der Linden and Veldkamp's (2004) shadow test approach. It has been also a driving force for new developments in multistage testing (Luecht & Nugester, 1998; Yan et al., 2014; Zenisky et al., 2010).

In addition to concerns about psychometric efficiency and content validity, other contextual considerations play a role as well. These considerations pertain to the testing conditions, transparency, and technical feasibility. First, in linear tests, test takers can easily skip questions and come back to them later. Research has shown that the possibility of review and changing answers had a positive effect on the results (Vispoel, 1998). Second, because CATs rely on IRT-based pattern scoring, the items may receive different weights in a CAT without the student being aware of it, rendering the scoring less transparent. With a fixed linear test, the student can be clearly communicated in advance how many score credits he/she needs to achieve a certain mastery level. Finally, computer-based testing may still not always be practically feasible, even in times where many students have laptops and tablets at their disposal.

A modest and accessible approach to adaptive testing for multiple purposes, including class-room settings, is the so-called targeting testing (e.g., Berger et al., 2019; Eggen & Verhelst, 2011; Mislavy & Wu, 1996; Wainer, 2000). The measurement instrument consists of a collection of linear tests, henceforth referred to as booklets, which cover the same content domain but differ in overall difficulty. Test takers are assigned to booklets using background information related to ability such as grades or class performance. In practice, it is usually the teacher who determines which booklet is to be administered to the student. Targeted testing is particularly interesting in classroom settings. By working with fixed booklets, teachers get easily experienced with the test materials and then can use those experiences to easily identify the learning objectives the student is struggling with (e.g., William & Leahy, 2015). Likewise, the teacher can better evaluate whether the performances are in line with expectations.

However, the effectiveness of targeting testing depends on the composition of the booklets and the accuracy by which booklets are optimally assigned to test takers.

Because booklets are assigned based on auxiliary background information, it cannot be guaranteed that the test takers always receive the booklet that best matches their ability. Therefore, it is important that the booklets in targeted test are generic enough to compensate modest misassignments, but specific enough to realize the benefits of an adaptive test. Finally, the booklets must be designed in such a way that valid conclusions can be reached at multiple levels (student, class, cohort).

This chapter explores targeting testing in applications where the main interest is in classifying the test takers into three (or more) ordered proficiency levels. These levels may pertain to educational achievement (e.g., below basic, basic, proficient) or used as input for making placement decisions (e.g., Berger et al., 2019). In particular, we are interested to what extent targeting testing allows consistent and accurate decisions – either at the individual level or the group level – using subjective booklet assignment and a limited number of items. The overarching goals of this chapter are twofold: first, to provide a better understanding of implementations of targeted testing for multiple purposes. Results of our study may help test publishers to improve their policies towards teachers with respect to the assignment of booklets with an optimal level of difficulty to each test taker. More insights into the extent to which results are sensitive to accurate assignment of the booklets will help in practice to further increase the efficiency of targeting testing. Second, this chapter illustrates a comprehensive framework to study the consistency and accuracy of tests for classification problems both at the individual level and group level. In particular, when using tests for individual decision-making, one should not be fooled by general measures of reliability because low reliability does not necessary disqualify a test for individual decision-making (see also Sijtsma, 2009, for a critical discussion).

15.2 Perspectives on Classification Consistency and Accuracy

Consistency and accuracy are two important indicators to gauge the psychometric quality of tests for classifying students into categories (e.g., Cheng & Morgan, 2013; Kim et al., 2006, Livingston & Lewis, 1995). Classification *consistency* refers to the agreement between observed classifications across *two* independent replications. The accuracy is the level of agreement between the classifications based on the observed scores in a *single* administration and those that would be obtained based on true (errorless) scores. Accuracy thus refers to test's ability to assign the candidate to the mastery level that adequately reflects his or her knowledge, that is, the extent to which the test yields the *correct* inferences about the person's mastery level. Note that in case of mastery testing, there is usually no gold standard, but the mastery criteria are defined by thresholds on the sum-score scale, as determined by experts (Cizek & Bunch, 2007). The experts basically determine what minimum true score is needed for each proficiency level. This means that if the test takers are classified by their true scores, the classifications will be correct by construction. As a result,

the consistency provides a lower bound for the accuracy, and both consistency and accuracy go to one as the reliability goes to perfect.

Furthermore, it is important to emphasize that common indices of consistency and accuracy (Cheng & Morgan, 2013; Livingston & Lewis, 1995) reflect how the test functions at the group level, but they do not provide a detailed picture of the functioning of the test at the individual level. To illustrate, consider as a simple measure of consistency the percentage of test takers consistently classified into the same category across two replications. Let say for the moment that there is 80% agreement. This means that in every (hypothetical) retest, about 80% of the test takers will be classified into the same category as they were before. However, in each replication, the composition of the group of test takers that is consistently classified may change. Hence, an overall percentage of 80% correct classifications does not imply 80% certainty of a consistent classification for each individual. More importantly, the probability of an individual student being consistently classified depends on the ability relative to the cutoffs. To decide whether a test is suitable for making individual decisions, we need to look at classification decisions as the individual level in addition to group-level classification consistencies.

15.2.1 Individual-Level Classification Certainty

It is generally accepted that test scores have occasion-specific random measurement errors. As a consequence, if we were able to test an individual many times, each time with a brainwash in between, we expect to observe a *distribution* of scores. Lord and Novick (1968, p. 30) refer to this (hypothetical) distribution as the *propensity distribution*. Using the propensity distribution, we can compute the certainty that the individual will be consistently classified into a particular category (Emons et al., 2007). For a test to function reliably, this certainty needs to exceed a certain lower limit. For example, a certainty of 0.8 or higher is deemed necessary for making important decisions. However, because in real life we cannot retest an individual under identical conditions, the certainty cannot be established empirically, and we have to infer those properties from group-level information.

To illustrate the concepts at hand, Fig. 15.1 shows the propensity distributions for a hypothetical student, for three different dichotomously scored tests of 30 items. The tests differ in the mean and range of the item difficulties. Furthermore, we assume that the students taking this test will be classified into one of three mastery levels, defined by carefully selected cutoffs (e.g., Cizek & Bunch, 2007). Panel A shows the propensity distribution if the person is measured using a “broad” booklet having items informative across the entire ability scale. Panels B and C show the distributions for an easy and difficult booklet, respectively. The cutoffs for the broad booklet were arbitrarily set at 5 and 15, and the cutoff scores for the other booklets were obtained using true-score equating (e.g., Kolen & Brennan, 2014). Based on the propensity distributions, we can infer that this particular student has 81% *certainty* of being *consistently* classified at level II with the broad booklet (Panel

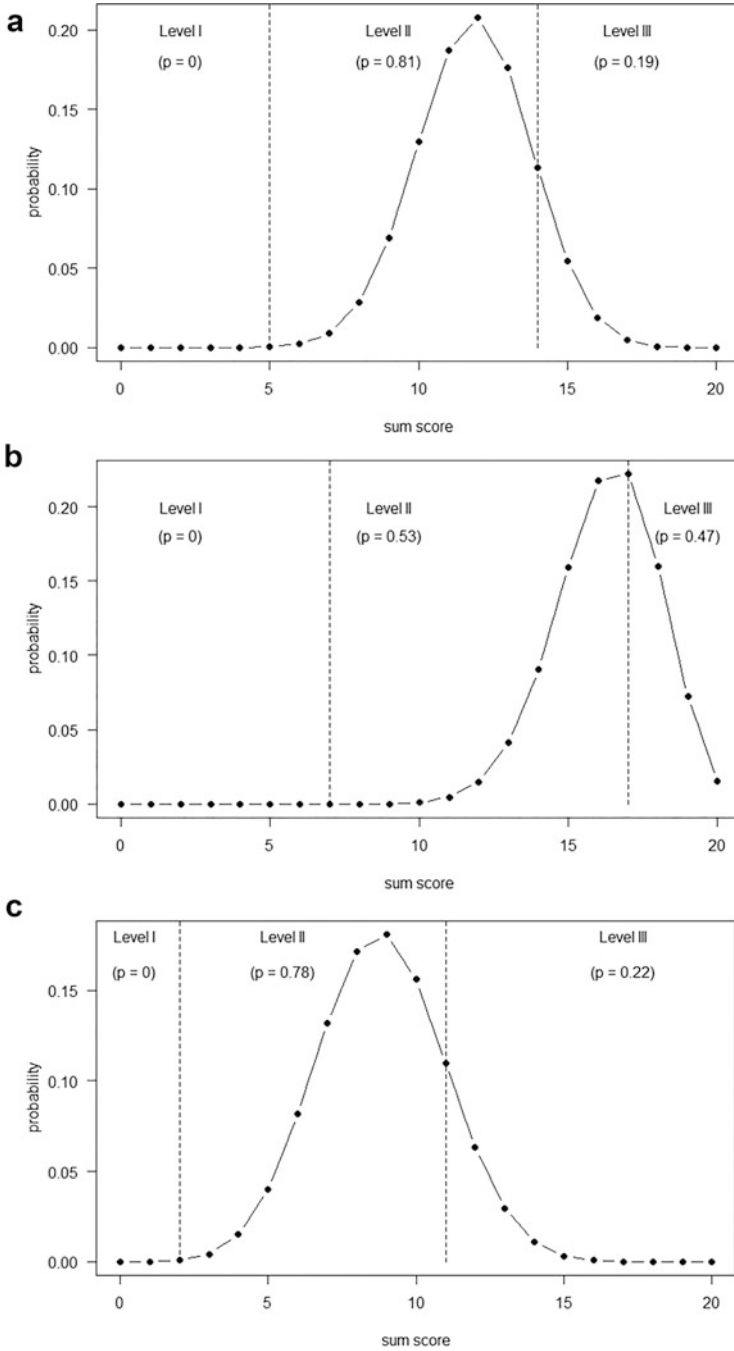


Fig. 15.1 Propensity distributions and classification certainties for a level-II student, for three different booklets: (a) a broad booklet, (b) an easy booklet, and (c) a difficult booklet

A), 53% *certainty* for the easy booklet (Panel B), and only 78% for the difficult booklet (Panel C). In general, for level II students, the certainty is highest if their ability lies halfway between the two cutoffs.

Two remarks are in order. First, for short (dichotomously scored) tests, the test scores are highly discrete. As a result, changing the cutoff – even if it is by only one score point – may affect the classification consistencies substantively. For example, if in Panel C the cut score was set at 12 instead of 11, then the certainty increases to 0.89. Second, as shown in Panel B, when tests are short and the difficulty does not match the ability, the propensity distribution is affected by floor or ceiling effects causing an asymmetry in the misclassification probabilities.

15.2.2 Group-Level Classification Statistics

To further quantify the reliability of tests for individual decision-making, Emons et al. (2007) introduced an index based on the individual classification certainties, which they called classification consistency. It is defined as follows. Let π^* denote a user-specified lower bound for the *desired certainty* of a *correct* classification. The choice of π^* depends on the consequences the test results can have for the individual. For high-stakes test and irreversible decisions, certainties of 0.8 or 0.9 are desired, whereas for low-stakes tests, lower values, say 0.7, may be deemed suffice. The CC_{π^*} is the proportion of individuals for whom the certainty of a correct classification exceeds π_c . For example, if $CC_{.8} = .65$, it means that about 65% of the test takers will be measured with at least a certainty of 0.8. Notice that Emons et al.'s CC considers individual-level consistency across *many* independent replications, not just two. Furthermore, it does not require a correction for chance as is the case with traditional group-level measures of consistency defined by two independent replications (e.g., Cheng & Morgan, 2013).

As noted above, accuracy refers to the extent to which classifications are correct given the true status of the students as reflected by their true scores. In the current context, accuracy can also be conceived as a measure for the (global) reliability of the test because the thresholds that mark the categories are defined in terms of the true score. Hence, misclassifications are due to random measurement errors only. However, when the test results are used for making inferences about educational achievement in populations (e.g., cohort studies), the inferences actually go into the other direction. Specifically, the question at hand is: Given the observed classifications, what proportion of students in each observed category truly has that mastery level? To answer this question, it is important to take not only the reliability but also the distribution over the categories in the population into account. By ignoring this population distribution while drawing conclusions from fallible information, one may fall prey to the base rate fallacy (e.g., Bar-Hillel, 1980).

An index that takes both reliability and the prior distribution into account is the *posterior predictive value* (PPV). The PPVs can be used to compare the practical use of different tests. For example, assume that students who score at mastery level

I are subjected to a tutoring program. If the PPV for mastery level I is 0.60, it means that for 60% of the students, the training was truly needed, but maybe not for the other 40%. It is important to emphasize that the students for whom the test result may be incorrect differ across replications, but with every repeated measurement, for about 40% students, the treatment may be ineffective.

15.3 Simulation Study

15.3.1 General Setup

We assume that students are classified into one of three ordered proficiency levels. The levels are referred to as mastery levels and labeled as *below basic*, *basic*, and *proficient*. In addition, we assume that different booklets are available, one general booklet covering all levels and a targeted booklet for each level separately. We will explore several compositions of the booklets with regard to the item difficulties and their match with the different mastery levels. Finally, we assume that booklets are assigned to pupils based on fallible background information such as teacher's appraisals. This means that the assignment of students to booklets is not necessarily optimal. In this simulation study, we will study the impact of different, possibly suboptimal, allocation strategies on consistencies and accuracy (to be explained below).

15.3.2 Data Generation

Data were generated using the one-parameter logistic model (1-PLM; Hambleton & Swaminathan, 1985). Let θ be a continuous latent variable (e.g., ability, proficiency). We arbitrarily assume that θ is standard normally distributed. The cutoffs that define the boundaries between mastery levels I and II and II and III were set at $\theta_1 = -0.674$ and $\theta_2 = 0.674$, respectively. As a result, about 25% in the population performs at level I (e.g., below basic), 50% at level II (e.g., basic), and 25% at level III (e.g., proficient). Furthermore, let X_j be the observed item response variable having realizations 1 for a correct response and 0 otherwise. The 1-PLM assumes unidimensionality, local independence, and response probabilities defined by

$$P_j(\theta) = P(X_j = x_j | \theta) = \frac{\exp[a(\theta - \delta_j)]^{x_j}}{1 + \exp[a(\theta - \delta_j)]}, \quad x_j = \{0, 1\}. \quad (15.1)$$

Parameter δ_j is the difficulty parameter. When $\theta = \delta_j$, we have $P_j(\theta) = 0.5$. Parameter a is the discrimination parameter. Notice that items are most informative for θ -levels around δ_j . Moreover, we assume that a is a constant for all items, which

amounts to assuming a Rasch model (Fischer & Molenaar, 1995). If $a = 1$, Eq. (15.1) reduces to the traditional Rasch model. The a -parameter is included to be able to easily manipulate the test-score reliability in the simulations.

15.3.3 Independent Variables

Number of Booklets Two conditions were considered. In the first condition, three targeted booklets were defined. One booklet consisted of items most informative around the lower cutoff, one booklet having items being informative around the upper cutoff, and one booklet that is a mixture of items informative about either the low or upper cutoff (generic booklet). Specifically, we sampled δs from the uniform distribution. For booklets 1 and 2, parameters δs were sampled from $U[-1.2, -0.2]$ and $U[0.2, 1.2]$, respectively. For booklet 3, half of the items were sampled from $U[-1.2, -0.2]$ and the other half from $U[-1.2, -0.2]$. In the second condition, we defined four booklets for the targeted test. One booklet having a majority of items for which the difficulty was below the lower cutoff, one booklet having predominantly items for which the difficulty was between the two cutoffs, one booklet for which the difficulty of *most* items was above the cutoff, and finally, one booklet where the item difficulties were spread out across the entire θ -scale. This setup was operationalized by sampling δs from $U[-2, -0.2]$ for booklet 1, $U[-1, 1.2]$ for booklet 2, $U[0.2, 2]$ for booklet 3, and $U[-2, 2]$ for booklet 4 (generic booklet).

Allocation Strategy We considered four scenarios for allocating booklets to the students: (1) the booklets were optimally matched with the true mastery level; (2) booklets were assigned randomly; (3) all students received the generic (non-mastery-level specific) booklet; (4) and booklets were assigned based on a deliberate *mismatch* with the true mastery level. The latter condition was implemented as follows. For the three-booklet case, the more difficult booklet was assigned to ability levels $\theta < 0$, and the easier booklet was assigned for ability levels $\theta > 0$. For the four-booklet case, ability levels $\theta < 0$ were assigned booklet 3 (most difficult), and all other students (i.e., mastery levels II and III) were assigned booklet 1 (easiest booklet).

Test Length Data were simulated for tests of 10, 20, and 30 items, respectively. According to Krueger et al. (2012), tests of at least 20 items are generally needed to have acceptable reliability for both individual decision-making and for group-level results. However, Krueger et al. did not specifically look at tailored testing, and their study was restricted to dichotomous classifications (e.g., accepting or rejecting applicants). See also Béguin and Straat (2019) for test length considerations in mastery testing. Using Bayesian analyses, they concluded that for dichotomous mastery decisions and a well-designed test, six items are the bare minimum to be able to decide about on mastery with enough certainty and, in general, that test lengths of 10 or more seem reasonable.

Reliability Reliability was manipulated by varying the a -parameter in the 1-PLM (Eq. 15.1). In particular we used $a = 1$ ('moderate' reliability) or $a = 1.5$ ('high' reliability). These a -levels resulted in tests with classical test-score reliabilities ranging from 0.65 to 0.92.

All factors were fully crossed. The result is a factorial design with $2 \times 4 \times 3 \times 2 = 48$ cells. Each cell in the design was replicated 500 times. In each replication, we drew new item parameters, so that the results can be generalized to a broader population of items and tests.

15.3.4 Dependent Variables

Individual-Level Classification Consistencies Let $f_{\theta}(X_b)$ be the conditional distribution of total score X_b for persons at ability θ completing booklet b ($b = 1, \dots, B$). For dichotomous items, the conditional X_b distribution is a compound binomial distribution (e.g., Lord, 1980, p. 45). Furthermore, as stated above, we assume that candidates are classified into one of three mastery levels, which are defined by two thresholds. The threshold values were -0.67 and 0.67 , respectively. The thresholds divide the θ -scale into three intervals, denoted θ_l ($l = 1, \dots, 3$), where $\theta_1 = (-\infty, -0.67)$, $\theta_2 = (-0.67, 0.67)$ and $\theta_3 = (0.67, \infty)$. Let p_l denote the population proportion of students at mastery level l ($l = 1, \dots, 3$). The thresholds on the θ -scale were defined such that p_1 and p_3 were about 0.25, and p_2 about 0.50. The latent thresholds are converted to thresholds on the sum score scale, denoted by c_l ($l = \{1, 2\}$), using the well-known relation $E(X) = \sum_{j=1}^J P_j(\theta)$ (e.g., Lord, 1980, p. 46). The three sum-score intervals defining the mastery levels, where the intervals are represented as $\mathbf{x}_1 = [0, c_1)$, $\mathbf{x}_2 = [c_1, c_2)$ and $\mathbf{x}_3 = [c_2, J]$.

Two remarks about the intervals \mathbf{x}_l are in order. First, the intervals are closed below, which means that when a person scores at the threshold, say c_l ($l = \{1, 2\}$), he or she will be classified at level $l + 1$. Second, because the sum scores represent a discretized measure of a continuous ability, the resulting population-level proportions of students at each mastery level may be slightly different from those obtained using θ . To signify the difference, we use p_l for the latent distribution of mastery levels and \tilde{p}_l for the distribution of mastery based on the sum-score distribution in the population.

For each mastery level, we can compute a classification consistency index for a desired certainty level. Let $\pi_{lb}(\theta)$ be the person-specific (conditional) certainty of being classified at mastery level l ($l = 1, \dots, 3$) when assessed with booklet b . This certainty is defined as

$$\pi_{lb}(\theta) = \sum_{x=0}^J I[x_+ \in \mathbf{x}_l] \cdot f_{\theta}(X_b = x), \tag{15.2}$$

where $I[\cdot]$ represents the indicator function taking the value 1 if the condition within the brackets is true, and 0 otherwise. Notice that $\sum_l \pi_{lb}(\theta) = 1$ because the categories are exhaustive and mutually exclusive. The certainties from Eq. (15.2) can be used to compute the classification consistencies (CC) given the minimum desired certainty level π^* ; that is,

$$CC_{\pi_c}(l) = \frac{1}{p_l} \int_{\theta} I[\pi_{lb}(\theta) > \pi^*] \cdot I[\theta \in \theta_l] d\theta. \tag{15.3}$$

We computed CCs for certainty levels of $\pi^* = 0.7$ and 0.8 . The integrand in Eq. (15.3) will be evaluated using 101 equidistant Gaussian quadrature points on the interval -3 to 3 . The CC conveys the proportion of students' mastery-level decisions is made with a minimum certainty.

Group-Level Consistency and Accuracy Evaluating targeted testing at the population level requires a formalization of the process by which booklets are assigned to test takers. Let $g(b|\theta)$ be the conditional probability of assigning booklet b to a student at level θ . The values of $g(b|\theta)$ depend on the specific mechanism by which booklets are assigned to students. For example, if the booklets would be assigned completely at random, we have $f(b|\theta) \equiv \frac{1}{B}$. However, if the same booklet would be assigned to all students at a certain ability level, we have $g(b|\theta) = 1$ for one particular booklet and 0 for all other booklets. Ideally, students are consistently assigned to the booklet that best matches their ability and thus for which the certainty of being assessed at the correct mastery level is highest.

Using the assignment probabilities $g(b|\theta)$, we can easily compute the relevant group-level statistics. Let $c(X)$ denote the *observed* mastery level given observed sum score X ; that is, $c(X) \in \{1, 2, 3\}$. Furthermore, let $c(\theta)$ be the true mastery level given θ ; that is, $c(\theta) \in \{1, 2, 3\}$. The marginal joint probability of *observing* mastery level l for respondents having a *true* latent mastery level k equals

$$p_{kl} \equiv p[c(X) = l, c(\theta) = k] = \int_{\theta} I[c(\theta) = k] \left(\sum_{b=1}^B [\pi_{lb}(\theta) \cdot g(b|\theta)] \right) d\theta. \tag{15.4}$$

From the marginal joint probabilities, we can compute the classification indices of interest.

Accuracy (ACC). The ACC expresses for each mastery level what proportion of the students would have an observed score indicating that particular mastery level; that is,

$$p[c(X_+) = l | c(\theta) = k] = \frac{p_{kl}}{\sum_{h=1}^3 p_{kh}}, \tag{15.5}$$

where $(k = 1, \dots, 3)$ and $l = k$.

Posterior predictive values (PPV): The PPV is the proportion of students who truly function at level m among the students for mastery level m is observed. The PPVs are obtained by

$$p[c(\theta) = k | c(X_+) = l] = \frac{P_{kl}}{\sum_{h=1}^3 P_{hl}} \quad (15.6)$$

where ($l = 1, \dots, 3$) and $k = l$.

15.4 Results

Table 15.1 shows the classification consistencies for the condition with three booklets, for two certainty levels $\pi^* = 0.7$ and 0.8, for low reliability (upper panel) and high reliability (lower panel). The results read as follows. Consider the value of 0.46 for mastery level I, booklet 1 and $J = 10$ items. This value suggests that when booklet 1 is used, about 46% of the students at mastery level 1 will be correctly classified at that level with a certainty of *at least* 0.70. Hence, booklet 1 is deemed reliable enough for individual decision-making for 46% of the students at mastery level I. For the other 54%, the measurement errors in the observed scores are too large to reach individual decisions at the desired certainty level. In this particular example, the difficulty of booklet 1 matches mastery level I. As the Table 15.1 shows, if booklet 2 is used for level I students, meaning a less optimal match between ability and difficulty, then only 32% of the students at mastery level I in the population would be classified with the desired certainty. Likewise, about 41% of students is assessed with the desired precision if the general booklet is used.

In general, our results suggest that the particular choice of the booklet may have a profound impact on the reliability with which *individual* decisions are made. The effect is largest for high-reliable short tests and decreases as test length grows, or as reliability gets lower. For example, consider the high reliable 10-item test with three booklets (Table 15.1; lower panel). Booklet 2 would only yield reliable decisions for about 28% of the test takers at level-1, whereas booklet 1, which difficulty matches the mastery level, reaches the desired reliability for about 60% of the test takers. Table 15.1 also shows that the general booklets were less reliable for individual decision-making than the targeted tests, but differences with the optimal situation were modest. Results thus suggest that the general booklet provides a safe choice if, for example, teachers feel insecure about their a priori ability estimate.

Table 15.1 further suggests that classification certainties for mastery level II students are smallest and for short tests and low reliability even dramatically low. This trend can be explained by a combination of limited scale range and the fact that the presence of measurement errors can play out in two ways. That is, the test-takers may obtain a score so high that they are rated at a higher mastery level than their true score justifies, but it might as well go the other way. As a result, the certainty that he/she score that exactly matches his/her mastery level may be low. This effect

Table 15.1 Classification consistencies for three Booklets, for low reliability (upper panel) and high reliability (lower panel)

Mastery level	Booklet:	Certainty level (π^*)					
		0.7			0.8		
		1	2	3	1	2	3
Low reliability							
$J = 10, \rho = 0.66$							
I		0.46	0.32	0.41	0.32	0.18	0.27
II		0.00	0.15	0.04	0.00	0.00	0.00
III		0.95	0.84	0.87	0.67	0.63	0.64
$J = 20, \rho = 0.80$							
I		0.63	0.53	0.59	0.50	0.37	0.45
II		0.57	0.64	0.61	0.27	0.33	0.32
III		0.87	0.84	0.85	0.67	0.68	0.68
$J = 30, \rho = 0.87$							
I		0.70	0.62	0.67	0.59	0.48	0.55
II		0.68	0.72	0.70	0.48	0.52	0.51
III		0.86	0.85	0.85	0.70	0.72	0.71
High reliability							
$J = 10, \rho = 0.80$							
I		0.60	0.28	0.51	0.48	0.14	0.37
II		0.47	0.70	0.60	0.21	0.42	0.32
III		1.00	0.90	0.93	0.82	0.74	0.75
$J = 20, \rho = 0.89$							
I		0.73	0.55	0.67	0.64	0.40	0.56
II		0.67	0.77	0.74	0.51	0.60	0.58
III		0.95	0.89	0.90	0.77	0.78	0.76
$J = 30, \rho = 0.92$							
I		0.79	0.66	0.75	0.70	0.54	0.65
II		0.74	0.81	0.79	0.60	0.67	0.66
III		0.92	0.89	0.90	0.78	0.80	0.79

becomes less severe when test length grows or reliability increases. Nonetheless, even with a reliable 30-item test, the amount of students for whom the test does not meet the desired certainty level is substantial.

Most remarkable trends were found for mastery level III. Results suggest that level-III test takers are more accurately assessed using booklet 1 than when the targeted booklet 2 was used, which is in contrast with our expectations. This trend can be explained by the fact that decisions are based on discrete scores, discrete cutoffs, and a limited score scale. Consider for example $J = 10$, which yields $X \in \{0, 1, 2, \dots, 10\}$. Now suppose that for booklet 1 test takers are assigned to mastery level III if their X -score is greater or equal to 9. With a relatively easy test, for almost all test takers at mastery level III, which are students having a high proficiency, it is likely that they will answer all items correct or at most miss one

item and thus the probability of having a score of *at least* 9 equals *at least* 0.70 for most of them.

Table 15.2 shows the group-level statistics. Column 1 shows the overall (marginal) proportion of correct classifications. For example, for $J = 10$, low reliability, and matched booklet assignment, we expect that 68% of the students in the population is categorized at the correct mastery level. The overall proportion of correct classifications increases with test length and with increasing reliability given test length. The effects of different assignment strategies were small.

Columns 2 through 4 show the conditional proportions of a correct classification. For example, we see that *of all* students whose ability is truly at mastery level I, about 70% will be categorized at level I based on their observed score X when the booklets are chosen such that match the true master level. Likewise, of all students at mastery level II, about 61% will be categorized at level II, and of all level-III students, about 84% will be correctly classified.

Comparison of Tables 15.1 and 15.2 shows the differences between consistency defined at the individual level and what is realized at the group level. For example, about 60% of mastery will end up in the correct level, but for few students, this correct classification can be assured even when the test would be replicated. Furthermore, comparing the results across different test lengths and reliabilities suggests that effects of booklet assignment are limited. Meaningful effects are only observed if students would collectively and systematically be allocated a booklet that does match the true mastery level.

The final three columns show the posterior predictive values. To illustrate, consider the value of 0.73 at mastery level I, for $J = 10$ and low reliability. This value suggests that of *all* students who *have been* categorized as a master level I student by their observed sum score, about 73% truly masters the content at level I. Just like the previous results, the booklet assignment mechanism had a minor effect. In many conditions, the broad booklets performed comparable to the situation where the booklets were optimally matched to the individual's abilities. Notice that the PPVs for mastery level III also show how using discrete scores affect the inferences. The PPVs at mastery level III are smallest because when the student has a score *at* the upper cutoff, the student migrates to level III. But if students score at the lower cutoff, they remain at level II. As a result, the chance of migrating upwards is higher than migrating downwards. Together the accuracy and PPV convey that if a student is a level-III student, we can be *certain* that he/she produces an observed test score corresponding to level III, but of all students for whom a test score indicating mastery level III is observed, a considerable percentage does not have a *true* mastery level III. Only for $J = 30$ do the percentages seem acceptable from a practical point of view. Tables 15.3 and 15.4 show the results for the four-booklet test. The trends are the same as for the condition with three booklets.

Table 15.2 Group-level statistics for three booklets, for different assignment Scenarios, for high reliability

Scenario	Marg. Corr. Class	Mastery level:	Accuracy			PPV		
			I	II	III	I	II	III
Low reliability								
$J = 10, \rho = 0.66$								
Matched	0.68		0.70	0.61	0.84	0.73	0.74	0.60
Random	0.66		0.67	0.58	0.84	0.73	0.71	0.58
Biased	0.65		0.62	0.59	0.83	0.71	0.70	0.57
Broad	0.67		0.65	0.58	0.85	0.74	0.70	0.59
$J = 20, \rho = 0.80$								
Matched	0.77		0.76	0.73	0.88	0.79	0.80	0.70
Random	0.75		0.75	0.71	0.88	0.78	0.80	0.68
Biased	0.74		0.71	0.71	0.88	0.77	0.78	0.68
Broad	0.75		0.75	0.72	0.88	0.78	0.80	0.70
$J = 30, \rho = 0.86$								
Matched	0.81		0.80	0.78	0.88	0.83	0.83	0.76
Random	0.80		0.80	0.78	0.88	0.83	0.83	0.76
Biased	0.78		0.76	0.75	0.88	0.79	0.81	0.72
Broad	0.80		0.80	0.78	0.88	0.83	0.83	0.76
High reliability								
$J = 10, \rho = .80$								
Matched	0.76		0.75	0.71	0.89	0.82	0.81	0.69
Random	0.74		0.69	0.70	0.88	0.80	0.79	0.64
Biased	0.70		0.59	0.67	0.90	0.72	0.78	0.58
Broad	0.74		0.71	0.71	0.88	0.81	0.78	0.66
$J = 20, \rho = .89$								
Matched	0.84		0.83	0.81	0.91	0.85	0.87	0.78
Random	0.81		0.78	0.78	0.89	0.84	0.83	0.75
Biased	0.77		0.71	0.75	0.88	0.81	0.79	0.71
Broad	0.82		0.80	0.79	0.89	0.84	0.84	0.75
$J = 30, \rho = .92$								
Matched	0.87		0.87	0.85	0.92	0.88	0.89	0.83
Random	0.84		0.84	0.81	0.91	0.85	0.87	0.79
Biased	0.81		0.80	0.77	0.90	0.83	0.84	0.74
Broad	0.85		0.84	0.82	0.91	0.86	0.88	0.80

15.5 Discussion

This chapter explored the efficiency of targeted testing for assessing educational achievements at the individual level and at aggregated levels. Targeting testing involves a modest degree of adaptivity in which fixed linear tests of varying difficulty are assigned to the test takers using background information related to their abilities. When the booklets are assigned optimally – i.e., such that the overall

Table 15.3 Classification consistencies for the four booklets test, for different assignment scenarios, for low reliability (upper panel) and high reliability (lower panel)

Mastery level	Certainty level (π^*)							
	0.7				0.8			
Booklet:	1	2	3	4	1	2	3	4
Low reliability								
$J = 10, \rho = 0.65$								
I	0.43	0.40	0.26	0.37	0.30	0.26	0.14	0.23
II	0.07	0.03	0.19	0.15	0.00	0.01	0.00	0.00
III	0.97	0.87	0.84	0.89	0.71	0.64	0.62	0.64
$J = 20, \rho = 0.78$								
I	0.61	0.59	0.46	0.56	0.48	0.45	0.31	0.42
II	0.42	0.61	0.58	0.55	0.04	0.32	0.20	0.16
III	0.89	0.85	0.83	0.84	0.67	0.68	0.67	0.367
$J = 30, \rho = 0.85$								
I	0.69	0.67	0.57	0.65	0.57	0.55	0.43	0.52
II	0.62	0.71	0.69	0.68	0.39	0.52	0.47	0.46
III	0.87	0.85	0.85	0.85	0.69	0.71	0.71	0.70
High reliability								
$J = 10, \rho = 0.77$								
I	0.57	0.51	0.32	0.45	0.45	0.37	0.17	0.33
II	0.17	0.60	0.40	0.49	0.02	0.33	0.04	0.15
III	1.00	0.93	0.89	0.94	0.85	0.75	0.73	0.74
$J = 20, \rho = 0.87$								
I	0.71	0.67	0.46	0.64	0.60	0.56	0.32	0.52
II	0.57	0.75	0.74	0.72	0.35	0.59	0.52	0.53
III	0.97	0.90	0.89	0.90	0.80	0.77	0.77	0.75
$J = 30, \rho = 0.92$								
I	0.77	0.75	0.57	0.72	0.68	0.65	0.42	0.61
II	0.68	0.80	0.79	0.77	0.53	0.67	0.63	0.63
III	0.95	0.90	0.88	0.90	0.78	0.79	0.79	0.78

difficulty matches the ability of the test taker – then the targeted test functions as an adaptive test by optimizing measurement precision with the same test length. Targeting testing tries to balance the psychometric benefits of computerized adaptive testing and practical benefits of linear tests. A good match ensures that tests are as accessible as possible and that testees experience a good balance between intrinsic and extraneous cognitive load throughout the test (Sweller, 1994).

One of the goals of our study was to provide some guidelines for practical use of targeted classification tests for multiple purposes. This line of research may help practitioners to find a better balance between the amount of test time and the quality of the test results. To accomplish our goal, we took two different perspectives. First, how do different implementations of targeted testing affect the reliability of individual decision-making? Results suggest that at the individual

Table 15.4 Group-level statistics for four booklets, for different assignment scenarios, for high reliability

Scenario	Marg. Corr. Class	Mastery level:	Accuracy			PPV		
			I	II	III	I	II	III
Low reliability								
$J = 10, \rho = 0.65$								
Matched	0.67		0.65	0.59	0.84	0.74	0.71	0.58
Random	0.65		0.64	0.57	0.84	0.73	0.70	0.57
Biased	0.63		0.61	0.53	0.88	0.70	0.71	0.54
Broad	0.65		0.64	0.57	0.84	0.73	0.70	0.57
$J = 20, \rho = 0.78$								
Matched	0.76		0.76	0.71	0.85	0.79	0.78	0.71
Random	0.74		0.72	0.70	0.84	0.78	0.76	0.68
Biased	0.71		0.68	0.65	0.84	0.77	0.73	0.62
Broad	0.74		0.72	0.70	0.84	0.78	0.76	0.68
$J = 30, \rho = 0.85$								
Matched	0.80		0.80	0.78	0.88	0.83	0.83	0.76
Random	0.79		0.76	0.75	0.88	0.79	0.81	0.73
Biased	0.76		0.75	0.72	0.88	0.82	0.80	0.69
Broad	0.79		0.76	0.75	0.88	0.79	0.81	0.73
High reliability								
$J = 10, \rho = 0.80$								
Matched	0.75		0.75	0.71	0.88	0.82	0.80	0.69
Random	0.72		0.69	0.65	0.88	0.78	0.74	0.66
Biased	0.68		0.65	0.55	0.90	0.81	0.67	0.63
Broad	0.72		0.68	0.67	0.88	0.81	0.76	0.62
$J = 20, \rho = 0.89$								
Matched	0.82		0.80	0.80	0.88	0.87	0.83	0.76
Random	0.79		0.76	0.76	0.88	0.83	0.81	0.73
Biased	0.75		0.71	0.69	0.88	0.81	0.78	0.66
Broad	0.80		0.76	0.76	0.88	0.83	0.81	0.73
$J = 30, \rho = 0.92$								
Matched	0.86		0.84	0.84	0.92	0.88	0.88	0.82
Random	0.83		0.80	0.82	0.92	0.87	0.85	0.79
Biased	0.79		0.75	0.76	0.88	0.86	0.81	0.71
Broad	0.83		0.80	0.82	0.88	0.87	0.84	0.79

level, the reliability of the conclusions drawn about the student’s mastery level is significantly impacted by how well the booklet level matches the student’s level. On the other hand, the results suggest that if there is no convincing information available to decide which targeted booklet is most appropriate, it is advised to choose the general booklet. Simulations also emphasize the importance of using enough items when students are categorized in three or more categories. This is especially important if the tests are used to make placement decisions. Furthermore,

test length requirements become more stringent if the number of categories grow (Cheng & Morgan, 2013). As an aside, if it turns out that lengthy tests are needed to be able to reliably classify students, it may also be a reason to take a closer look at the categories and to re-evaluate whether the categories are not defined too narrow.

Second, we also explored to what extent group-level results are affected by different implementations of the targeted test and suboptimal assignment of forms. These results have practical consequences when targeted tests are used for, for example, evaluating educational achievements at the school or national level, or when results are used for educational policy analysis. A specific example from the Netherlands may be a point in case. Reference levels have been developed for Dutch and arithmetic in order to be able to evaluate educational progress, either at the individual level, the school level, or the national level. These reference levels reflect the desired levels for Dutch language proficiency and arithmetic/math students should have at different points in their educational career. Hence, these reference levels form benchmarks against which individual or general learning performance can be measured. A similar approach has been developed for English proficiency (i.e., Common European Framework of Reference for Languages, CEFR). The percentages of students that reach different levels are monitored over time to identify important trends. Standardized tests are available for each level. However, most often students are assigned a test given the number of years of education completed, rather than the booklet that presumably matches their current ability level. This system of student monitoring may gain effectivity when the level of the administered tests better match the individual mastery levels of the students.

This research is only a start of hopefully a line of research that provides more guidance in general design issues for test administrations. As a follow-up on this study, we see four possible directions for future research. First, the importance of the conclusions drawn may differ greatly. For a single cut score, a test material can be optimized for measurement precision at that single point, but in this line of research, we are mainly interested in more precise measurement across the entire scale. The focus was now on correct prediction of one out of three levels, but future research may focus on more levels, the minimum level – i.e., prevention of under estimation – or precise point estimation across the entire scale.

Second, the present research may be extended to other forms of adaptive systems, such as computerized adaptive testing and multistage testing. Test adaptivity has a clear advantage in that it follows the student's ability level during the test administration. Hence, there is less need for a high-quality a priori indication of a student's ability level. On the other hand, when the high-quality a priori indication is available, a targeted test may be more efficient because all administered items are then at the student's ability level. It is interesting to also take this trade-off into account in future research.

Third, in our simulations we used the (unweighted) sum scores as the basis for mastery-level decisions. The use of sum scores has some practical advantages. First and foremost they are easy to communicate to students. Especially when questions are scored differently, it is important that students know in advance how many points they can earn with each question so that they can divide their attention efficiently

and effectively. From a statistical perspective, taking the sum may not be the most efficient scoring algorithm as it ignores *which* items have been answered correctly, although correlations between the optimally weighted and unweighted scores are usually high (>0.95). Nevertheless, one may use weighted scoring as is implicitly done when using, for example, IRT scoring under the 2-PLM or generalized partial credit model (Van der Linden & Hambleton, 1997). However, given the high correlations between sum scores and IRT scores, it is indifferent whether sum scores or estimated abilities are used for decision-making. Therefore, we expect our results will be largely generalizable under IRT *scoring*.

As a final remark, we may add that this study only considered a single test administration. As mentioned before, student monitoring systems commonly consist of a complete testing program with multiple follow-up tests. If a single test administration leads to a wrong conclusion about the student's ability level, can a student easily recover from this single measurement error, or may this wrong classification cause misclassification on follow-up tests? Future research is needed to better understand how a testing program may affect the observed learning trajectory of a student, especially when previous measurement outcomes are used as entry level for the next assessment.

References

- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211–233. [https://doi.org/10.1016/0001-6918\(80\)90046-3](https://doi.org/10.1016/0001-6918(80)90046-3)
- Béguin, A. A., & Straat, J. H. (2019). On the number of items in testing mastery of learning objectives. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 121–134). Springer International Publishing. https://doi.org/10.1007/978-3-030-18480-3_6
- Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontiers in Education*, 4(1). <https://doi.org/10.3389/feduc.2019.00001>
- Cheng, Y., & Morgan, D. L. (2013). Classification accuracy and consistency of computerized adaptive testing. *Behavior Research Methods*, 45(1), 132–142. <https://doi.org/10.3758/s13428-012-0237-6>
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage.
- 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(5), 713–734. <https://doi.org/10.1177/00131640021970862>
- Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicologica: International Journal of Methodology and Experimental Psychology*, 32(1), 107–132.
- Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, 12(1), 105–120. <https://doi.org/10.1037/1082-989X.12.1.105>
- Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to schools closures during the COVID-19 pandemic. *PNAS*, 118(17), 1–7. <https://doi.org/10.1073/pnas.2022376118>

- Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. Springer-Verlag.
- Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, 21(2), 449–460. <https://doi.org/10.2307/1162454>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Academic Publishers.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams. *Applied Measurement in Education*, 19(3), 203–220. https://doi.org/10.1207/s15324818ame1903_3
- Kim, D., Choi, S. W., Um, K. R., & Kim, J. (2006). *A comparison of methods for estimating classification consistency [paper presentation]*. Annual meeting of the National Council of Measurement in Education.
- Kolen, M. J., & Brennan, B. L. (2014). *Test equating, scaling and linking: Methods and practices*. Springer.
- Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, 12(4), 321–344. <https://doi.org/10.1080/15305058.2011.643517>
- Lek, K., Feskens, R., & Keuning, J. (2020). *Het effect van Afstandsonderwijs op Leerresultaten in het PO [effects of distance learning on educational achievement in primary education] [research report]*. Cito.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), 179–199. <https://doi.org/10.1111/j.1745-3984.1995.tb00462.x>
- Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, 36(3), 227–242. <https://doi.org/10.1007/BF02297844>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Laurence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35(3), 229–249. <https://doi.org/10.1111/j.1745-3984.1998.tb00537.x>
- Mislevy, R. J., & Wu, P. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Reports Series No. RR-96-30-ONR). Educational Testing Service.
- Sijtsma, K. (2009). Correcting fallacies in validity, reliability and classification. *International Journal of Testing*, 9(3), 167–194. <https://doi.org/10.1080/15305050903106883>
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer New York. <https://doi.org/10.1007/978-0-387-85461-8>
- Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer.
- Van der Linden, W. J., & Veldkamp, B. P. (2004). Constrained item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, 29(3), 273–291. <https://doi.org/10.3102/10769986029003273>
- Vispoel, W. P. (1998). Review and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, 35(4), 328–345. <https://doi.org/10.1111/j.1745-3984.1998.tb00542.x>
- Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates.

- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4), 473–492. <https://doi.org/10.1177/014662168200600408>
- William, 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>
- William, D., & Leahy, S. (2015). *Embedding formative assessment; practical techniques for K-12 classrooms*. Learning Sciences International.
- Yan, D., Von Davier, A. A., & Lewis, C. (Eds.). (2014). Computerized multistage testing: Theory and applications. *CRC Press*. <https://doi.org/10.1201/b16858>
- Ysseldyke, J., Spicuzza, R., Kosciulek, S., & Boys, C. (2003). Effects of a learning information system on mathematics achievement and classroom structure. *The Journal of Educational Research*, 96(3), 163–173. <https://doi.org/10.1080/00220670309598804>
- Zenisky, A. L., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). Springer.

Part IV
New Psychometrics

Chapter 16

The Hierarchical Model for Response Times: Advantages, Limitations, and Risks of Its Use in Measurement Practice



Jesper Tijmstra and Maria Bolsinova

Abstract With the advance of computerized testing in educational and psychological measurement, the availability of response time data is becoming commonplace, and practitioners are faced with the question if and how they should incorporate this information into their measurement models. For this purpose, the use of the hierarchical model is often considered, which promises to improve the precision of measurement and has various other appealing properties. However, practitioners also need to be aware of the several limitations and risks involved when using this model, which have been covered less extensively in the literature. This chapter covers both the advantages and disadvantages of using the hierarchical model, to allow practitioners to form a balanced assessment of the potential use of the hierarchical model for their testing application.

16.1 Introduction

The testing of abilities and skills has a long history in both psychology and educational measurement. While until recently the default administration form of such tests was paper and pencil, with the advance of computerized testing in many fields of psychology and educational measurement, it is becoming commonplace to administer tests digitally. One clear benefit of this digital administration of tests is the potential availability of process data that can be registered in addition to the registration of the response that is provided (Goldhammer & Zehner 2017). These process data can come in many forms, ranging from registering the number of attempts made on an item to data based on advanced mouse- and eye-tracking techniques. However, by far, the most commonly considered type of process data

J. Tijmstra (✉) · M. Bolsinova

Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands
e-mail: J.Tijmstra@tilburguniversity.edu; M.A.Bolsinova@tilburguniversity.edu

is the registering of the response time (RT, the time that passes between reaching an item and providing the final response), a measure that is generally considered to at least potentially contain information that is relevant in a wide range of testing settings.

While there have been many different ways of looking at and using RTs proposed in the literature over at least the last 70 years (Gulliksen 1950; van der Linden 2009), one relatively new method that has gained a lot of attention in recent years is the hierarchical model (van der Linden 2007), which jointly models the RTs together with the correctness of the responses. Partly due to its relatively simplicity, and partly due to its promise to improve the precision of measurement, practitioners are not only becoming aware of the existence of this model but are taking steps toward implementing this model as part of their measurement practice. While the model itself is rather simple and well known, the challenges that one should be aware of when using this model in practice are both less straightforward and less well known. This chapter aims to address these issues by providing a comprehensive overview of what the hierarchical model for RTs has to offer for measurement practice, what its limitations are (and how some of these limitations can be addressed), and what the risks are of using this model in practice.

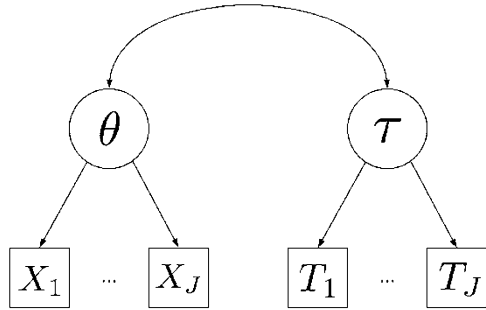
16.2 An Overview of the Hierarchical Model and Its Advantages

The hierarchical model consists of two measurement models, one concerning the accuracy of the response (RA, for item j denoted by X_j) and one concerning the speed of the response (RT, for item j denoted by T_j). The measurement model for RA concerns the latent ability parameter θ , while the measurement model for RT concerns the latent speed parameter τ .¹ The modeling framework leaves it open which specific measurement models are used for modeling RA and RT and as such is neutral with respect to the particular relationship that is expected between the response data and the latent variables in the model. In practice, standard IRT models are commonly considered for modeling RA, and RTs are often modeled through a lognormal model (van der Linden 2006).

Regardless of which particular measurement models are chosen, both measurement models are connected at a higher level, through the inclusion of correlations between the different item parameters (e.g., item difficulty and item time intensity) and the inclusion of a correlation between the person parameters θ and τ . It is through these correlations that the hierarchical model can explain possible associations observed at the response level between RA and RT. The general

¹ For reasons of simplicity but without loss of generality, we will only consider versions of the hierarchical model that have a single ability and a single speed parameter.

Fig. 16.1 The general structure of the hierarchical model



structure of the hierarchical model is presented in Fig. 16.1, which remains neutral with respect to the choice of measurement models for RA and RT.

16.2.1 Using RTs to Improve the Precision of Measurement

When contrasting the hierarchical model for RT and RA with standard IRT models that consider only RA, one clear advantage of the hierarchical model becomes readily apparent: In addition to the information about ability that is captured by the IRT measurement model that considers the RA, the hierarchical model also considers information about ability that is contained in the RTs (van der Linden et al. 2010). As Fig. 16.1 shows, the RTs are indirectly linked to ability, through the latent speed variable τ . Thus, if in the population speed and ability are correlated, the measurement model for speed provides collateral information for the estimation of ability, on top of what is provided by standard IRT models.

The correlation between speed and ability can take on any value between -1 and 1 , and in practice, positive values (e.g., see Loeys et al. 2011; Wang & Xu 2015; Meng et al. 2015), negative values (e.g., see Klein Entink et al. 2009; Goldhammer & Klein Entink 2011; Scherer et al. 2015), and values close to 0 (e.g., see van der Linden et al. 1999; Bolsinova et al. 2017; Shaw et al. 2020) have been observed. Rather intuitively, the amount of information that the RTs can provide for improving the precision with which ability is estimated is bounded by the size of this correlation: If there is only a weak correlation between speed and ability, even a perfectly estimated speed latent variable will only be able to explain a small part of the variance in the latent ability variable. This also means that the marginal amount of information about ability that is gained through the measurement model of speed by adding items to the test quickly decreases as the test increases in length: Once speed is estimated with a reasonable amount of precision, for the precision with which ability is measured, the gain of reducing the measurement

error with which speed is measured will be minimal (Ranger 2013).² This is in contrast with the measurement model for ability, where each new item contains new and independent information about ability that continues to increase precision as test length increases. Effectively, the RAs on all the items together with speed provide information about ability in the hierarchical model, and the relative relevance of the speed latent variable decreases as more RAs are observed, even if the latent speed dimension does end up being measured with lower measurement error as the test length increases. The consequence of this is that the biggest relative gains of using the hierarchical model instead of a “RA-only” model in terms of improving precision can be expected to be found for relatively short tests, where the added explanatory power of including an additional (imperfectly measured) predictor can be expected to matter the most.

16.2.2 Relevance of RT for Test Construction and Analysis

In addition to improving the precision of measurement of ability, the hierarchical model also provides the user with a more extensive toolbox to evaluate the quality of the test, the individual items, and the performance of individuals. In this sense, it can provide practitioners with more options for critically evaluating items during test construction, for evaluating the performance of an existing test, and for detecting aberrant responding.

Since for every item not only characteristics in the measurement model of ability are considered, but also characteristics in the model for speed are measured, a more complex and more complete picture emerges of the properties of the different items on the test. Not only is it possible to determine which items are relatively time intense, but it is also possible to assess the relationship between the different item characteristics in the two measurement models. Since in the context of the hierarchical model all commonly considered measurement models for ability and for speed contain a location parameter, this is also the most commonly studied association between the item characteristics (van der Linden 2009). Not unsurprisingly, the correlation between item difficulty and item time intensity is generally found to be positive, with more difficult items requiring on average more time from the respondents to be solved. While this pattern may not be unexpected, it is something that test constructors should keep in mind when designing a test, especially when there will be strict limits to the amount of testing time. Less studied, but equally relevant, is the relationship between time intensity and item discrimination in the RA model: Do items on the test that respondents spend more time on provide us with more information about ability than items that are answered more rapidly? If the answer is no for a particular testing setting, it may make sense

² This limiting aspect of the hierarchical model is addressed in one of its extensions, as will be discussed in the next section.

for test constructors to focus on designing items with at most a moderate time intensity, to optimize the total testing time or the precision of measurement of ability obtained within a certain time limit.

On the person side, a similar picture emerges. Not only do we obtain information about the speed with which different individuals answer items on the test, but we also gain insight into the relationship between the speed with which persons take the test and their overall performance (as captured by their estimated ability). Since this speed-ability correlation takes on wildly different values in practice, studying that correlation can be considered important for getting a better picture of the response processes of different types of respondents who take the test: Do people who work fast on average show a better or worse performance than those that take more time on the test? It is important to stress that since this correlation considers a *between-person* association, it should not be confused with the often studied “speed-accuracy” trade-off (Heitz 2014): the well-known phenomenon that increasing the speed with which one executes cognitive tasks generally decreases the accuracy of the outcome of that task. This speed-accuracy trade-off (which in the context of IRT might better be considered in the form of a “speed-ability trade-off”; van der Linden 2009) describes a negative *within-person* association, which does not need to translate to a negative association at the between-person level. That is, the speed-ability trade-off is only one of the factors that contributes to the between-person association between speed and ability. Another phenomenon that contributes to this association is well known from cognitive psychology: More competent persons may be able to execute a task both faster and with higher accuracy than less competent persons (i.e., have a speed-accuracy trade-off curve that is positioned above those of other respondents). This explains why it is possible for the between-person association of speed and ability to be positive, even though the within-person speed-ability trade-off pushes this association in the negative direction. When the speed-ability trade-off is the main factor driving between-person differences, a negative correlation between speed and ability will emerge. In those cases, one could be worried about the validity of measurement of ability, since it means that many respondents performed suboptimally on the test (i.e., unnecessarily sacrificed performance in favor of speed). This phenomenon may be especially prevalent in low-stakes assessment, where it may not be safe to assume that all respondents are fully engaged with the test and where differences in observed performance (as captured by estimated ability) could possibly to a large extent be attributable to differences in engagement rather than to differences in actual ability.

Finally, the hierarchical model extends the possibilities for detecting aberrant persons and items on the test, compared to what is possible using standard IRT models (van der Linden & Guo 2008). When using the hierarchical model, in addition to determining whether (a set of) responses should be considered an outlier in terms of the observed RAs, other outliers can be studied. On the person side, outliers in RTs on the full set of items could suggest that the person may not be taking the test seriously (in case of both overly fast or overly slow responses). On the item side, observing overly fast or overly slow responses for a significant portion of the respondents could suggest problems with that item, such as possible

guessing (in case of many fast responses) or possible issues with the clarity of the item (in case of many slow responses). Since the hierarchical model considers RAs and RTs simultaneously, these cases can be studied in further detail by considering whether the conjunction of the RA and RT of a (set of) response(s) should be considered an outlier. For example, observing many fast incorrect responses on an item might suggest that guessing is prevalent, while many fast correct responses might suggest that item preknowledge is a problem or that it can be solved using an unintended heuristic. While these patterns can to some extent be studied without the use of advanced psychometric models, the advantage of using the hierarchical model is that one can truly consider whether a (combination of) response(s) should be considered an outlier, since one can determine whether a (set of) residual(s) is extreme compared to what is expected under the model. This makes it possible to contrast an item that is simply so easy that many people provide a fast correct response to it with an item where a part of the population provides unexpectedly fast responses with an unexpectedly high rate of success.

16.2.3 Simple Structure and Flexibility

A final major advantage of using the framework of the hierarchical model is its relative simplicity and flexibility, which go hand in hand. The framework's flexibility comes from the fact that a simple structure is assumed and the two measurement models are separated and are only linked through correlations at the higher level. Because of this, one can consider a wide range of models for the RA side (including all commonly considered IRT models) and independent of that choice also consider different models for the RT side of the model. This makes it possible to choose a model specification that is tailored to the specific needs of the testing context that is considered.

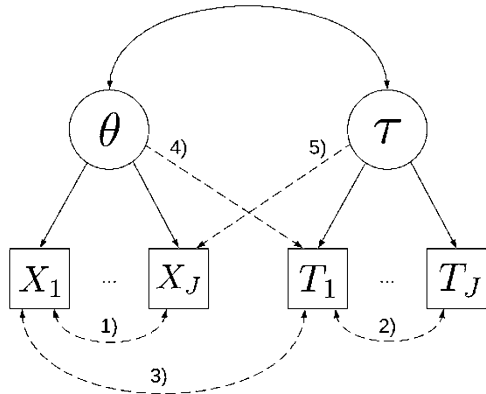
On the interpretation side, the simplicity that is entailed by this simple structure is also beneficial. On the RA side of the model, one generally uses one of the common IRT models for dichotomous or polytomous data, with the standard interpretation of both the item and the person parameters remaining applicable, unaffected by the fact that RTs are considered elsewhere in the model. Similarly, on the RT side, item and person parameters are considered that keep their standard interpretation and only relate to the RTs. The connection between the two measurement models is likewise easy to understand, since correlations between the different item parameters and between the different person parameters are considered. All of this can be considered to be an advantage for practitioners, both who themselves have to fully understand the workings of the model and who will need to be able to effectively communicate findings based on the models to stakeholders.

16.3 Limitations: A Range of Conditional Independence Assumptions

While the simplicity of the hierarchical model is often considered as one of its selling points, this simplicity is at the same time at the root of a set of limitations that have both an important practical and theoretical impact. That is, the assumption of a simple structure can often be considered problematic in practice, not only in the sense that the model shows less than perfect fit but also in the sense that important patterns may be overlooked or even that bias may occur in one of the outcome measures (e.g., the ability estimates or the estimated precision). It is therefore of great importance that practitioners are aware of these limitations before they consider applying the framework in practice.

The different limitations of the hierarchical model that will be considered in this section all relate to different conditional independence assumptions that are made by (all standard versions of) the hierarchical model. The hierarchical model as it was presented graphically in Fig. 16.1 shows that various variables in the model are not directly connected to each other, although all of them are indirectly connected. Figure 16.2 provides a graphical overview of the five different forms of conditional independence that are assumed by the model, where dashed lines indicate a residual correlation of 0 (i.e., conditional independence). A violation of any of these conditional independence assumptions constitutes a violation of the hierarchical model, which can result in various issues beyond simply a reduced model fit, all of which will be covered in this section.

Fig. 16.2 The hierarchical model and its five conditional independence assumptions. All conditional independence assumptions are indicated by numbered broken lines, which indicate the assumed absence of a relationship



16.3.1 *Conditional Independence of the RAs*

The first conditional independence assumption considered is that of the RAs given the latent variables:

$$P(\mathbf{X}|\theta, \tau) = P(X_1|\theta, \tau)P(X_2|\theta, \tau) \dots P(X_J|\theta, \tau),$$

where \mathbf{X} is the vector containing all the item responses X_1, \dots, X_J . Since the hierarchical model assumes a simple structure, the RAs do not depend on speed given ability, so this assumption reduces to the standard local independence assumption considered in IRT:

$$P(\mathbf{X}|\theta) = P(X_1|\theta)P(X_2|\theta) \dots P(X_J|\theta).$$

Compared to the other four assumptions of conditional independence, violations of local independence and their impact have been studied rather extensively (Yen 1984; Wainer & Thissen 1996; Chen & Thissen 1997; Hoskens & De Boeck 1997; Zenisky et al. 2001). Since this form of conditional independence is shared by almost all commonly used IRT models, and since one can in principle use a measurement model for RA that allows for local dependence, this conditional independence assumption will not be discussed here extensively. It is however important to note that the presence of local dependence generally results in an underestimation of the standard error of ability (e.g., see Zenisky et al. 2001), such that in its presence one overestimates the precision with which ability is measured.

16.3.2 *Conditional Independence of the RTs*

Similar to the assumption of conditional independence of the RAs, standard versions of the hierarchical model assume conditional independence of the RTs given the latent variables:

$$P(\mathbf{T}|\theta, \tau) = P(T_1|\theta, \tau)P(T_2|\theta, \tau) \dots P(T_J|\theta, \tau) = P(T_1|\tau)P(T_2|\tau) \dots P(T_J|\tau),$$

where \mathbf{T} is the vector containing the RTs T_1, \dots, T_J . Effectively, this assumption tells us that the RT of a response only depends on the overall speed of the respondent (and the item parameters in the RT model), but not on the RT of the previous response or of any other response.

While this assumption has not been studied extensively in the context of the hierarchical model, it links directly to the extensive literature on RT modeling. For example, the phenomenon of speeding on the test is well established in many testing settings with effective time limits (e.g., see Lu & Sireci 2007).

Similarly, it is well known that respondents generally spend a relatively long time answering the first few items presented on a test. Both of these phenomena concern violations of the assumption of the hierarchical model that the latent variables are “stationary” throughout the test (Fox & Marianti 2016). This non-stationarity of speed throughout the test may lead to conditional dependence between the RTs in two ways. Firstly, respondents may differ in the extent to which they work with a slow start and speeded conclusion on the test, which should result in positive dependence between the RTs of adjacent responses of items in the beginning or at the end of the test. Secondly, if the items are presented in booklets, the item position will likely be different for different respondents, and hence respondents will differ in whether they encounter the item in the beginning, in the middle, or at the end of the test. In that case, even if all respondents show the exact same pattern of slowing down in the beginning and speeding up near the end of the test, positive residual dependencies will remain between adjacent items (i.e., between items in a booklet).

While the impact of unmodeled conditional dependence between the RTs in the hierarchical model has to our knowledge not been studied, one can be hopeful that in practice its impact is relatively limited. That is, one can expect an impact similar to what is commonly found in IRT models where unmodeled local dependencies are present: an underestimation of the standard error of the latent variable in the measurement model. While this may be undesirable, its impact in settings where one mainly uses the hierarchical model for improving the precision of measurement of ability can be expected to be minor, since it only directly concerns the precision with which speed is estimated. It does however mean that there is relevant model misfit and that one misses potentially relevant information about the response processes. If getting a more complete picture of these processes is considered desirable, one could consider working with a more complex measurement model for RT that allows for local dependencies.

16.3.3 *Conditional Independence of RT and RA*

While the previous two forms of conditional independence both only concerned one of the two measurement models, the remaining three forms of conditional independence all concern the relationship between the RA and RT side of the hierarchical model. In this sense, the remaining three forms of conditional independence can be considered to be unique to models that jointly consider RA and RT. The most well known and well studied of these assumptions is conditional independence of RT and RA:

$$P(\mathbf{X}, \mathbf{T} | \theta, \tau) = P(\mathbf{X} | \theta, \tau) P(\mathbf{T} | \theta, \tau).$$

This assumption of conditional dependence thus states that once the latent variables are taken into account, the accuracy of the response is not linked to the RT:

Unexpectedly fast or slow responses cannot be expected to be more (or less) likely to be correct, and vice versa.

Conditional dependence between RA and RT implies that the association between RA and RT that is observed is not fully explained by the two latent variables in the model and hence that unexplained patterns remain. Since the hierarchical model purports to fully explain the observed association between RA and RT, this form of conditional dependence can be considered conceptually important. As will be discussed below, its presence both poses risks for the model inferences and creates opportunities to gain better insight into the response processes for specific items and for specific persons.

Bolsinova et al. (2017) have provided an extensive overview of the various possible sources of positive and negative conditional dependence, which will briefly be summarized here. As they point out, conditional dependence may both be present in situations where all individuals answer the items in similar ways (i.e., homogeneous response processes) and when individuals differ in how they answer the items (i.e., heterogeneous response processes).

When respondents take the test in similar ways, conditional dependence may occur due to between-person differences in the item parameters (i.e., differential item functioning). If differential item functioning (DIF) is present, an item may be relatively more difficult for one respondent than for another respondent with the same ability level. Since time intensity is generally positively correlated with item difficulty, it is reasonable to expect the item time intensity to similarly show DIF, meaning that respondents for whom the item is relatively difficult may also spend a relatively large amount of time on solving the item, introducing DIF for the item time intensity parameter as well. This covariation of item difficulty and item time intensity will generally result in negative conditional dependence, since those persons who find the item more difficult are both expected to provide a less accurate and slower response to the item. While DIF is normally only studied in the context of contrasting specific subgroups in the population that is tested, the negative conditional dependence described here can occur even if there is no DIF that links specifically to group membership, but only concerns “unexplained” between-person variation in the item parameters (e.g., the item having a higher difficulty parameter for one respondent than for another, without this difference being attributable to group membership). Such DIF is not studied in practice for the obvious reason that there is always too little data to consider it (since it concerns person-by-item interactions rather than group-by-item interactions), but this does not mean that such between-person variation in the item parameters should not be expected, as Bolsinova et al. explain (2017). Thus, any between-person covariation of item difficulty and item time intensity is sufficient for causing negative conditional dependence, and this covariance can be present even if the DIF on the RA and on the RT side average out at the level of the different groups and hence is not detected. This means that standard DIF analysis (even if extended to the hierarchical model) will not be able to show that such DIF is not present, since it only considers variation in the item parameter(s) across a small prespecified set of respondent

groups. Unfortunately, this means that in practice excluding the possibility of this kind of DIF is empirically practically infeasible.

Additionally, conditional dependence may occur due to non-stationarity of the two latent variables. That is, while the hierarchical model assumes that all persons work at a constant speed and with a constant ability level, this assumption may often be unrealistic in practice. On any test with an effective time limit, speeding near the end of the test will occur for at least a subset of the respondents, meaning that their effective speed for those later items is higher than it was for the earlier items. Due to the speed-accuracy trade-off, we can expect responses to those later items to be both faster (i.e., negative residual RT) and more often incorrect (i.e., negative residual RA), resulting in positive dependence.

While the abovementioned sources of conditional dependence between RA and RT concern situations where respondents still take the test in comparable ways, additional sources of conditional dependence may play a role when there are qualitative differences in how respondents take that test. That is, when the response processes of respondents differ for a particular item, these differences can be expected to result in conditional dependence between the RA and RT of responses to that item. The most obvious example is rapid responding, which means that some respondents provide low-quality fast responses to the item, introducing positive dependence. In contrast, slow disengaged or unmotivated responding would result in negative dependence. Additionally, when engaged respondents show differences in their answer strategy, conditional dependence can be expected. For example, when some respondents produce the answer to an item through heuristics, while others solve the item algorithmically, both differences in the expected RA and the expected RT will be present, leading to dependence.

With all these different possible sources of conditional dependence between RT and RA, it should not come as a surprise that this assumption often appears to be violated in practice (Ranger & Ortner 2012; Meng et al. 2015; Bolsinova et al. 2017; Bolsinova et al. 2017; Bolsinova & Molenaar 2018). It should also be noted that both positive and negative conditional dependence between RA and RT can be observed within the same test. This will, for example, be the case if a heuristic approach leads to the correct response on one item, while it leads to an incorrect response on another item. Thus, conditional dependence between RA and RT should always be studied at the item level.

It may be noted that in addition to a possible dependence between the RA and RT on the same item, dependencies across items are also possible. For example, the well-studied phenomenon of post-error slowing (Rabbitt & Rodgers 1977; Laming 1979) suggests that there may often be a negative dependence between the RA of one response and the RT of the subsequent response. To our knowledge, this phenomenon has not been studied in the context of the hierarchical model, but it seems reasonable to assume that the impact of this kind of violation of conditional dependence will be similar to the impact of conditional dependence between RA and RT of the same item.

Beyond the fact that misfit shows that the model inadequately captures the patterns observed in the data, the presence of conditional dependence between RA

and RT suggests that there may be important aspects of the response process that are not captured by the model or perhaps even misrepresented. Thus, a variety of extensions of the hierarchical model have been considered (Ranger & Ortner 2012; Meng et al. 2015; Bolsinova et al. 2017; Bolsinova et al. 2017) that attempt to incorporate possible residual dependencies between RA and RT in the model. These models generally provide a more flexible toolkit for jointly modeling RA and RT, allowing users to get a more complete picture of the response processes and item and person characteristics, at the cost of increased model complexity. Thus, it can be considered important to first critically test for the possible presence of conditional dependence between RA and RT (e.g., using the test proposed by Bolsinova & Tijmstra 2016) and subsequently explore the use of one of the extensions of the hierarchical model if needed and desired.

16.3.4 Conditional Independence of RT and Ability

In addition to the RT of a response possibly depending on the RA of that response or the RT of other responses, there is also the possibility that RT depends on ability. That is, there may be difference between persons of different ability levels in terms of how much time they spend on each item, beyond what can be explained through their overall speed. This would entail a violation of the following conditional independence assumption:

$$P(\mathbf{T}|\theta, \tau) = P(\mathbf{T}|\tau).$$

This possibility was considered by Bolsinova and Tijmstra (2018).

Conceptually, the possibility of ability being linked to how much time a respondent spends on one item, relative to the other items, makes a lot of sense. Low-ability respondents in all likelihood realize that some of the more difficult items are too difficult for them to solve and may decide to allocate most of their limited time to solving the easier items, where they do stand a reasonable chance of finding the right answer. In contrast, high-ability respondents likely do not need to spend a lot of time in solving easy items and allocate most of their time to tackling the more difficult items. Effectively, the hierarchical model states that throughout the entire test, there will be no difference in how high-ability respondents allocate their time, compared to low-ability respondents. This assumption may not be plausible in most practical testing settings.

The ignored possibility of conditional dependence between RT and ability is not only a limitation for the standard hierarchical model in the sense that it introduces model misfit, but it also means that not all relevant information about ability that is contained in the RTs is utilized by the model. That is, conditional dependence between RT and ability means that there is collateral information in the RTs for the estimation of ability, beyond that which is contained in the overall correlation between speed and ability. Bolsinova and Tijmstra (2018) developed a model that

allows for this kind of conditional dependence and found that in practice the gain in precision with which ability is estimated when allowing for this dependence can be notable and may exceed the original gain in precision when moving from an IRT model to the standard hierarchical model (i.e., from including speed as a predictor of ability). This is especially likely for larger tests, since in the extended model the collateral information in RT for the estimation of ability effectively increases linearly with every additional item, while in terms of collateral information, the standard hierarchical model can never do better than the inclusion of a single perfectly estimated covariate (i.e., speed). Thus, if one's main reason for using the hierarchical model is to increase the precision with which ability is estimated, it makes sense to explore whether the extended model proposed by Bolsinova and Tijmstra makes better use of the collateral information from the RTs than the standard hierarchical model.

16.3.5 Conditional Independence of RA and Speed

In addition to the RA of a response possibly depending on the RAs of other responses and the RTs, it may also be the case that under the hierarchical model, a residual association remains between RA and speed. In that case, one is dealing with a violation of the following conditional independence assumption:

$$P(\mathbf{X}|\theta, \tau) = P(\mathbf{X}|\theta).$$

Such violations can be expected when the effect of “operating speed” on the probability of success is not the same for all items. For example, it may be realistic that some items can be solved rather easily using heuristics, in which case a high speed would not necessarily lead to a low expected RA or a lower expectation than what is expected for respondents operating at lower speed levels. If there are other items on the same test where using heuristics does not lead to the correct (and possibly to an incorrect) answer, respondents who operate at that same high speed level would now be expected to do relatively worse compared to respondents operating at a lower speed level. This differential impact of speed on the expected accuracy of the response for different items would show up as a negative residual dependence between speed and RA.

Extensions of the hierarchical model that specifically attempt to address possible residual dependencies between RA and speed have to our knowledge not been developed. Additionally, no formal study into the possible presence of this kind of dependence in real life data has to our knowledge been conducted, nor have tests been developed that specifically aim to detect such possible dependence. However, an approach similar to the one proposed by Bolsinova and Tijmstra (2018) for dealing with conditional dependence between RT and ability could be explored. While such an extension would not lead to a notable improvement in the precision with which ability is estimated, it would provide users with relevant information

about how different items function, which can be considered relevant for testing practice and especially test design (e.g., intentionally ex- or including items where fast operating speed improves the expected accuracy).

16.4 Risks of Using the Hierarchical Model in Practice

In addition to the formal and practical limitations of the standard hierarchical modeling framework discussed above, there are important risks and misconceptions of the framework that should be well understood by practitioners before they choose to use the model in practice, which will be covered in this section.

One important misconception that should be avoided concerns the interpretation of the correlation between speed and ability in the model. Given that in standard formulations of the model, τ effectively captures the (weighted) average RT on the test what could be called “effective speed,” while θ captures “effective ability” (i.e., overall performance on the test), all that this correlation tells us is whether persons who provide answers faster generally do so with higher or lower accuracy than those who provide answers more slowly (Tijmstra & Bolsinova 2018). While it may be tempting to take this between-person association and assume that it informs us what would happen to the expected performance of respondents if they would provide answers more (or less) quickly, no such inferences can be made, since this concerns a (counterfactual) within-person association that cannot be assessed based on the model. Fundamentally different models and a fundamentally different testing setting are needed if one wants to assess this within-person speed-ability trade-off (e.g., see Goldhammer 2015), which require respondents to operate at different levels of effective speed.

Another potential risk of the hierarchical model is that unlike standard IRT models, the estimates of ability depend on more than just the accuracy of the responses, since the correlation between speed and ability means that speed estimates affect ability estimates. Of course, this was also one of its main selling points, but the inclusion of speed as an additional predictor of ability does run the risk of introducing bias. That is, while the precision of measurement will increase through the inclusion of this additional predictor, if the actual relationship between these two variables does not fully match their relationship in the model, we will introduce bias in the ability estimates that would not have been there if we had used a “RA-only” model. With the complexity of standard test taking settings in mind, it may not be overly realistic to assume that the simple linear relationship between speed and ability completely and correctly captures the relationship between RT and RA, meaning that at least some degree of bias in the estimate of ability should be expected. Thus, the risk of introducing systematic bias is prominent if the actual relationship between speed and ability is not captured well by a linear correlation. This will, for example, be the case when respondents differ in how they take the test and, for example, a subset of the respondents provide fast disengaged responses. It is therefore important to ascertain that respondents all took the test in similar ways

(e.g., with similar levels of engagement and using similar response processes). This will of course be difficult to actually establish in testing practice, where there is only a limited amount of information available per respondent.

There is another risk that follows from using speed as an additional predictor of ability that specifically applies to high-stakes testing. Since the speed with which responses are given will have an influence on the estimated ability, it may be possible to optimize one's speed to maximize one's estimated ability. Since the association between speed and ability is assumed by the model to be linear, this is simply a matter of responding as fast as possible in case of a positive association between speed and ability and as slow as possible when the association is negative. Giving very fast responses will likely result in a strong reduction in the accuracy of the responses, meaning that this strategy will likely not be very effective in case of a positive association between speed and ability. However, if the association is negative, there is nothing stopping a well-informed respondent from giving slow responses to all of the items (to the extent that the time limit allows) to obtain a speed estimate that is as high as possible. While one could partially address this issue by not informing respondents of how their speed will affect their estimated ability, this would mean that the scoring rule cannot be communicated to respondents before or during the test, which may also be problematic. These issues, together with the general possibility of introducing bias discussed in the previous paragraph, make it that using the hierarchical model for improving the precision of ability estimates in high-stakes testing settings may be ill-advised.

In contrast, using the model in low-stakes testing settings may be more defensible, since the introduction of some degree of bias in the individual ability estimates could be considered acceptable there if it leads to a relevant increase in precision of those ability estimates. However, in these settings, the risk of heterogeneous response processes will be more prominent, since unlike in high-stakes testing, settings there likely will be a relevant subset of respondents who are providing fully or partially disengaged responses. If these "deviant" responses and respondents are not detected and excluded from the analysis, they will likely have a notable impact on the estimated correlation between speed and ability. Concretely, when many fast disengaged responses are present, the correlation between speed and ability will likely be more negative than it would be if those disengaged responses would be excluded from the analysis. While RTs may provide relevant information for determining disengaged responding (e.g., see Goldhammer et al. 2016; Nagy & Ullitzsch 2021), it is unlikely that any method will succeed in detecting disengaged responses with such a degree of accuracy that their presence no longer biases the estimate of the correlation between speed and ability.³ Consequently, there remains

³ It is important to stress here that the estimated ability effectively just summarizes the observed performance on the test, meaning that it only captures "effective ability." Since the effective ability of a respondent who is not fully engaged on the test will be lower than their actual ability level (i.e., the ability level that they would display when being fully engaged), these respondents should ideally be excluded from the analysis, since their estimated ability will be a (potentially highly) biased estimate of their actual ability level. With this in mind, they should therefore also not be included in the analysis when determining the correlation between speed and ability, which one

a risk of introducing notable bias in the estimate of ability for engaged respondents due to the failure to sufficiently exclude disengaged respondents and responses from the analysis.

Even if all disengaged responses and respondents can be eliminated from the analysis, the possibility of heterogeneous response processes remains. For example, if respondents differ in the extent to which they work heuristically versus algorithmically, this will affect both their expected RAs and RTs on the test. If one ignores these differences, one ends up with one overall association between speed and ability that aggregates the patterns found for the two styles of taking the test, which will likely not adequately represent the association between speed and ability in either of the subgroups, and hence potentially introduces bias in the ability estimates. While one would ideally study each subgroup separately, there may often be a variety of differences between persons in how they take the test, and adequately capturing this heterogeneity in the response processes will often not be feasible in practice. Thus, the possibility of heterogeneous response processes poses a challenge for the use of the hierarchical model, in both low- and in high-stakes testing settings.

An additional risk of bias lies in the assumption of the model that RTs are informative of ability (through speed) regardless of the accuracy of the response. Bolsinova and Tijmstra (2019) have found that in some settings, it may be plausible that only the RTs of correct responses are informative of ability. They proposed the possibility of separately measuring the speed with which correct and incorrect responses are given, respectively. Since the standard hierarchical model assumes that there is a single speed latent variable that explains the RTs and that RT (conditionally) does not depend on RA, it is not equipped to deal with this possibility. By combining the RTs of correct and incorrect responses together in a single latent speed variable, bias in the estimated ability may be introduced. This makes carefully checking whether there is indeed a single latent speed variable that explains the RTs important when using the hierarchical model in practice.

Finally, it may be relevant to point out the importance of distinguishing between θ and the construct of interest that the test is supposed to measure. While ideally the two overlap perfectly, even in the best of settings, it may be realistic to assume that there is some degree of construct-irrelevant variance present in the true value of θ of different respondents, meaning that θ is not a perfect proxy for the construct of interest even if there were no uncertainty in its estimates. For example, in addition to depending on ability, someone's test performance might be influenced by their experience in test taking or by their general reading skill. Such construct-irrelevant factors that influence θ could easily affect the expected RTs as well. Thus, there is the risk that the predictive power of speed is especially linked to this construct-irrelevant variance in θ , which would mean that using the hierarchical model instead of a standard IRT model would exacerbate the confounding of measurement that

wants to establish for the population of respondents who provided "normal" engaged responses to the items.

occurs. That is, the precision with which θ is estimated would increase, but at the cost of increasing the discrepancy between θ and the construct that the test was intended to measure. Using the hierarchical model therefore requires users to be confident that there is no issue with construct-irrelevant variance in θ , which may be difficult to establish in practice.

References

- Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, 82(4), 1126–1148. <https://doi.org/10.1007/s11336-016-9537-6>
- Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology*, 9, 1525. <https://doi.org/10.3389/fpsyg.2018.01525>
- Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics*, 41(2), 123–145. <https://doi.org/10.3102/1076998616631746>
- Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, 71(1), 13–38. <https://doi.org/10.1111/bmsp.12104>
- Bolsinova, M., & Tijmstra, J. (2019). Modeling differences between response times of correct and incorrect responses. *Psychometrika*, 84(4), 1018–1046. <https://doi.org/10.1007/s11336-019-09682-5>
- Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 70(2), 257–279. <https://doi.org/10.1111/bmsp.12076>
- Bolsinova, M., Tijmstra, J., Molenaar, D., & De Boeck, P. (2017). Conditional dependence between response time and accuracy: an overview of its possible sources and directions for distinguishing between them. *Frontiers in Psychology*, 8. Retrieved from <https://www.frontiersin.org/article/10.3389/fpsyg.2017.00202>
- Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289. <https://doi.org/10.2307/1165285>
- Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, 51(4), 540–553. <https://doi.org/10.1080/00273171.2016.1171128>
- Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement*, 13, 133–164. <https://doi.org/10.1080/15366367.2015.1100020>
- Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, 39(2), 108–119. <https://doi.org/10.1016/j.intell.2011.02.001>
- Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. *OECD Education Working Papers*, doi = 10.1787/5jltzfl6fhex2-en(133).
- Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives*, 15(3–4), 128–132. <https://doi.org/10.1080/15366367.2017.1411651>
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley. <https://doi.org/10.1037/13240-000>
- Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8, 150. <https://doi.org/10.3389/fnins.2014.00150>

- Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, 2(3), 261. <https://doi.org/10.1037/1082-989X.2.3.261>
- Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, 14(1), 54. <https://doi.org/10.1037/a0014877>
- Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica*, 43(3), 199–224. [https://doi.org/10.1016/0001-6918\(79\)90026-X](https://doi.org/10.1016/0001-6918(79)90026-X)
- Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika*, 76(3), 487–503. <https://doi.org/10.1007/s11336-011-9211-y>
- Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, 26(4), 29–37. <https://doi.org/10.1111/j.1745-3992.2007.00106.x>
- Meng, X. B., Tao, J., & Chang, H. H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, 52(1), 1–27. <https://doi.org/10.1111/jedm.12060>
- Nagy, G., & Ulitzsch, E. (2021). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement*, 1–35. <https://doi.org/10.1177/00131644211045351>
- Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error? An analysis of response programming. *Quarterly Journal of Experimental Psychology*, 29(4), 727–743. <https://doi.org/10.1080/14640747708400645>
- Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika*, 78(3), 538–544. <https://doi.org/10.1007/s11336-013-9324-6>
- Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, 54(2), 128–148.
- Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, 48, 37–50. <https://doi.org/10.1016/j.intell.2014.10.003>
- Shaw, A., Elizondo, F., & Wadlington, P. L. (2020). Reasoning, fast and slow: How noncognitive factors may alter the ability-speed relationship. *Intelligence*, 83, 101490. <https://doi.org/10.1016/j.intell.2020.101490>
- Tijmstra, J., & Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in psychology*, 9, 964. Retrieved from <https://doi.org/10.3389/fpsyg.2018.00964>
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioural Statistics*, 31(2), 181–204. <https://doi.org/10.3102/10769986031002181>
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308. <https://doi.org/10.1007/s11336-006-1478-z>
- van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46(3), 247–272. <https://doi.org/10.1111/j.1745-3984.2009.00080.x>
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365–384. <https://doi.org/10.1007/s11336-007-9046-8>
- van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). Irt parameter estimation with response times as collateral information. *Applied Psychological Measurement*, 34(5), 327–347. <https://doi.org/10.1177/0146621609349800>
- van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 195–210. <https://doi.org/10.1177/01466219922031329>

- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? what is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15(1), 22–29. <https://doi.org/10.1111/j.1745-3992.1996.tb00803.x>
- Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 68(3), 456–477. <https://doi.org/10.1111/bmsp.12054>
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145. <https://doi.org/10.1177/014662168400800201>
- Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2001). *Effects of local item dependence on the validity of IRT item, test, and ability statistics* (MCAT Monograph. 5). Association of Medical Colleges. Retrieved from <https://eric.ed.gov/?id=ED462426>

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 licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence 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

Computer-Adaptive Testing with Fewer Assumptions



Jules L. Ellis 

Abstract Two methods for computer-adaptive testing are being developed, based on monotone homogeneity. The first method uses the latent or observed item difficulties, and the second method is based on item-rest regressions. These methods can be used for the scaling of subjects and/or the selection of items. Seven combinations of nonparametric scaling and selection are studied and compared with a parametric method in various item banks. The nonparametric method based on item-rest regressions, for both scaling and selection, performs almost as good as the parametric method for computer-adaptive testing.

17.1 Computer-Adaptive Testing with Fewer Assumptions

This chapter will explore two methods of computer-adaptive testing based on monotone homogeneity. Computer-adaptive testing (CAT) is defined here as a test method where each subject may be exposed to a different combination of items from a given pool of test items. CAT is usually based on parametric item response theory (IRT) models such as the Rasch model and the 2-parameter logistic (2PL) model (e.g., van der Linden & Glas, 2010), and relatively few attempts have been made to base it on nonparametric IRT (e.g., Chiu & Chang, 2021). This chapter will use the nonparametric IRT model of *monotone homogeneity* developed by Mokken (1971; Sijtsma, 2005) for binary variables $X_j, j = 1, \dots, J$. This model has relatively few assumptions, namely:

- Unidimensionality: there is a real-valued latent variable (denoted as θ).
- Monotonicity: the item response functions (IRFs) $P(X_j = 1 | \theta)$ are increasing in θ .
- Conditional independence: the item scores $(X_j)_{j=1}^J$ are independent given θ .

J. L. Ellis (✉)

Behavioural Science Institute, Radboud University Nijmegen, Nijmegen, Netherlands
e-mail: jules.ellis@ru.nl

Authors on monotone homogeneity often use the term “monotonically nondecreasing” instead of “increasing,” but the conventional mathematical definition of both terms is the same (a function $f()$ is *increasing* if $x > y \Rightarrow f(x) \geq f(y)$ for all x, y in the domain of $f()$).

Developing a CAT method on the basis of monotone homogeneity may have two advantages. The first advantage is that the use of such a method is easily defensible in situations where monotone homogeneity holds while it is known that more specific models such as the 2PL model are violated. In such situations it would probably still be technically possible to base the CAT method on the 2PL model: the ordinary estimation algorithms used in CAT will produce estimates for the item and subject parameters even if the model is wrong. However, it would be hard to defend that these estimates are useful and to base decisions on them, if it is known that the underlying model is wrong. The second advantage of developing a CAT method based on monotone homogeneity is that it can be used to study the robustness of outcomes produced by a CAT method based on a parametric model. That is, in a situation where there is no clear violation of the 2PL model, one may still wonder whether similar outcomes will be obtained with less specific assumptions.

Two different problems of CAT can be distinguished. The first problem is that of *scaling the subjects*: how to assign scale values to subjects if they have responded to different items? The second problem is that of *selecting the items*: how to select the next item for a subject, given the previously administered items and the subject responses to them? In a CAT algorithm, these problems are usually addressed repeatedly and in sequence: (1) an item is selected, and the subject responds; (2) the scale value of the subject is estimated; (3) repeat.

The outline of the paper is as follows. The following sections will briefly review some examples of CAT that have been developed in nonparametric IRT and explicate the assumptions and objectives of this chapter. The subsequent section develops the nonparametric CAT methods that will be studied. After this, we describe the design of the simulation study and present the results of the simulation study. The last section is the discussion.

17.2 CAT in Nonparametric IRT

Nonparametric IRT has a longstanding relation with cognitive diagnosis modeling (Junker & Sijtsma, 2001; van der Ark et al., 2019). Chang et al. (2019) and Chiu and Chang (2021) discuss CAT with cognitive diagnosis models, where “the latent attribute profile of examinee i is a $K \times 1$ vector denoted as $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK})^T$. The latent space expanded by the K attributes thus contains 2^K latent proficiency classes, and the ultimate goal of CD is to assign examinees to the proficiency class to which they belong” (Chang et al., 2019, p. 545). Of special interest here is the nonparametric classification (NPC) method (Chiu & Douglas, 2013):

The NPC method classifies examinees by evaluating the distance between the observed and ideal item responses. Specifically, for the 2^K possible attribute profiles, the corresponding ideal response patterns are denoted as $\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(2^K)}$ where $\eta^{(m)} = (\eta_1^{(m)}, \dots, \eta_J^{(m)})$ for $m = 1, \dots, M = 2^K$. The examinee's attribute profile is then estimated by minimizing the distance between the observed and ideal item responses, $d(Y_i, \eta^{(m)})$, where $m = 1, \dots, M$. For binary data, a natural and frequently used distance measure is the Hamming distance, which simply counts the number of times that the entries in two vectors disagree. (Chang et al., 2019, p. 546; boldface in one formula omitted)

For example, if latent attribute 1 indicates whether the subject knows everything about Julius Caesar, and latent attribute 2 indicates whether the subject knows everything about Napoleon, and the first three items are questions about Julius Caesar and the fourth question is about Napoleon, and the fifth item requires knowledge of both Julius Caesar and Napoleon, then the ideal response pattern of someone with latent attribute profile $(1, 0)^T$ is $(1, 1, 1, 0, 0)$, and the ideal response pattern of someone with attribute profile $(0, 1)^T$ is $(0, 0, 0, 1, 0)$. These ideal response patterns are latent too; the observed response pattern on the five items can deviate from the ideal patterns because of mistakes and guessing.

The first CAT method that will be developed in this chapter is a special case of this, where the ideal response patterns form a Guttman scalogram. However, items that satisfy monotone homogeneity do not necessarily fit into this cognitive diagnosis model, and one can wonder how good this CAT technique performs if the items actually satisfy a 2PL model.

A second approach to CAT in nonparametric IRT is to estimate the IRFs (Xu & Douglas, 2006), based on, for example, the kernel smoothing (Ramsay, 1991). Douglas (1997) showed that the estimates of θ and the IRFs are consistent if both the number of subjects and the test length go to infinity. The second method that will be developed in this chapter will use the item-rest regressions instead of the kernel-smoothed IRF estimates. The item-rest regressions have the advantage that it is known that they have to be increasing under monotone homogeneity for binary items (Junker & Sijtsma, 2000), and this is true even for a finite number of items. Moreover, they are computationally very easy to obtain.

A third approach to CAT in nonparametric IRT is the use of a monotonic polynomial model (Falk & Feuerstahler, 2022). This model is based on the 2PL model, but each item can have additional parameters that accommodate deviations of the IRF from the logistic shape. This approach will not be explored in this chapter.

17.3 Assumptions and Objectives

17.3.1 Assumptions

- (1) There is a large pool of test items that can be selected. All items are binary: each answer on an item is either correct or incorrect. We will denote the score

on item j as X_j , with values 1 (subject gave a correct answer) or 0 (subject gave an incorrect answer).

- (2) The items have previously been administered in a large group of subjects, and from this the item means $\mathbb{E}(X_j)$, and the item-rest regressions $\mathbb{E}(X_j|R_{(j)})$, where $R_{(j)} := \sum_{k=1}^J X_k - X_j$, are known. In parametric examples, the latent item parameters are known too.
- (3) The items satisfy monotone homogeneity.
- (4) For each item the IRF has an infimum less than 0.50 and a supremum larger than 0.50.
- (5) The item selection in the CAT method depends only on the item parameters such as $\mathbb{E}(X_j)$ and $\mathbb{E}(X_j|R_{(j)})$, and not on the substantive domain of the items. (In practical applications, it is common practice to require that the items are balanced across certain domains. This will be avoided here because it obscures the differences between various methods.)
- (6) Each subject receives the same number of items. That is, the termination criterion is simply the number of items which the subject answered. (As a consequence, the CAT method will not directly render the test shorter for some subjects, but it may make the test more reliable because a smarter subset of items is used for each subject. And, knowing this, the test administrator may decide to decrease the general test length.)

17.3.2 Objectives

We seek nonparametric CAT methods based on monotone homogeneity. The monotone homogeneity model is more general than the 2PL model, and therefore the nonparametric methods based on monotone homogeneity should at least work in cases where the 2PL model holds, and preferably they should work in more situations. If the 2PL model holds, one can apply a parametric CAT method based on the 2PL model, and that will presumably be optimal; it is not expected that our nonparametric methods will outperform the parametric methods in that case. However the nonparametric method should produce outcomes that are very similar to the outcomes of a parametric method. Thus, even though the objective of this chapter is to develop a nonparametric method, without the 2PL model, its effectiveness will be studied both inside and outside the 2PL model.

If the accuracy of subject scale values (i.e., subject parameter estimates) is studied, this will be based on rank correlations rather than product-moment correlations. It is often argued that the subject parameters of the Rasch model have only ordinal meaning, and no one ever argued that the subject parameters of the monotone homogeneity model have more than ordinal meaning; therefore a rank correlation is more appropriate than a product-moment correlation.

With respect to the selection of items, it would be futile to require that parametric and nonparametric CAT methods select the same items, since even two parametric

methods will often not do so. What counts is how precise the ensuing subject estimates are. The error variance is not a good measure here because it depends on the scale, which is only ordinal. Therefore, the rank correlation between the true subject parameters and the estimated subject parameters will be used as a measure of *ordinal reliability*.

In sum, the nonparametric CAT method will be developed with the following objectives in mind:

- (1) If the test items satisfy the 2PL model, it will be required that the ordinal reliability of the new method is close to that of the parametric method.
- (2) If the test items do not satisfy the 2PL method, the outcomes of our new method should still be good; the ordinal reliability should still be high.

17.4 Description of the CAT Methods

17.4.1 *Nonparametric CAT Method 1: Latent and Observed Difficulty Matching*

The first new method is based on the idea that the items can be sorted on difficulties, and that the response pattern of a subject suggests the subject's location in relation to the difficulties. For example, if the items have increasing difficulty, and the subject has response pattern 11100, one would infer that the subject is positioned between the third and fourth item.

In the context of monotone homogeneity, item difficulty is often defined with an additional assumption, called *uniform relative difficulty* (Rosenbaum, 1987) or *invariant item ordering* (Sijtsma & Junker, 1996), which means that the IRFs do not intersect. However, in the present paper, we *do not assume invariant item ordering*. Let us therefore define a concept of item difficulty in monotone homogeneity *without invariant item ordering*. It was assumed that the items satisfy monotone homogeneity and that the IRFs increase from some value below 0.50 to some value above 0.50. Denote the IRFs as $P_j(\theta) = P(X_j = 1 | \Theta = \theta)$. If the IRF assumes the value 0.50 at exactly one value of θ , we can in theory define a difficulty parameter for the item as the value of θ for which $P_j(\theta) = 0.50$. If the IRF assumes the value 0.50 for multiple θ , we can still define a difficulty as the average value of θ for which $P_j(\theta) = 0.50$. If the IRF does not assume the value 0.50 anywhere, we can still define the item difficulty as the infimum of all θ for which $P_j(\theta) > 0.50$. Denote the difficulty of the j -th item as δ_j .

If a subject answers item j correctly, then the most likely inference from that observation is that $\theta > \delta_j$. If a subject answers item j incorrectly, then the most likely inference from that observation is that $\theta < \delta_j$. Therefore, the most likely answer patterns would be that of a Guttman scale; that is, if the items are ordered from easiest to most difficult by their δ_j , then the answer patterns are most likely of the form 1...111000...0. The most relevant property of θ would be its rank relative to

the δ_j . This will now be used as the scale on which estimates of θ are expressed. Similarly, the item difficulties will now be expressed as ranks, which means that the easiest item has $\delta_j = 1$ and the most difficult item has $\delta_j = J$ if the pool has J items. The index used to label the items is quite arbitrary; therefore we will simply assume that they are indexed in order of difficulty; thus $\delta_j = j$ for $j = 1, \dots, J$.

Suppose the subject answers only a few of the items, for example, the items with rank 3, 5 and 9 (from easy to difficult) and that the score pattern is 110. From this one would infer that the subject's θ is between 5 and 9, and we will then take the average of these values: $\hat{\theta} = (5 + 9) / 2$.

For a more general estimator, define new functions $g_j(\cdot)$ that will be called the *quasi-IRFs*. They are supposed to indicate the *modal* response on each item for each possible value of $\theta = 0, 1, 2, \dots, J$. Let them be defined as follows:

$$g_j(\theta) = 1 \text{ if } \theta \geq \delta_j$$

$$g_j(\theta) = 0 \text{ if } \theta < \delta_j$$

These quasi-IRFs are similar to the IRFs that one would have in a Guttman scalogram. They describe the ideal response patterns of a cognitive diagnosis model (Chang et al., 2019). Now, the reader may frown upon the usage of this model and believe that our theory is developing in the wrong direction, in that the model is getting more strict instead of less strict. However, it is not assumed that the g_i specify the IRFs; they are rather used as a simplification of the IRFs where the probabilities are rounded to the nearest integer value. That being said, we are now going to treat them as if they are the IRF anyways. Define the discrepancy function

$$\text{dis1}(\theta) = \sum_j (x_j - g_j(\theta))$$

where the summation runs across all items that were answered by the subject. The estimated value $\hat{\theta}_1$ of θ is now defined as the mean value of all θ for which $|\text{dis1}(\theta)|$ is minimal. This describes how the subject parameter, on the scale of the item ranks, can be estimated from an incomplete response pattern. This method utilizes the order of the item difficulties on the scale of the latent variable, θ . This estimation method will be called *latent difficulty matching*.

Note the similarity of this estimation method to the maximum likelihood estimate in the Rasch model. In the Rasch model, the derivative of the log likelihood of the response pattern would be $\sum_j (x_j - P_j(\theta))$ (e.g., Rose, 2010), and the maximum likelihood estimate would be obtained by setting it equal to 0 and solving in θ . This will be discussed further at the end of this section.

Although we want to avoid the assumption of invariant item ordering, one may wonder whether it is implicitly used in latent difficulty matching. One of the questions in this chapter, however, is whether latent difficulty matching can produce acceptable estimates even without invariant item ordering, that is, with intersecting IRFs. One reason for this question is this. In case of the 2PL model

with discrimination parameters α_j , if the subject answered all items, the weighted sum score $\sum_j \alpha_j x_j$ is known to be a sufficient statistic for θ (Birnbaum, 1968, p. 429). However, the unweighted sum score $\sum_j x_j$ often correlates very highly with a positively weighted sum score (Wilks, 1938, p. 27), and it has monotone likelihood ratio with θ (Grayson, 1988; Huynh, 1994; Unlü, 2008). Thus, the discrimination and difficulty parameters seem not very important for a rough, ordinal estimate of θ . In case of a CAT, where some items are unanswered, the difficulties may be more important, but it remains to be seen whether invariant item ordering is important.

The latent difficulty matching method can be applied only if one knows the rank order of the items by latent difficulty. The item means and regressions, known by assumption 2, do not provide sufficient information for the latent difficulties. Alternatively, the items can be ranked based on the *observed difficulties* $1 - \mathbb{E}(X_j)$ instead of the latent difficulties. More sophisticated estimates may be created, but the idea of difficulty matching is to keep things simple. If the difficulty ranks based on $1 - \mathbb{E}(X_j)$ are labeled δ_j^* , then the corresponding quasi-IRFs are $g_j^*(\theta) = \mathbf{1}[\theta \geq \delta_j^*]$, where $\mathbf{1}$ is the indicator function, and then the discrepancy is

$$\text{dis2}(\theta) = \sum_j (x_j - g_j^*(\theta)).$$

The estimated value $\hat{\theta}_2$ of θ is now defined as the mean value of all θ for which $|\text{dis2}(\theta)|$ is minimal. This estimation method will be called *observed difficulty matching*.

If the IRFs are non-intersecting (i.e., with invariant item ordering), the observed item difficulties δ_j^* will agree with the latent difficulties δ_j . However, we are not assuming invariant item ordering, and then the two difficulty rankings can be different. We will study how well observed difficulty matching performs without this assumption.

The next question is how items can be selected during the CAT administration. For this I suggest to pick, from the items that have not been used so far for the subject, the item with the difficulty rank closest to the estimated subject parameter. This method to select items will also be labelled latent difficulty matching or observed difficulty matching.

17.4.2 Nonparametric CAT Method 2: Item-Rest Regressions

In this second new method we use a second kind of quasi-IRFs, defined by the item-rest regressions. They are supposed to indicate the expected response on each item for each possible value of $r = 0, 1, \dots, J - 1$. Let them be defined as follows:

$$h_j(r) = \mathbb{E}(X_j | R_{(j)} = r)$$

Again, we know that these quasi-IRFs are generally not equal to the true IRFs, but they will be used as an approximation nonetheless. Indeed, other IRF estimates (e.g., Ramsay, 1991; Douglas, 1997 (both are used in the program TestGraf)) could be used too. The item-rest regressions have the advantage that they are easily computed, and they are known to be increasing under monotone homogeneity of binary items (Junker & Sijtsma, 2000), and consequently, isotonic regression can be used to smoothen sample estimates (although this was not used in the simulations of section 5 and 6).

Similar to the difficulty matching method, we define the discrepancy function

$$\text{dis3}(r) = \sum_j (x_j - h_j(r))$$

where the summation runs across all items that were answered by the subject. The estimated value $\hat{\theta}_3$ of θ should now be one of the values of r with $|\text{dis3}(r)|$ minimal. If only a few items have been administered, there can be many values of r with $\text{dis3}(r) = 0$, and in that case I suggest to take $\hat{\theta}_3$ close to the middle of the range of the rest scores, which is $(J - 1)/2$. The estimated value $\hat{\theta}_3$ of θ is thus defined as the value of r for which $|\text{dis3}(r)|$ is minimal and for which, given this restriction, $|(J - 1)/2 - r|$ is minimal. This describes how the subject parameter, on the scale of the item ranks, can be estimated from an incomplete response pattern.

Note that, in comparison with difficulty matching, the scale of the subject estimates is now changed from $J + 1$ possible levels of $\hat{\theta}_1$ to J possible levels of $\hat{\theta}_3$, because the rest-scores run from 0 to $J - 1$. Furthermore, the rest-score regressions are not entirely equivalent: if we start with a model in which θ is defined by, say, the 2PL model, then the posterior distribution of θ given $R_{(j)} = r$ depends on the item, j . These differences are ignored in the definition of $\hat{\theta}_3$. For a large number of items, these differences are presumably negligible. For example, if a subject has a rest score of 115 in a pool of 300 items, it does not matter very much which item was omitted in the rest score of 115; the posterior distributions of θ are probably very similar.

The next question is how items can be selected during the CAT administration. For this I suggest the item with the largest slope of the item-rest regression in a neighborhood of the estimated subject parameter. That is, for some value $\varepsilon > 0$, define a neighbourhood of $\hat{\theta}_3$ by the lower bound $x_1 := \max(\hat{\theta}_3 - \varepsilon, 0)$ and upper bound $x_2 := \min(\hat{\theta}_3 + \varepsilon, (J - 1))$, and define the local slope of each item j as

$$\text{slope}(j) := \frac{h(x_2) - h(x_1)}{x_2 - x_1}$$

Now pick, from the items that have not yet been used for the subject, the item with the largest slope. Since the estimated θ ranges between 0 and $J - 1$, it may be wise to let the value of ε depend on the number of items, but we do not yet have

general recommendations for it. In the simulations reported below, $\varepsilon = 10$ was being used, with $J = 162$ and $J = 347$, which seemed to work slightly better than $\varepsilon = 1$.

17.4.3 Parametric CAT Method

As specified in the objectives, the two new nonparametric methods will be compared with a well-established parametric CAT method for the 2PL model. For this parametric method, I choose the following specification. Subject parameters are estimated by Warm's (1989) weighted maximum likelihood, which will be denoted as $\hat{\theta}_4$. The items are selected on basis of maximum information given the current estimate of θ .

17.4.4 Comparison with Expected Moments and Maximum Likelihood Estimation of Subjects

If the IRFs are known, then estimation of θ by the expected moments method, applied to the total score, entails setting $\sum_j x_j = \sum_j P_j(\theta)$, and therefore solving

$$\sum_j (x_j - P_j(\theta)) = 0$$

That is, the expected moments estimate is a value of θ for which the average residual $x_j - P_j(\theta)$ equals zero. The nonparametric CAT methods to estimate θ , introduced here, are very similar to this, but replace the IRFs by discretized and/or estimated IRFs.

Next, consider maximum likelihood estimates. Under monotone homogeneity, if the derivative of P_j exists and is denoted as P'_j , the derivative of the log likelihood of the response pattern is

$$\sum_j (x_j - P_j(\theta)) \frac{P'_j(\theta)}{P_j(\theta)(1 - P_j(\theta))}$$

This can be viewed as a positively weighted average of residuals, with the remark that the weights $w_j(\theta) := P'_j(\theta) / [P_j(\theta)(1 - P_j(\theta))]$ depend on θ . The maximum likelihood estimate is a value of θ for which $\sum_j (x_j - P_j(\theta)) w_j(\theta) = 0$, that is, a weighted mean of the residuals equals 0.

A reviewer asked why we do not use weights similar to the maximum likelihood equation in the nonparametric CAT methods. For the discretized IRFs, this would be impossible because the denominator of $w_j(\theta)$ would become zero everywhere.

For the item-rest regression method, the IRFs P_j would be replaced by the item-rest regressions h_j in the computation of w_j , and the item-rest regression would become very influential if it is close to 0 or 1. This happens at extreme values of θ , where there are usually few observations, rendering the estimate of the item-rest regression unreliable. Therefore, I do not expect much benefits from including such weights.

In a 2PL model, $w_j(\theta) \equiv \alpha_j$, which depends on the item but not on θ . This might seem an advantage of the 2PL model in comparison to monotone homogeneity, but note that if there are few observations with extreme θ , one does not really know whether the 2PL model holds exactly for such θ , and then the increased precision of using the maximum likelihood equation with $w_j(\theta) \equiv \alpha_j$ is based on speculation.

17.5 Design of Simulation Study

Monte Carlo simulations were conducted with four item banks that will be described below. The first two item banks satisfy the 2PL model; the last two item banks satisfy monotone homogeneity but not the 2PL model. The scaling of subjects and the selection of items are not necessarily based on the same information: for example, the subjects can be estimated by item-rest regression, while the items are selected by observed difficulty matching. Additionally, the possibility to select the items randomly without replacement was studied as a kind of baseline. Table 17.1 lists the combinations of methods that were studied. Each combination will be referred to as a ‘‘CAT.’’

In each item bank and for each combination of methods, the CAT was simulated for 1000 subjects with a standard normal distribution for θ . The test length was fixed a priori at values of 20, 40, or 80 items. After the simulation, the rank correlation between the true values of θ and the estimates was computed.

Table 17.1 List of methods used in the simulations

CAT	Scaling of subject	Selection of items
1	Latent difficulty matching	Random
2	Latent difficulty matching	Latent difficulty matching
3	Observed difficulty matching	Random
4	Observed difficulty matching	Observed difficulty matching
5	Item-rest regression	Random
6	Item-rest regression	Observed difficulty matching
7	Item-rest regression	Item-rest regression
8	Parametric	Parametric
9	Parametric + observed difficulty matching ^a	Parametric

^aIn CAT 9, the item selection and the intermediary subject estimates are based on the parametric method, but in the last phase, after the subject completed the last item, the subject parameter is estimated by the observed difficulty matching method. This method is studied in order to distinguish the effects of item selection method and subject estimation method

Table 17.2 Descriptive statistics of item parameters of item bank 2

	Discrimination α_j	Difficulty δ_j
Mean	1.131	-0.646
SD	0.354	1.643
Minimum	0.383	-4.909
Quartile 1	0.874	-1.943
Median	1.151	-0.550
Quartile 3	1.360	0.513
Maximum	3.003	3.971

17.5.1 Description of Item Banks

17.5.1.1 Item Bank 1 (2PL, Artificial Parameters)

This item bank was simulated with a 2PL model, $P(X_j = 1|\theta) = (1 + \exp(\alpha_j(\theta - \delta_j)))^{-1}$. The item bank contained 162 items with discrimination parameter α_j equal to 1 or 2, and difficulty parameters δ_j evenly spaced between -2 and 2, where each value of the difficulty parameter occurred once with $\alpha_j = 1$ and once with $\alpha_j = 2$. These were created by setting for odd j , $1 \leq j \leq 162$: $\alpha_j = 1$, $\alpha_{j+1} = 2$, $\delta_j = -2 + (j - 1) * 0.05$, $\delta_{j+1} = \delta_j$.

17.5.1.2 Item Bank 2 (2PL, Realistic Parameters)

This item bank was simulated with a 2PL model, $P(X_j = 1|\theta) = (1 + \exp(\alpha_j(\theta - \delta_j)))^{-1}$. The item bank contained 347 items. The item parameters were obtained from a real item bank with Arithmetic items that was used in 2021 as test in Dutch primary schools. In the Dutch school system, children are required to make a test at the end of primary school that serves as the basis for advisements for secondary school. Several commercial test developers may provide their own version of an end test, of which the quality is evaluated by a government committee. The estimated item parameters of one such commercial test developer were used. These parameters were estimated on the basis of data obtained from 4223 children, who made the test as a parametric CAT. Each item was answered by at least 800 children. Descriptive statistics of the item parameters are given in Table 17.2.

17.5.1.3 Item Bank 3 (Non-2PL)

This item bank was simulated with IRFs that have a 2PL shape for low and high values of θ , but a plateau with constant value 0.50 in a middle region of θ . The width of the middle region is modeled by a third item parameter γ_j . The IRFs are

thus defined by these rules:

$$P(X_j = 1 | \theta) = (1 + \exp(\alpha_j(\theta - \delta_j)))^{-1} \text{ if } \theta \leq \delta_j.$$

$$P(X_j = 1 | \theta) = 0.50 \text{ if } \delta_j < \theta < \delta_j + \gamma_j$$

$$P(X_j = 1 | \theta) = (1 + \exp(\alpha_j(\theta - \delta_j - \gamma_j)))^{-1} \text{ if } \theta > \delta_j + \gamma_j$$

In this item bank, 347 items were used with the same α_j and δ_j parameters as in item bank 2, and with $\gamma_j = 0.50$ for all items.

17.5.1.4 Item Bank 4 (Non-2PL)

This item bank was simulated with IRFs that are the mean of two 2PL-shaped IRFs with the same discrimination parameter but different difficulty parameters. The difference between the two difficulty parameters is modeled by a third item parameter γ_j . The IRFs are thus defined by these rules:

$$P(X_j = 1 | \theta) = 0.5(1 + \exp(\alpha_j(\theta - \delta_j)))^{-1} + 0.5(1 + \exp(\alpha_j(\theta - \delta_j - \gamma_j)))^{-1}$$

In this item bank, 347 items were used with the same α_j and δ_j parameters as in item bank 2 and with $\gamma_j = 0.50$ for all items.

17.5.2 Estimation of Observed Means and Item-Rest Regressions

In each item bank, the items were ranked based on $1 - \mathbb{E}(X_j)$, estimated by 1000,000 draws of standard normal θ . The empirical rest score regressions, $\mathbb{E}(X_j | R_{(j)})$ were estimated based in 1000,000 draws. Groups of $R_{(j)}$ with less than 100 subjects were deleted. The function h_j was now defined as $h_j(r) = \mathbb{E}(X_j | R_{(j)} = r)$ on the values of r where the latter quantity was defined. For values of r for which $\mathbb{E}(X_j | R_{(j)} = r)$ was unknown, h_j was set to 0 for low values of r (i.e., values r for which $\mathbb{E}(X_j | R_{(j)} = r')$ was unknown for all $r' < r$) and set to 1 for high values of r (i.e., values r for which $\mathbb{E}(X_j | R_{(j)} = r')$ was unknown for all $r' > r$), and interpolated on the middle values of r .

17.6 Results

Table 17.3 shows the ordinal reliabilities of the various CATs in the four item banks for varying test lengths.

CAT 7, where both subject scaling and item selection are based on the item-rest regressions, has the highest ordinal reliability among the nonparametric CATs (1–7) in all four item banks. The ordinal reliabilities of this CAT are very close to the

Table 17.3 Simulation results

Item bank	CAT	Scaling of subject	Selection of items	J = 20	J = 40	J = 80
1	1	Latent difficulty matching	Random	0.878	0.948	0.972
	2	Latent difficulty matching	Latent difficulty matching	0.847	0.927	0.980
	3	Observed difficulty matching	Random	0.884	0.946	0.970
	4	Observed difficulty matching	Observed difficulty matching	0.854	0.932	0.980
	5	Item-rest regression	Random	0.909	0.954	0.975
	6	Item-rest regression	Observed difficulty matching	0.936	0.967	0.983
	7	Item-rest regression	Item-rest regression	0.963	0.979	0.986
	8	Parametric	Parametric	0.971	0.984	0.988
	9	Parametric	Parametric + observed difficulty matching	0.949	0.967	0.979
2	1	Latent difficulty matching	Random	0.816	0.899	0.946
	2	Latent difficulty matching	Latent difficulty matching	0.578	0.699	0.830
	3	Observed difficulty matching	Random	0.809	0.895	0.944
	4	Observed difficulty matching	Observed difficulty matching	0.695	0.784	0.888
	5	Item-rest regression	Random	0.764	0.877	0.932
	6	Item-rest regression	Observed difficulty matching	0.818	0.915	0.967
	7	Item-rest regression	Item-rest regression	0.943	0.968	0.983
	8	Parametric	Parametric	0.954	0.972	0.984
	9	Parametric	Parametric + observed difficulty matching	0.933	0.958	0.972
3	1	Latent difficulty matching	Random	0.780	0.881	0.934
	2	Latent difficulty matching	Latent difficulty matching	0.495	0.620	0.771
	3	Observed difficulty matching	Random	0.777	0.882	0.935
	4	Observed difficulty matching	Observed difficulty matching	0.602	0.692	0.820

(continued)

Table 17.3 (continued)

Item bank	CAT	Scaling of subject	Selection of items	$J = 20$	$J = 40$	$J = 80$
	5	Item-rest regression	Random	0.744	0.869	0.925
	6	Item-rest regression	Observed difficulty matching	0.790	0.861	0.939
	7	Item-rest regression	Item-rest regression	0.911	0.953	0.978
	8	Parametric	Parametric	0.919	0.952	0.976
	9	Parametric	Parametric + observed difficulty matching	0.896	0.939	0.966
4	1	Latent difficulty matching	Random	0.809	0.898	0.948
	2	Latent difficulty matching	Latent difficulty matching	0.592	0.706	0.845
	3	Observed difficulty matching	Random	0.807	0.897	0.945
	4	Observed difficulty matching	Observed difficulty matching	0.697	0.775	0.887
	5	Item-rest regression	Random	0.763	0.877	0.933
	6	Item-rest regression	Observed difficulty matching	0.841	0.919	0.969
	7	Item-rest regression	Item-rest regression	0.938	0.966	0.981
	8	Parametric	Parametric	0.954	0.973	0.984
	9	Parametric	Parametric + Observed difficulty matching	0.935	0.961	0.975

ordinal reliabilities obtained by the parametric CAT (CAT 8), with the difference at most 0.017 and the ratio at least 0.983.

CAT 9, which utilizes parametric item selection but nonparametric subject estimation in the final phase, has also high ordinal reliabilities that are close to the fully parametric method. This suggests that once the items are selected, the nonparametric estimation of subjects by observed difficulty matching is almost as good as the parametric estimation.

CATs 5 and 6, based only partially on item-rest regressions, have considerable lower ordinal reliability than CAT 7 in most item banks.

CATs 1–4, based on difficulty matching, consistently have the lowest ordinal reliability, and substantially lower than CAT 7. In these cases, the CATs with randomly selected items outperform the CATs with items that were selected by difficulty matching – which is counterintuitive.

17.7 Discussion

The results are encouraging. A nonparametric CAT can be based on item-rest regressions, and in the cases studied here, it proved to be almost as reliable as a parametric CAT based on the 2PL model.

An almost equally high performance was obtained in the combination method, where the items are selected by the parametric method and the final subject estimates are obtained with observed difficulty matching. The fact that the combination procedure performs so well can be used in robustness studies to corroborate subject estimates after a parametric CAT has been administered. It might also be used to explain to lay people the basic idea of subject scaling in a CAT.

Future research may study how various smoothing methods for the item-rest regressions affect the performance, how to pick the optimal value of epsilon (the radius of the neighborhood on which the slope is determined), and the effect of sample fluctuations in the prior calibration phase in which item-rest regressions are estimated. Item-rest regressions based on smaller samples might perform much worse than observed here with $N = 10^6$. In an explorative simulation with ten independent calibrations of $N = 10,000$ each, we found an average correlation of 0.900 with standard error 0.009 for CAT 7 with maximum test length 20 and item bank 2 (this correlation was 0.943 in Table 17.3 with $N = 10^6$). The item-rest regressions in the simulations with $N = 10,000$ were smoothed by deleting groups of $R_{(j)}$ with group size less than 20, and the result may improve further if the smoothing is based on isotonic regression instead (e.g., Tijmstra et al., 2012). Thus, the nonparametric method based on item-rest regressions may still work for a realistic size of N , but obviously it will break down if N becomes too small. A fair comparison would also study the effect of estimation error of item parameters in parametric methods. Note that if the calibration sample size is so small that the IRFs cannot be estimated accurately, assuming that the IRFs are logistic seems premature.

For practical applications it should also be studied how content constraints can be added and which termination criteria can be used to create CATs with flexible test length. A more elaborate study can investigate how the item-rest regression method performs in comparison to other methods for CAT in nonparametric IRT, notably cognitive diagnosis CAT (Chiu & Chang, 2021), kernel smoothing of IRFs (Xu & Douglas, 2006), and monotonic polynomial models (Falk & Feuerstahler, 2022). Finally, as pointed out by a reviewer, our estimation equation can be rewritten as $\sum_j x_j - \sum_j \mathbb{E}(x_j|\theta) = 0$, and this may work for polytomous items too. This requires further investigation.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison Wesley. <https://ci.nii.ac.jp/naid/10011544105/>
- Chang, Y.-P., Chiu, C.-Y., & Tsai, R.-C. (2019). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*, 43(7), 543–561. <https://doi.org/10.1177/0146621618813113>
- Chiu, C.-Y., & Chang, Y.-P. (2021). Advances in CD-CAT: The general nonparametric item selection method. *Psychometrika*, 86(4), 1039–1057. <https://doi.org/10.1007/s11336-021-09792-z>
- Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30(2), 225–250. <https://doi.org/10.1007/s00357-013-9132-9>
- Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, 62(1), 7–28. <https://doi.org/10.1007/bf02294778>
- Falk, C. F., & Feuerstahler, L. M. (2022). On the performance of semi- and nonparametric item response functions in computer adaptive tests. *Educational and Psychological Measurement*, 82(1), 57–75. <https://doi.org/10.1177/00131644211014261>
- Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, 53(3), 383–392. <https://doi.org/10.1007/BF02294219>
- Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent bernoulli random variables. *Psychometrika*, 59(1), 77–79. <https://doi.org/10.1007/BF02294266>
- Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, 24(1), 65–81. <https://doi.org/10.1177/01466216000241004>
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272. <https://doi.org/10.1177/01466210122032064>
- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56(4), 611–630. <https://doi.org/10.1007/bf02294494>
- Rose, N. (2010). Maximum likelihood and Bayes modal ability estimation in two-parametric IRT models: derivations and implementation. <https://www.kompetenztest.de/download/szbf-2010-rose-n-maximum-likelihood.pdf>
- Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika*, 52(1), 217–233. <https://doi.org/10.1007/BF02294236>
- Sijtsma, K. (2005). Nonparametric item response theory models. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 875–882). Elsevier. <https://doi.org/10.1016/B0-12-369398-5/00459-X>

- Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, 49(1), 79–105. <https://doi.org/10.1111/j.2044-8317.1996.tb01076.x>
- Tijmstra, J., Hessen, D. J., van der Heijden, P. G. M., & Sijtsma, K. (2012). Testing manifest monotonicity using order-constrained statistical inference. *Psychometrika*, 78(1), 83–97. <https://doi.org/10.1007/s11336-012-9297-x>
- Unlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 61(1), 179–187. <https://doi.org/10.1348/000711007X173391>
- Van der Ark, L. A., Rossi, G., & Sijtsma, K. (2019). Nonparametric item response theory and Mokken scale analysis, with relations to latent class models and cognitive diagnostic models. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 21–45). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_2
- Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450. <https://doi.org/10.1007/bf02294627>
- Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 3(1), 23–40. <https://doi-org.ru.idm.oclc.org/10.1007/BF02287917>
- Xu, X., & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, 71(1), 121–137. <https://doi.org/10.1007/s11336-003-1154-5>

Chapter 18

Validity Indices for Interpreting Informant Discrepancies in ADHD Assessment



Judith M. Conijn, Mengdi Chen, Hanneke van Ewijk,
and L. Andries van der Ark

Abstract In ADHD assessment, discrepancies between informants' test scores complicate decision-making with respect to treatment and educational adaptations. These informant discrepancies may be due to meaningful differences such as variations in assessed behavior across settings (i.e., school vs. home) but may also be due to response biases or unsystematic error. We propose using response-pattern-based validity indices for studying which is the most plausible of the two explanations. These indices detect invalid test scores through identifying extreme, repetitive, or inconsistent response patterns. To illustrate the use of validity indices for interpreting informant discrepancies, we used a subset of data ($N = 431$) from the self-report and parent-report version of Conners ADHD rating scales collected in the NeuroIMAGE study. Pairs of self-report and parent-report scores were classified as either discrepant or non-discrepant, and validity indices were applied to classify self- and parent-report scores as suspect (i.e., potentially invalid due to rater effects) or not. Finally, we demonstrate how information from validity indices can be taken into account in diagnostic decision-making.

J. M. Conijn (✉)

Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands

Kohnstamm Institute, University of Amsterdam, Amsterdam, The Netherlands

e-mail: jconijn@kohnstamm.uva.nl

M. Chen · L. A. van der Ark

Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands

e-mail: m.chen@uva.nl; L.A.vanderArk@uva.nl

H. van Ewijk

Curium-LUMC, Department of Child and Adolescent Psychiatry, Leiden University Medical Center, Leiden, The Netherlands

e-mail: h.van_ewijk.cur@curium.nl

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

L. A. van der Ark et al. (eds.), *Essays on Contemporary Psychometrics*,

Methodology of Educational Measurement and Assessment,

https://doi.org/10.1007/978-3-031-10370-4_18

18.1 Introduction

Attention-deficit hyperactivity disorder (ADHD) is defined as a disorder with symptoms present in two or more settings (such as home, school, or work) with functional impairment in at least one of these settings. In an educational setting, a student's inattentiveness symptoms of ADHD, for example, may lead to difficulties in following instructions, organizing, and completing assignments. For diagnosing ADHD, data are commonly collected from multiple sources, for example, self-reports and informant reports from teachers, parents, or peers (Achenbach, 2006; De Los Reyes, 2011; Dirks et al., 2012). Typically, informants assess student's inattentiveness symptoms of ADHD on a multi-item checklist, and the item scores are then added to a total symptom score. Collecting data from multiple informants is crucial for decision-making considering that one of the criteria for an ADHD diagnosis is the presence of symptoms across multiple contexts.

An important problem for researchers, school psychologists, and clinicians working with data on symptom severity from multiple informants is that different informants may not provide the same score, which is known as informant discrepancy (De Los Reyes et al., 2019; De Los Reyes & Kazdin, 2005; Martel et al., 2017). For example, Nelson and Lovett (2019) found that students' self-reports on ADHD symptoms showed only moderate correlations ($r = .3 - .5$) with parent reports. An analysis of response validity showed that a large part of the discrepancies could be explained by students and parents reporting inconsistent symptoms or students overreporting symptoms to obtain the academic benefits of an ADHD diagnosis (e.g., extended testing time for taking exams).

In scientific research on ADHD, informant discrepancies may lead to biased research conclusions because informants may provide different or even contradictory information. However, in educational practice informant discrepancies are particularly problematic. They complicate individual treatment decisions and decisions regarding academic benefits, as clinicians or school psychologists need to base their diagnostic decisions on discrepant information (De Los Reyes & Kazdin, 2005; De Los Reyes et al., 2019; Smith, 2007). Nevertheless, as Martel et al. (2017) already noticed, there have been conducted only few studies that propose and evaluate solutions for handling discrepancies in practical settings.

In the current study, we propose that the use of validity indices can solve part of the difficulties due to informant discrepancies in individual decision-making. This approach can be applied to the assessment of ADHD—as demonstrated in the example below. However, the approach is general and can be applied to all sorts of contexts in which educational—and also clinical—decision-making is based on multiple informants providing data on the same set of symptoms. Before explaining our approach, we first discuss several existing methods for dealing with informant discrepancies, which also depend on the specific cause of the discrepancy.

18.1.1 Approaches and Explanations for Informant Discrepancies

A first strategy to mitigate the effect of informant discrepancies is to combine the item scores of multiple informants into a single item score. For example, there are approaches that minimize the Type I error (an item score equals 1, if and only if all informants rated the symptom as present) and approaches that minimize the Type II error (an item score equals 1, if at least one informant rated the symptom as present). Alternatively, Martel et al. (2017) used the average across respondents of the total symptom score. In an empirical comparison of these three approaches, Martel et al. (2015) found the averaging method to be most valid for determining the total number of children's ADHD symptoms and taking diagnostic ADHD decisions. A second strategy is to select the score of a specific informant that has been shown to produce the most useful or valid score in empirical research. This informant is called the *optimal informant* (Bird et al., 1992; Kraemer et al., 2003). Research has shown that the optimal informant depends on the characteristics of the informants and constructs measured. For example, Smith (2007) developed a model that combines three characteristics (age, setting, and construct type) to help in deciding the optimal informant in the assessment of child psychopathology based on previous research results (e.g., incremental validity coefficients).

A third approach for dealing with discrepancies is the general framework of De Los Reyes et al. (2013, 2019): the *operations triad model*. In this approach, the chosen method for handling informant discrepancies depends on the main explanation for the informant discrepancies. The first possible explanation is the presence of *meaningful variation* across reports due to differences in observed behavior. For example, if a child's behavior varies between school and home, a possible discrepancy between parents and teacher's informant reports may be due to true differences in observed behavior (Dirks et al., 2012; De Los Reyes et al., 2009; Martel et al., 2017). Secondly, *error sources* may explain informant discrepancies. Error sources may be due to poor psychometric properties of a given informant measure (i.e., poor reliability or validity of a questionnaire) in a specific sample, or methodological issues such as differences in item content between the two versions of a questionnaire (De Los Reyes et al., 2013).

Based on the two main explanations for informant discrepancies, the operations triad model outlines how researchers may deal with informant discrepancy, for example, as evidenced by an unexpected low correlation between scores of different informants that should assess the same construct. In short, it is argued that a lack of correspondence between scores of different informants *may* be meaningful and, in that case, the scores reported by all informants should be used as separate variables in a study. However, before concluding that informant discrepancies are meaningful and adjusting the analyses accordingly, researchers should first systematically rule out that informant discrepancy is due to error sources. To this end, researchers should examine reliability indices for each of the informant measures and validity indices such as correlations representing convergent or divergent validity. If the

psychometric properties of both informant measures are favorable in the population of interest and no other methodological issues can be identified (e.g., differences in the item wording of the two informant measures) that can explain informant discrepancies, support for meaningful differences between informant scores is found (De Los Reyes et al., 2013, 2019). Instead of being averaged or combined into a single variable, the scores should then be treated as separate yet dependent observations in further analyses, for example, extensions of the generalized linear model that account for correlations among dependent variables.

18.1.2 Rater Effects as an Additional Explanation for Informant Discrepancy

The operations triad model could be regarded as the more sophisticated and advanced approach for dealing with informant discrepancy since it takes the source of the discrepancy into account. Moreover, the psychometric analyses and methodological checks proposed in the operations triad model seem to be an appropriate approach for assessing whether error sources can explain informant discrepancy in group-level research. On the other hand, these analyses do not provide a sufficient means to assess whether the response pattern of a specific individual respondent is reliable and valid. Even when a measure shows adequate psychometric properties in the population of interest, *rater effects* may compromise the accuracy of the test score of an individual respondent (e.g., Martel et al., 2015).

Rater effects can be manifested as response biases or unsystematic error in a response pattern. Response biases, for example, include social desirability bias, overreporting, underreporting, or malingering (Martel et al., 2017; Smith, 2007). Unsystematic error can be due to careless responding when respondents face complicated questions, lack of motivation, a too lengthy test, or environmental distractions (Meade & Craig, 2012). Low cognitive or reading skills and idiosyncratic interpretation of item content may also result in unsystematic error (e.g., Meijer et al., 2008; Smith et al., 2010). If rater effects are substantial, they result in invalid test scores and are likely to generate informant discrepancies (e.g., Martel et al., 2015).

Rater effects are mainly problematic for individual decision-making. In group-level research, the different types of rater effects that occur across respondents partly cancel each other out and therefore may not strongly affect group-level statistics such as group means (e.g., Conijn et al., 2019). In contrast, in individual decision-making, rater effects due to for example overreporting or careless responding can result in severely biased decisions. In this study, we therefore propose that when *individual* decision-making is hampered by informant discrepancy, an additional psychometric approach (i.e., next to the standard validity and reliability checks using group-level data) is needed to separate informant discrepancy due to error sources from informant discrepancy due to meaningful differences. The goal of the current study is therefore to describe an additional approach that is useful for

interpreting informant discrepancy in an individual decision context. The first part of this paper provides a general description of the validity-index approach. The second part provides an illustration of the method by means of an application to a dataset of self-reported and parent-reported ADHD symptoms.

18.2 The Validity-Index Approach

The validity-index approach can be applied in cases that individual decision-making is hampered by a discrepancy between test scores of different informants regarding the same construct, for example, a college student reporting a clinical level of symptom severity on a self-report questionnaire and his parents reporting a non-clinical level of symptom severity on the same questionnaire. This discrepancy interferes with decision-making regarding whether or not the student has a right to receive educational adaptations. The specific definition of informant discrepancy will depend on the particular decision-making context and is actually not relevant for how to use the validity-index approach. Informant discrepancy may, for example, also be defined as a specific size of a test score difference between two informants that is clinically relevant. Alternatively, it may be a score difference that is significant given the estimated reliability of the measures.

For a given pair of discrepant test scores, the main goals of the validity-index approach are to categorize response patterns of each of the informants as likely invalid due to rater effects or likely valid and to take that categorization into account in individual decision-making. In the validity-index approach, the categorization of response patterns is done using response-pattern-based validity indices. Response-pattern-based validity indices are computed using the observed response pattern and quantify the degree that an individual's pattern of responses is aberrant (Meade & Craig, 2012; Niessen et al., 2016; Wanders et al., 2017). Response patterns may be aberrant because they include inconsistent item scores, excessive repetition (i.e., long sequences of the same item score), or many extreme item scores. Such aberrant response patterns may indicate careless responding, malingering, or another type of response bias or unsystematic error. The corresponding test scores are therefore likely to be invalid. Examples of validity indices are the item response theory (IRT)-based I_z person-fit index (Drasgow et al., 1985; Emons, 2008) and the long-string index of repetitive responding (Johnson, 2005).

To classify response patterns as potentially invalid, a cutoff value for the validity index is needed. Such a cutoff value is commonly based on the distribution of the indices in a large sample of representative respondents. For example, Meijer et al. (2016) suggested using a cutoff value corresponding to the 5% most extreme observed values on a validity index. Several other more sophisticated methods based on IRT-based simulations have also been proposed (see Conijn et al., 2019, for an overview). The choice for the specific classification method depends on properties of the questionnaire data (e.g., the use of IRT-based simulations requires that an IRT model fits the data) but also on the stakes of the decisions that are taken using

the validity-index method (i.e., higher stakes require more sophisticated methods). If liberal cutoff values are used, auxiliary information such as interviews with the respondents should also be collected to confirm invalid responding after the initial round of screening (Meijer et al., 2016).

In short, application of the validity-index approach to a discrepant set of informant scores includes the following steps:

1. Assess which validity indices are appropriate for the data. This depends on the properties of the informant and/or self-report measures such as the response scale of the items and the wording of the items (Conijn et al., 2019). In general, it is useful to select different types of validity indices representing different types of response biases and unsystematic error.
2. Establish appropriate cutoffs for the validity indices that are used to classify respondents as having a “suspect” (i.e., likely invalid) test score.
3. Compute the selected validity indices for each of the response patterns, and apply the cutoff values to classify informant scores as suspect or not suspect. This results in three different categories: discrepancy that is likely due to rater effects of one of the informants, discrepancy that is likely due to rater effects of multiple informants, or discrepancy that is likely due to meaningful differences. In the latter category, none of the informants produced a response pattern that is suspect of problematic rater effects.
4. Account for the validity information in individual decision-making, for example, in a *diagnostic algorithm* (i.e., the approach for combining the different pieces of information on symptom severity into a diagnosis).

The exact procedure in Step 4 depends on the specific diagnostic algorithm or decision-making rule used in the study, but we provide some basic suggestions: If the results show that one of the multiple informant scores is suspect, the other informant score(s) should gain most weight in (clinical) decision-making. If each of the informant scores is flagged as suspect, retesting should be done, or clinical interviews should be conducted. Finally, if neither of the informant scores is classified as suspect, the discrepancy is likely due to meaningful variation in informant scores such as variation in the subjects’ observed behavior. Such explanations can then be further explored.

As a final note, the validity-index approach presupposes that reliable and valid informant and self-report measures are used in the assessment and there are no important methodological issues in the research design that can explain informant discrepancies. Given these assumptions, the discrepancies between scores of different informants can only be attributable to either rater effects or to true and meaningful differences in observed behavior. The validity-index approach could therefore also be regarded as the next step after the group-level analyses that were proposed in the operations triad model have shown to support the validity and reliability of the measures used and have not identified other methodological issues.

18.2.1 Illustrative Example

We illustrate the proposed validity-index approach using data from 441 self-reports and corresponding parent proxy-reports on Conners ADHD Rating Scales from the NeuroIMAGE study (Von Rhein et al., 2015). Conners Rating Scales include scales for self-report, parent-report, and teacher report of ADHD symptoms. In educational contexts, the scales are commonly used to assess whether ADHD causes functional impairment at school (e.g., Purpura & Lonigan, 2009).

We used two selected subscales among Conners ADHD Rating Scales, the Inattentive symptoms scale and the Hyperactive-Impulsive symptoms scale, to study discrepancies between self-reports and parents' informant-reports. First, we present the dataset, the measures, and descriptive statistics on invalid responding in the NeuroIMAGE dataset. Next, we illustrate how the validity-index approach can be used in individual decision-making by accounting for validity information in the diagnostic algorithm of the NeuroIMAGE study.

18.3 Method

18.3.1 Participants

We used secondary data collected between 2009 and 2012 in the NeuroIMAGE study (Von Rhein et al., 2015). The NeuroIMAGE study includes ADHD ratings for 1978 studied participants. Ratings came from different informants: parents and either teachers (if participants were still in school and less than 18 years old) or self-reports (if participants were either at least 18 years old or no longer in school). For the current study, we used only a small subset of NeuroIMAGE data. Specifically, our inclusion criteria were the following: The rated participant had (a) available self-report data on Conners Adult ADHD Rating Scales—Self-Report: Long Version (CAARS- S: L), (b) available informant report data from one of their parents on Conners Parent Rating Scale (CPRS), and (c) no more than 5% missing data on both the CAARS and the CPRS. Cases with more than 5% missing data were excluded because missing values may interfere with the comparability of validity-index values across respondents with and without missing item scores.

Our subsample consisted of $N = 431$ participant pairs, including adolescents as well as young adults and one corresponding parent providing the informant data about their child. Among the participants being rated, 56% were male and the mean age was 20.3 ($SD = 2.4$; range = 15 – 30). Following the diagnostic algorithm used in the NeuroIMAGE study, participants were diagnosed as affected with ADHD (37%; $n = 160$), as unaffected with ADHD (45%; $n = 196$), or they were labeled as “subthreshold” because they did not meet the criteria for either being affected or unaffected (13%; $n = 57$). For 18 (4%) of the rated participants, the diagnosis was missing.

Table 18.1 Description of the Conners ADHD rating scales

Rating scale	Factor-analysis-derived subscales	DSM-IV symptom-count subscales (<i>T</i> scores ≥ 65 indicate a positive screening outcome)
CAARS (66 items; self-report)		
	Inattention-Memory problems (12 items)	Inattentive symptoms (9 items)
	Hyperactivity-Restlessness (12 items)	Hyperactive- Impulsive symptoms (9 items)
	Impulsivity/Emotional lability (12 items)	Total symptoms (sum of the two specific symptom scales)
	Problems with Self-Concept (6 items)	
CPRS (80 items; parent-report)		
	Cognitive problems (10 items)	Inattentive symptoms (9 items)
	Oppositional (12 items)	Hyperactive- Impulsive symptoms (9 items)
	Hyperactivity-Impulsivity (9 items)	Total symptoms (sum of the two specific symptom scales)
	Anxious-shy (8 items)	
	Perfectionism (7 items)	
	Social problems (5 items)	
	Psychosomatic problems (6 items)	

Note. Post hoc validity indices were computed using the respondents' complete response pattern on the CAARS/CPRS; informant discrepancy was computed only for the DSM Inattentive symptoms subscale and the DSM Hyperactive-Impulsive symptoms subscale

18.3.2 Conners ADHD Rating Scales: CAARS and CPRS

Participants completed the CAARS (Conners et al., 1999), and one of their parents completed the CPRS (Conners et al., 1998) about the participant. The CAARS and the CPRS are designed to screen for ADHD but also include items to assess a range of externalizing (e.g., aggression) and internalizing (e.g., anxiety) symptomatology (Table 18.1). The CAARS is a 66-item self-report measure for adults; 42 of the items belong to one of the 4 factor-analysis-derived subscales (Conners et al., 1999). The CPRS is an 80-item standardized behavior rating scale designed to be completed by parents; 56 items are included in the seven factor-analysis-derived subscales (Conners et al., 1998). Respondents indicated on a 4-point response scale how frequently they (or their child) experience(s) the symptom described in the item: *seldom/never* (0), *sometimes* (1), *quite often* (2), or *very often* (3). For all the items, higher scores indicate more severe psychopathology.

Next to the factor-analysis-derived subscales, both the CAARS and the CPRS include subscales that provide a count of Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) ADHD symptoms: the Inattentive symptoms subscale (DSM Inattentive; nine items), the Hyperactive-Impulsive symptoms subscale

(DSM Hyperactive-Impulsive; nine items), and the Total symptoms subscale (i.e., the inattentive and hyperactive-impulsive symptoms combined; 18 items). The three subscales are used to screen for the diagnosis of the inattentive ADHD subtype, hyperactive-impulsive ADHD subtype, and the combined ADHD subtype (i.e., both inattentive and hyperactive-impulsive), respectively. Test scores on the three subscales are transformed to T scores ($M = 50$; $SD = 10$).

18.3.3 Informant Discrepancy

We investigated informant discrepancies on both the DSM Inattentive subscale and the DSM Hyperactive-Impulsive subscale. We defined pairs of informants to have a discrepant set of scores on the DSM Inattentive subscale and the DSM Hyperactive-Impulsive subscale, respectively, when they showed a 15-point difference on the T score scale (i.e., corresponding to 1.5 SD difference on the T score scale). A difference of 1.5 SD seemed both clinically relevant and unlikely to occur due to chance but at the same time did result in a substantial proportion of discrepant cases for further analysis. Our chosen definition of informant discrepancy was pragmatic considering that it served an illustrative example and is not part of the validity-index approach. Previous studies that investigated informant discrepancy on other psychopathology scales defined informant discrepancy using cutoffs of 1 or 2 SD score-point difference between two informants (e.g., Conijn et al., 2018; Dorz et al., 2004).

The DSM Inattentive subscale and the DSM Hyperactive-Impulsive were chosen to define informant discrepancy because they are used for diagnostic decisions rather than for descriptive purposes. Furthermore, the value of Cronbach's alpha equaled 0.84 for the DSM Inattentiveness subscale and 0.88 for the DSM Hyperactive-Impulsive subscale, suggesting that test-score reliability was no cause of informant discrepancy. Also, differences in item content between the two versions of the subscales were unlikely to cause substantial informant discrepancy. Item wording of the CPRS and CAARS version of these subscales is not exactly equal, but item content is similar: for the DSM Inattentive subscale, each CPRS item can be paired with a CAARS item, and for the DSM Hyperactive-Impulsive subscale, seven out of nine items can be paired.

18.3.4 Validity Indices

Although we investigated informant discrepancies on the DSM symptom scales, we used the respondents' response pattern on the complete set of CPRS/CAARS items to determine whether their response pattern was suspect. The reason for doing so was that the use of more item scores renders the validity indices more reliable (i.e., resulting in higher sensitivity and specificity; Conijn et al., 2019). Moreover, we expected invalid response behavior to be consistent across different subscales because items from different subscales were presented in a mixed order.

Table 18.2 Description of the validity indices and their cutoff values

Validity index	Short description	Cutoff method	Cutoff value ^a	
			CPRS	CAARS
l_z^p person-fit index	Detects inconsistent responding with respect to the graded response IRT model. First computed for separate subscales, next averaged into a single index	95th percentile value	0.51	0.89
Mahalanobis distance (MD)	Detects multivariate outliers; quantifies the distance between an observed response pattern and the remaining response patterns in the sample	95th percentile value	223.0	123.9
Long-string index (L_{max})	Detects repetitive responding by the maximum length of strings of consecutive identical answers. Computed for separate response options	2.5% most extreme observed values		
0 score			64	24
1 score			6	8
2 score			4	5
3 score			4	4
Overreporting index	Detects overreporting using the CAARS Infrequency Index (CII) for the CAARS and the percentage of 3 scores for the CPRS	Cutoff value CII	N/A	20
		98th percentile value	24%	29%
Underreporting index	Detects underreporting by the percentage 0 scores	95th percentile value	96%	82%

Note.^aResponse patterns with index values larger than the specified cutoff value are classified as suspect

Five validity indices were applied to the CPRS and CAARS data resulting in an index-specific validity classification for each informant. If at least one of the index-specific classifications was suspect, the response pattern was overall classified as “suspect.” Table 18.2 summarizes indices and methods for establishing cutoff values. Next, we describe the validity indices in detail.

Mahalanobis Distance The Mahalanobis distance (MD) is a multivariate outlier statistic (e.g., Johnson & Wichern, 2008). When used as a validity index, MD

quantifies the distance between an observed response pattern and the remaining response patterns in the sample while taking the inter-item correlation matrix into account. Several studies have found the MD index appropriate for detecting random responding (Curran, 2016; DeSimone, et al., 2015). We used the 95th percentile value as a cutoff value for classifying response patterns as suspect. To account for missing values, we using the R package **modi** (Hulliger, 2018) to compute an adapted MD in which the missing values are ignored and a correction factor is applied to MD based on the number of observed values.

l_{zm}^p Person-Fit Index The parametric l_z^p person-fit index (Drasgow et al., 1985) is the standardized log-likelihood of a polytomous response pattern given the estimated unidimensional item response theory (IRT) model. The l_z^p index mainly detects random or inconsistent responding but can also pick up other types of invalid response styles if these lead to deviations with respect to the IRT model (Emons, 2008).¹ We computed l_z^p with respect to the graded-response IRT model, using the R package **Perfit** (Tendeiro et al., 2016). We took the negative of the l_z^p index so that a higher value of the index was indicative of a more inconsistent response pattern. As the l_z^p index should be computed for unidimensional subscales, we first computed the index for each factor-derived subscale (Table 18.1) and next averaged the subscale l_z^p values into an overall multiscale validity index l_{zm}^p (e.g., Conijn et al., 2014; Niessen et al., 2016). Items that did not belong to a CPRS/CAARS factor-derived subscale were excluded from computing l_z^p . We used the 95th percentile value as a cutoff value for classifying response patterns as suspect (Meijer et al., 2016). Missing item scores were imputed by the default non-parametric single imputation method in the R package **Perfit**. We regarded the single imputation method sufficient for our purpose because we were only interested in quantifying validity. Imputed item scores are not informative about the validity of the response pattern, even if a superior multiple imputation method would have been used (e.g., Van Ginkel et al., 2007).

Long-String Index: L_{max} Long-string indices count the length of strings of consecutive identical answers to detect repetitive careless responding (Johnson, 2005; Kam & Chan, 2018; Meade & Craig, 2012). L_{max} equals the maximum length of a string of consecutive identical answers and has been found to have higher power to detect careless responding compared to other long-string indices (Meade & Craig, 2012; Niessen et al., 2016). We computed L_{max} for each of the four score options separately, resulting in four response-option-specific L_{max} values for each respondent. Different cutoff values for the different response options were used to take into account the skewed item-score distribution (Conijn et al., 2019; Johnson, 2005). The four cutoff scores were based on the 2.5% most extreme observed L_{max}

¹ We used the l_z^p index instead of other possible person-fit indices (e.g., the Guttman person-fit indices) because previous research suggests that l_z^p (1) has relatively high power to detect careless responding, (2) is least confounded with the substantive trait measured, and (3) is least strongly correlated with the MD index (Conijn et al., 2019).

values for that response option. Missing values were ignored in the computation of L_{\max} , meaning that the length of strings of identical answers was computed after excluding the missing values from the response pattern.

Overreporting Index To detect overreporting and malingering on the CAARS, we used the CAARS Infrequency Index (CII) index (Suhr et al., 2011). This index equals the sum score of 12 items that were endorsed infrequently in the Suhr et al. (2011) study. Scores greater than 20 are classified as suspect (Suhr et al., 2011). For the CAARS data, the CII index classified 2% ($n = 9$) response patterns as suspect of overreporting. No overreporting indices have been developed for the CPRS. We therefore used the percentage of 3 scores as an indicator of possible overreporting. Based on the 2% suspect cases that were identified using the CII in the CAARS data, we used a cutoff value equal to the 98th percentile value in the CPRS data.² Missing values were ignored in the computation of the percentage value.

Underreporting Index To detect underreporting, no specific indices have been proposed for the CAARS or CPRS. We therefore used the percentage of 0-scores as an indicator of possible underreporting, using the 95th percentile value as a cutoff value for classifying response patterns as suspect. Missing values were ignored in the computation of the percentage value, meaning that percentages were computed after excluding the missing values from the response pattern.

18.3.5 Statistical Analyses

In preliminary analyses, we used the full dataset, including the data of informant pairs that were not discrepant. In this full dataset, we inspected descriptive statistics for informant discrepancy and the validity indices, and we analyzed the relationship between informant discrepancy and validity-index classification. Informant pairs were categorized into four categories: (a) non-discrepant and no suspect response pattern(s), (b) discrepant and no suspect response pattern(s), (c) discrepant and suspect response pattern(s), (d) non-discrepant and suspect response pattern(s).

The main analyses illustrate how the validity-index approach can be used to improve diagnostic decision-making in the presence of informant discrepancy. For the two categories of informant pairs having discrepant response patterns [i.e., categories (b) and (c)], we describe the outcomes of the application of validity indices and discuss possible rater effects that may have caused the informant discrepancy. We also illustrate how the validity-index information can be taken into account in a diagnostic algorithm to establish ADHD diagnosis. To this end, we applied both the original “basic” diagnostic algorithm used in the NeuroIMAGE study and an adapted diagnostic algorithm that takes into account the validity-

² In the CAARS data, we assessed the correspondence between classifications based on the CII index and the percentage of 3 scores. We found moderate agreement (Cohens Kappa: 0.58) between the CII classification and the alternative classification based on the percentage of 3 scores.

index information to data of discrepant informant pairs. The difference between the basic and adapted algorithm was that the latter excluded the rating scale data (either CAARS or CPRS or both) that was classified as suspect by at least one of the validity indices. We provide simple descriptive statistics to summarize the differences between diagnostic outcomes from the basic and the adapted diagnostic algorithm. The diagnostic algorithms are explained in detail in the following section.

18.3.6 Diagnostic Algorithm

The diagnostic algorithm in the NeuroIMAGE study (Von Rhein et al., 2015) for establishing an ADHD diagnosis was based on a combination of data from diagnostic interviews and data from CAARS and/or CPRS. Specifically, a semi-structured clinical interview, the Schedule for Affective Disorders and Schizophrenia—present and lifetime version (K-SADS; Kaufman et al., 1997)—was used. The K-SADS was conducted with participants and, to provide an informant interview, with one of their parents. From these two K-SADS interviews and the impression of the interviewer, a K-SADS ADHD symptom count was derived.

Basic Algorithm First, a combined symptom count was calculated by counting a symptom as present if it was scored as present in the K-SADS symptom count or the self-report CAARS data. Based on this symptom count, the CAARS and the CPRS, criteria for being considered affected with ADHD were the following:³ (a) combined symptom count ≥ 5 symptoms of inattentive or hyperactive/impulsive behavior, and (b) T score ≥ 63 on at least one of the CPRS or CAARS ADHD symptom scales (filled in about a period without medication): the Inattentive symptoms subscale, the Hyperactive-Impulsive Symptoms subscale and the Total symptoms subscale. Criteria for being considered unaffected with ADHD were: (a) ≤ 2 symptoms derived from the combined symptom count and (b) T scores < 63 on each of the CPRS or CAARS ADHD symptom scales. Table 18.3 (left-hand column) shows a summary. Participants who did not meet the requirements for the affected or unaffected status were classified as “subthreshold.” Additional to this basic algorithm, in the NeuroIMAGE study, cases with inconsistent information were evaluated by a team of experts to derive a consensus (best-estimate) diagnosis. We did not use these (clinical) adjustments in the current study because we were interested in the difference between a basic algorithm without and with adjustment for validity information.

Adapted Algorithm To assess the impact of taking into account validity information in diagnostic decisions, we defined an adapted algorithm that took into account

³ The other basic requirements for diagnosis were as follows: an age of onset before 12, meeting the DSM criteria for pervasiveness and impairment, and symptoms are not better accounted for by another disorder.

Table 18.3 Basic and diagnostic algorithm used in this study

Algorithm		
	Basic	Adapted
Diagnosis		
ADHD	$S \geq 5 \cap \{T_I \geq 63 \cup T_H \geq 63 \cup T_T \geq 63\}$	$S^* \geq 5 \cap \{T_I^* \geq 63 \cup T_H^* \geq 63 \cup T_T^* \geq 63\}$
No ADHD	$S \leq 2 \cap \{T_I < 63 \cap T_H < 63 \cap T_T < 63\}$	$S^* \leq 2 \cap \{T_I^* < 63 \cap T_H^* < 63 \cap T_T^* < 63\}$

Note. S = Combined symptom count of inattentive or hyperactive/impulsive behaviour based on the K-SADS data and the CAARS. S^* = combined symptom count that excludes the CAARS subscale data if classified suspect. $T_I = T$ score on the CAARS or CPRS Inattentive symptoms subscale; $T_H = T$ score on the CAARS or CPRS Hyperactive/impulsive symptoms subscale; $T_T = T$ score on the CAARS or CPRS Total symptoms subscale. $T_{H/I/T}^* = T$ score on the CAARS or CPRS symptoms subscales but excluding T scores of informants whose data is classified suspect

the validity-index information. In contrast to the original algorithm, the adapted algorithm completely excluded the item-response patterns (either CAARS or CPRS or both) that were classified as suspect by at least one of the validity indices. This also applied to the combined symptom count. So, if the CAARS data was found to be suspect, an “adapted” combined symptom count was computed using only the K-SADS interview data. If the CPRS data was classified as suspect, the combined symptom count was not affected (i.e., since it was not taken into account in the total symptom count in the basic algorithm either). The criteria for being affected with ADHD were equal to those of the basic algorithm, but now the CPRS/CAARS data of informants that generated a suspect response pattern were excluded from the algorithm (Table 18.3, right-hand column).

18.4 Results

18.4.1 Preliminary Analyses

Both the DSM Hyperactive-Impulsive subscale score and the Inattentive subscale score of the CAARS (self-report) correlated .62 with the corresponding CPRS (parent-report) subscale scores. The percentage of informant pairs that was classified discrepant was 22.3% for the DSM Hyperactive-Impulsive subscale and 22.6% for the DSM Inattentive subscale. For 8.5% of the informant pairs, the test scores were discrepant for both the Inattentiveness and the Hyperactive-Impulse subscales.

For the CPRS and CAARS, Table 18.4 shows the percentages of response patterns classified as suspect by each separate validity index and the percentages of response patterns classified as suspect based on at least one of the indices (Table 18.4, last row). The percentage of response patterns classified as suspect by at least one validity index was 19.7% for CAARS and 20.0% for CPRS. These percentages are very similar as for each validity index, the cutoff scores for CAARS

Table 18.4 Percentage of response patterns classified as suspect by each separate index and by at least one of the indices

Validity classification as suspect	% Classified suspect	
	CAARS self-report	CPRS parent-report
Index-specific classification		
I_z^p	5.1	5.1
MD	5.1	5.1
L_{\max}	8.4	9.5
Overrep. index	2.1	2.3
Underrep. index	5.8	5.1
Overall classified as suspect (≥ 1 index-specific)	19.7	20.0

Note. The overall classification as suspect is used in the subsequent and main analyses

Table 18.5 Cross tabulation of validity and informant discrepancy

Validity classification ^a	Informant discrepancy ^b		Total
	Non-discrepant	Discrepant	
DSM Hyperactive-Impulsive subscale			
Not suspect	237 (0.81)	54 (0.19)	291 (1.00)
Suspect	100 (0.71)	40 (0.29)	140 (1.00)
Total	337 (0.78)	94 (0.22)	431 (1.00)
DSM Inattentive symptoms subscale			
Not suspect	233 (0.80)	58 (0.20)	291 (1.00)
Suspect	100 (0.71)	40 (0.29)	140 (1.00)
Total	333 (0.77)	98 (0.23)	431 (1.00)

Note. ^aSuspect when at least one response pattern within an informant pair classified as suspect by at least one validity index; ^bDiscrepant when the difference in T scores is >15

and CPRS were determined using the same percentile score (95 for MD, I_{zm}^p , and underreporting index; 97.5 for L_{\max} , and 98 for overreporting index).

Table 18.5 shows a cross-tabulation of informant discrepancy and the presence of at least one suspect response pattern within that informant pair. Informant pairs with at least one suspect response pattern (either the CAARS pattern, the CPRS pattern, or both; Table 18.4) are more likely to have a discrepant set of scores on the DSM Inattentive subscale ($X^2 = 4.02$, $df = 1$, $p = .045$) and the DSM Hyperactive-Impulsive subscale ($X^2 = 5.56$, $df = 1$, $p = .018$). Phi coefficients equal 0.09 for the DSM Inattentive subscale and 0.11 for the DSM Hyperactive-Impulsive subscale, indicating a (very) weak relationship between the discrepancy and the “suspect” classifications (Cohen, 1988). The positive relationships between the discrepancy and the “suspect” classifications suggest that part of the discrepancy is due to rater effects. However, the weak relationships also suggest that most discrepancies in the data may be due to meaningful test-score differences between participants and their parents.

18.4.2 *The Validity-Index Approach in Diagnostic Decision-Making*

We applied the adapted algorithm to the data of 169 informant pairs with discrepant test scores on either the Hyperactivity/Impulsivity subscale or the Inattentive subscale. Hundred and eight pairs did not have a suspect response pattern, and the adapted algorithm therefore produced the same diagnostic ADHD decision as the basic algorithm. Ten informant pairs had suspect response patterns for both the CAARS and the CPRS. For these informant pairs, the adapted algorithm used only the K-SADS data, which resulted in a different diagnosis (“subthreshold” instead of “affected”) for one of the ten pairs. For 8 of the 29 informant pairs with a suspect CPRS score but a valid CAARS score, we found that the adapted diagnostic algorithm produced a different diagnosis (either “unaffected” instead of “threshold” or “threshold” instead of “affected”). For 10 out of the 22 informant pairs with a suspect CAARS score and a valid CPRS score, the adapted diagnostic algorithm produced a different diagnosis (either “unaffected” instead of “threshold” or “threshold” instead of “affected”). The diagnosis from the adapted algorithm was always more conservative than the diagnosis of the original algorithm because it used a compensatory rule, defining symptoms to be present if at least one informant rates the symptom to be present. As an illustration, Table 18.6 describes three parent-child pairs who had a discrepant set of test scores and who had one of the two Conners rating-scale response patterns classified as suspect. In the following, we describe each case in more detail.

Participant A The rated participant was a 22-year-old female. The self-report scores on the Hyperactive-Impulsive and Inattentive CAARS subscales were approximately 30 *T* score units higher than for the corresponding CPRS parent-report. The self-report pattern was classified as suspect based on four different validity indices: I_{zm}^p , CII/overreporting index, L_{max} , and MD. The CII index classified the response pattern as suspect of overreporting. Consistently, we found that for 48% of the 66 items the respondent selected the response option “very frequent.” Furthermore, the L_{max} index suggested a repetitive response bias, and the I_z^p values for separate subscales indicated severe response inconsistency on the CAARS Inattention/Memory problems subscale and milder response inconsistency on the other three subscales (for a description of the subscales, see Table 18.1). An example of response inconsistency is that the respondent indicated to “always plan things in advance,” but also indicated to often be “disorganized” and “dependent on others organizing my life and helping focusing on details.”

Based on three CAARS *T* scores ≥ 82 and a combined symptom count of 9 for both Hyperactive-Impulsive symptoms and Inattentive symptoms, the original diagnostic algorithm resulted in an “affected with ADHD” diagnosis. In the adapted algorithm, the CAARS subscale scores were excluded from the algorithm, and none of the CPRS subscale scores were high enough to meet the criteria for the affected status. So, the adapted algorithm resulted in a “subthreshold ADHD” diagnosis.

Table 18.6 Data of three illustrative cases and outcomes of a basis diagnostic algorithm and an adapted algorithm that takes into account validity information

Construct/subscale	Informant	Variable	Participant A Suspect self-report	Participant B Suspect self-report	Participant C Suspect parent-report
Hyperactive- Impulsive symptoms	Parent	CPRS <i>T</i> score	53	43	90
	Self	CAARS <i>T</i> score	82	44	61
	Parent + self	K-SADS symptom count	6	1	6
	Parent + self	S (S*)	9 (6)	2 (1)	8 (8)
Inattentive symptoms	Parent	CPRS <i>T</i> score	56	45	90
	Self	CAARS <i>T</i> score	85	74	63
	Parent + self	K-SADS symptom count	8	4	7
	Parent + self	S (S*)	9 (8)	6 (4)	8 (8)
Total symptoms	Parent	CPRS <i>T</i> score	55	44	90
	Self	CAARS <i>T</i> score	88	61	64
Outcome of diagnostic algorithm	Basic algorithm		Affected	Affected	Affected
	Adapted algorithm		Subthreshold	Subthreshold	Affected

Note. The basic diagnostic algorithm comes from the NeuroIMAGE study and does not take into account validity information. The adapted algorithm takes into account validity information. Only the bold data is used in the adapted algorithm for a specific case. *S* = Combined symptom count of inattentive or hyperactive/impulsive behavior based on the K-SADS data and the CAARS. *S** = Combined symptom count that excludes the CAARS subscale data if classified suspect

Given this outcome, a clinician may decide to interview the participant and the parent about the aberrant response pattern and the discrepancy in their scores to understand which of the two diagnostic outcomes is most appropriate.

Participant B The rated participant was a 20-year-old male. There was no informant discrepancy for the Hyperactive-Impulsive subscale, but the *T* score for the CAARS self-report Inattentive subscale was 29 points higher than the corresponding CPRS parent-report. The self-report response pattern was classified suspect based on the l_{zm}^p statistic. The respondent showed inconsistent responding particularly on the Inattention/Memory problems and Impulsivity/Emotional lability subscales. Examples of inconsistent responding were that the respondent indicated that he quite

often “explodes,” “has bad tempers,” and “is easily irritated,” but on the other hand also indicated that his temper is not at all unpredictable.

Based on the self-report CAARS T score of 74 and a combined symptom count of 6, the original diagnostic algorithm resulted in an “affected with ADHD” diagnosis. Using the adapted algorithm, both the total symptom count and the CPRS T scores were not high enough for meeting the affected status, but they were also not low enough to meet the unaffected status. The adapted algorithm therefore resulted in a “subthreshold ADHD” diagnosis. Similarly as for Participant A, the clinician may decide to collect additional (interview) data to make a final diagnostic decision.

Participant C The rated participant was a 20-year-old female. The T scores for each three parent-report CPRS subscales were at the maximum value (i.e., T score = 90), while the self-report CAARS scores were at least 26 score points lower. The parent report was classified suspect based on L_{max} and the overreporting index. The L_{max} index indicated several long strings of 3 scores (i.e., the response option “very frequent”) with a maximum length of seven. The overreporting index indicated that 50% of all item scores were in the “very frequent” category, which is particularly notable because the CPRS measures a wide range of different symptoms (Table 18.1). So, the validity information suggested that the parent exaggerated the symptoms and used a repetitive response style.

Based on three CPRS T scores of 90, CAARS T scores of 63 (Inattentive symptoms) and 64 (Total symptoms), and a combined symptom count of eight for each specific type of symptoms, the participant was diagnosed as “affected with ADHD.” The adapted algorithm resulted in the same diagnosis: although the suspect CPRS scores were excluded from the algorithm, the CAARS T scores on the total symptoms subscale were high enough for meeting the affected ADHD status.

18.5 Discussion

We illustrated how post hoc validity indices can be used for studying whether informant discrepancies may be due to invalid test scores caused by rater effects of individual respondents. In the example dataset from the NeuroIMAGE study, we found a weak relationship between informant discrepancies and rater effects, suggesting that most discrepancies are due to meaningful test score differences between informants. Furthermore, we provided an example on how information about the validity of an individual’s response pattern can be taken into account into a diagnostic algorithm. Here we found that for 19 (11%) out of 169 informant pairs with discrepant test scores, an adapted diagnostic algorithm led to a different diagnostic outcome.

Nelson and Lovett (2019) studied invalid responding on the CAARS ADHD symptom subscales for the same pairs of informants as we studied (young adults and their parents). Compared to their results, the level of informant discrepancy

was low in our illustrative dataset. We found cross-informant correlations of .64, whereas Nelson and Lovett (2019) found correlations between parent and self-report scores on exactly the same CAARS subscales ranging from .37 to .42. The difference in results is likely due to the context in which data were collected. In the Nelson and Lovett study, data were collected in a naturalistic environment where symptom exaggeration could lead to benefits for students, whereas data collection in the NeuroIMAGE study was primarily done for scientific research purposes. In both our study and the Nelson and Lovett study, the diagnostic outcome was generally more conservative for informant pairs with no evidence of an invalid report. Both studies thus suggest that invalid responding is likely to lead to an overestimation of the prevalence of ADHD.

The strengths of our study are the following. Firstly, we demonstrated how a widely applicable, general approach to check validity of a response pattern can be used to study and take into account informant discrepancies. This is valuable because most psychopathology scales do not have built-in validity scales. Furthermore, our illustrative example is particularly useful as we used an ADHD dataset. In the context of diagnostic ADHD assessment, informant discrepancies as well as systematic rater effects (e.g., overreporting and underreporting) are important and often studied topics (Luderer et al., 2019; Martel et al., 2015; Sibley et al., 2012; Walls et al., 2017). Second, unsystematic error (e.g., inconsistent and random responding) can also be expected to be problematic given the core symptom of ADHD such as inattentiveness and impulsivity (Nelson & Lovett, 2019; Raiker et al., 2012; Sibley et al., 2019). Third, this study provides one the first application examples of response-pattern based validity indices to individual decision-making. A previous study showed how the IRT-based I_z^p person-fit index can be used to detect inconsistent responding on a depression measure and provide psychiatrists with valuable information for treatment and diagnostic decision-making (Wanders et al., 2017). Our study adds to Wanders et al. (2017) by using multiple validity indices to evaluate a response pattern and by applying validity indices to solve informant discrepancies.

There are also several important limitations to this study. Because we used secondary data, we could not show to the full potential of the validity-index approach for handling informant discrepancies. The validity-index information can be used optimally if the response patterns identified as inconsistent, repetitive, or extreme can be discussed with the patient or the other informant (Nelson & Lovett, 2019; Wanders et al., 2017). Such qualitative interview data can provide additional evidence for the invalidity of a test score, information on the reasons underlying the unexpected response pattern, or may lead to the conclusion that the validity-index approach resulted in a false-positive classification. A second limitation is that the percentile-based method that we used for establishing cutoff values for the validity indices was relatively simple. In previous research, for some of the validity indices, more sophisticated methods have been suggested (Conijn et al., 2019; De la Torre & Deng, 2008). For example, one approach for establishing a cutoff-value is to generate multiple (e.g., 20) “clean” datasets based on an IRT model estimated in the empirical questionnaire dataset and use the average 95th

percentile value of a specific validity index across the simulated datasets as a cutoff value (Conijn et al., 2019). We did not use this approach in the current study because it requires the questionnaire items to belong to a unidimensional (sub)scale. Third, a limitation specific to using data from the NeuroIMAGE study was a mismatch between respondents and questionnaires used in that study. The CAARS, a rating scale designed for adults, was administered to children (12% of the sample was 15–17 years old). Likewise, the CPRS, an informant scale for parents to rate children up to 17 years old, was applied in the study although 88% of the rated children were actually young adults.

Finally, we provide several suggestions for future research. First, future research could conduct a pilot implementation of the validity-index method in a clinical or educational practice where ADHD assessment is complicated by informant discrepancies. By following and interviewing the clinicians working with the method, the study could evaluate the practical value of using the validity indices. Second, if a dataset is available where different informants (e.g., children or parents) completed exactly the same questionnaire, validity indices can be used to investigate which informant shows most rater effects (Conijn et al., 2020). This type of study may point to one type of informant being more accurate in responding than the other. Third, investigating the relationship between respondents' ADHD symptom severity and validity indices can provide insight into the extent that ADHD symptoms interfere with valid self-report responding (Sibley et al., 2019).

References

- Achenbach, T. M. (2006). As others see us clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, 15(2), 94–98. <https://doi.org/10.1111/j.0963-7214.2006.00414.x>
- Bird, H. R., Gould, M. S., & Staghezza, B. (1992). Aggregating data from multiple informants in child psychiatry epidemiological research. *Journal of the American Academy of Child & Adolescent Psychiatry*, 31(1), 78–85. <https://doi.org/10.1097/00004583-199201000-00012>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- Conijn, J. M., Emons, W. H., & Sijtsma, K. (2014). Statistic l_2 -based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement*, 38(2), 122–136. <https://doi.org/10.1177/0146621613497568>
- Conijn, J. M., Emons, W. H., Page, B. F., Sijtsma, K., Van der Does, W., Carlier, I. V., & Giltay, E. J. (2018). Response inconsistency of patient-reported symptoms as a predictor of discrepancy between patient and clinician-reported depression severity. *Assessment*, 25(7), 917–928. <https://doi.org/10.1177/1073191116666949>
- Conijn, J. M., Franz, G., Emons, W. H. M., de Beurs, E., & Carlier, I. V. E. (2019). The assessment and impact of careless responding in routine outcome monitoring within mental health care. *Multivariate Behavioral Research*, 19(4), 1–19. <https://doi.org/10.1080/00273171.2018.1563520>
- Conijn, J. M., Smits, N., & Hartman, E. E. (2020). Determining at what age children provide sound self-reports: An illustration of the validity-index approach. *Assessment*, 27(7), 1604–1618. <https://doi.org/10.1177/1073191119832655>

- Conners, C. K., Sitarenios, G., Parker, J. D., & Epstein, J. N. (1998). The revised Conners' Parent Rating Scale (CPRS-R): Factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology*, 26(4), 257–268. <https://doi.org/10.1023/A:1022602400621>
- Conners, C. K., Erhardt, D., & Sparrow, E. (1999). *Conners' adult ADHD rating scales: Technical manual*. Multi-Health Systems.
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. <https://doi.org/10.1016/j.jesp.2015.07.006>
- De la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, 45(2), 159–177. <https://doi.org/10.1111/j.1745-3984.2008.00058.x>
- De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, 40(1), 1–9. <https://doi.org/10.1080/15374416.2011.533405>
- De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, 131(4), 483–509. <https://doi.org/10.1037/0033-2909.131.4.483>
- De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology*, 37(5), 637–652. <https://doi.org/10.1007/s10802-009-9307-3>
- De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, 9, 123–149. <https://doi.org/10.1146/annurev-clinpsy-050212-18561>
- De Los Reyes, A., Cook, C. R., Gresham, F. M., Makol, B. A., & Wang, M. (2019). Informant discrepancies in assessments of psychosocial functioning in school-based services and research: Review and directions for future research. *Journal of School Psychology*, 74, 74–89. <https://doi.org/10.1016/j.jsp.2019.05.005>
- DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, 36(2), 171–181. <https://doi.org/10.1002/job.1962>
- Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual research review: Embracing not erasing contextual variability in children's behavior—theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry*, 53(5), 558–574. <https://doi.org/10.1111/j.1469-7610.2012.02537.x>
- Dorz, S., Borgherini, G., Conforti, D., Scarso, C., & Magni, G. (2004). Comparison of self-rated and clinician-rated measures of depressive symptoms: A naturalistic study. *Psychology and Psychotherapy: Theory, research and practice*, 77(3), 353–361. <https://doi.org/10.1348/1476083041839349>
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67–86. <https://doi.org/10.1111/j.2044-8317.1985.tb00817.x>
- Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, 32(3), 224–247. <https://doi.org/10.1177/0146621607302479>
- Hulliger, B. (2018). *modi: Multivariate outlier detection and imputation for incomplete survey data* (Version 1.6.1) [Computer software]. <https://CRAN.R-project.org/package=modi>
- Johnson, J. A. (2005). Ascertain the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39(1), 103–129. <https://doi.org/10.1016/j.jrp.2004.09.009>
- Johnson, R. A., & Wichern, D. W. (2008). *Applied multivariate statistical analysis*. Pearson.
- Kam, C. C. S., & Chan, G. H. H. (2018). Examination of the validity of instructed response items in identifying careless respondents. *Personality and Individual Differences*, 129, 83–87. <https://doi.org/10.1016/j.paid.2018.03.022>

- Kaufman, J., Birmaher, B., Brent, D., Rao, U. M. A., Flynn, C., Moreci, P., Williamson, D., & Ryan, N. (1997). Schedule for affective disorders and schizophrenia for school-age children present and lifetime version (K-SADS-PL): initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry*, 36(7), 980–988. <https://doi.org/10.1097/00004583-199707000-00021>
- Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, 160(9), 1566–1577. <https://doi.org/10.1177/0734282906296233>
- Luderer, M., Kaplan-Wickel, N., Richter, A., Reinhard, I., Kiefer, F., & Weber, T. (2019). Screening for adult attention-deficit/hyperactivity disorder in alcohol dependent patients: Underreporting of ADHD symptoms in self-report scales. *Drug and Alcohol Dependence*, 195, 52–58. <https://doi.org/10.1016/j.drugalcdep.2018.11.020>
- Martel, M. M., Schimmack, U., Nikolas, M., & Nigg, J. T. (2015). Integration of symptom ratings from multiple informants in ADHD diagnosis: A psychometric model with clinical utility. *Psychological Assessment*, 27(3), 1060–1071. <https://doi.org/10.1037/pas0000088>
- Martel, M. M., Markon, K., & Smith, G. T. (2017). Research review: Multi-informant integration in child and adolescent psychopathology diagnosis. *Journal of Child Psychology and Psychiatry*, 58(2), 116–128. <https://doi.org/10.1111/jcpp.12611>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–453. <https://doi.org/10.1037/a0028085>
- Meijer, R. R., Egberink, I. J., Emons, W. H., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, 90(3), 227–238. <https://doi.org/10.1080/00223890701884921>
- Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, 23(1), 52–62. <https://doi.org/10.1177/1073191115577800>
- Nelson, J. M., & Lovett, B. J. (2019). Assessing ADHD in college students: Integrating multiple evidence sources with symptom and performance validity data. *Psychological Assessment*, 31(6), 793–804. <https://doi.org/10.1037/pas0000702>
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, 63, 1–11. <https://doi.org/10.1016/j.jrp.2016.04.010>
- Purpura, D. J., & Lonigan, C. J. (2009). Conners' teacher rating scale for preschool children: A revised, brief, age-specific measure. *Journal of Clinical Child & Adolescent Psychology*, 38(2), 263–272. <https://doi.org/10.1080/15374410802698446>
- Raiker, J. S., Rapport, M. D., Kolfer, M. J., & Sarver, D. E. (2012). Objectively-measured impulsivity and attention-deficit/hyperactivity disorder (ADHD): Testing competing predictions from the working memory and behavioral inhibition models of ADHD. *Journal of Abnormal Child Psychology*, 40(5), 699–713. <https://doi.org/10.1007/s10802-011-9607-2>
- Sibley, M. H., Pelham, W. E., Jr., Molina, B. S., Gnagy, E. M., Waxmonsky, J. G., Waschbusch, D. A., Derefinko, K. J., Wymbs, B. T., Garefino, A. C., Babinski, D. E., & Kuriyan, A. B. (2012). When diagnosing ADHD in young adults emphasize informant reports, DSM items, and impairment. *Journal of Consulting and Clinical Psychology*, 80(6), 1052–1061. <https://doi.org/10.1037/a0029098>
- Sibley, M. H., Campeze, M., & Raiker, J. S. (2019). Reexamining ADHD-related self-reporting problems using polynomial regression. *Assessment*, 26(2), 305–314. <https://doi.org/10.1177/1073191117693349>
- Smith, S. R. (2007). Making sense of multiple informants in child and adolescent psychopathology: a guide for clinicians. *Journal of Psychoeducational Assessment*, 25(2), 139–149. <https://doi.org/10.1177/0734282906296233>

- Smith, A. F., Baxter, S. D., Hardin, J. W., Guinn, C. H., & Royer, J. A. (2010). Relation of Children's dietary reporting accuracy to cognitive ability. *American Journal of Epidemiology*, *173*(1), 103–109. <https://doi.org/10.1093/aje/kwq334>
- Suhr, J. A., Buelow, M., & Riddle, T. (2011). Development of an infrequency index for the CAARS. *Journal of Psychoeducational Assessment*, *29*(2), 160–170. <https://doi.org/10.1177/073428291038019>
- Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5), 1–27. <https://doi.org/10.18637/jss.v074.i05>
- Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research*, *42*(2), 387–414. <https://doi.org/10.1080/00273170701360803>
- Von Rhein, D., Mennes, M., van Ewijk, H., Groenman, A. P., Zwiers, M. P., Oosterlaan, J., Helslenfeld, D., Franke, B., Hoekstra, P. J., Faraone, S. V., Hartman, C., & Buitelaar, J. (2015). The NeuroIMAGE study: A prospective phenotypic, cognitive, genetic and MRI study in children with attention-deficit/hyperactivity disorder. Design and descriptives. *European Child & Adolescent Psychiatry*, *24*(3), 265–281. <https://doi.org/10.1007/s00787-014-0573-4>
- Walls, B. D., Wallace, E. R., Brothers, S. L., & Berry, D. T. (2017). Utility of the Conners' adult ADHD rating scale validity scales in identifying simulated attention-deficit hyperactivity disorder and random responding. *Psychological Assessment*, *29*(12), 1437. <https://doi.org/10.1037/pas0000530>
- Wanders, R. B., Meijer, R. R., Ruhé, H. G., Sytema, S., Wardenaar, K. J., & de Jonge, P. (2017). Person-fit feedback on inconsistent symptom reports in clinical depression care. *Psychological Medicine*, *48*(11), 1–9. <https://doi.org/10.1017/S003329171700335X>

Chapter 19

Computerized Adaptive Testing Without IRT for Flexible Measurement and Prediction



L. Andries van der Ark and Niels Smits

Abstract In education, testing procedures can be lengthy. The long duration takes up precious time and affects the quality of responses, possibly resulting in a biased diagnosis or wrong treatment. The problem can be reduced using computer adaptive testing (CAT). However, three issues prevent the use of traditional CAT: (1) the type of tests and questionnaires we focus on do not allow for the construction of large item banks, (2) the test data are usually not (approximately) unidimensional, and (3) the aim of the researchers may not only be measurement but also prediction. We propose a flexible generalization of CAT to accommodate these three issues, coined FlexCAT. First, FlexCAT estimates the (discrete) density of item-score vectors (denoted \mathbf{p}) using any convenient model that provides a good description of \mathbf{p} ; this need not be an IRT model. Second, FlexCAT estimates test scores from $\hat{\mathbf{p}}$. In contrast to traditional CAT, the test score need not be a latent trait but can also be the total score, ordinal scores such as percentiles, or external criteria that the test aims to predict. We introduce FlexCAT for the case that a latent class model is used to estimate \mathbf{p} , and the total score is used as a test score. Using a real-data example, we compare the accuracy of FlexCAT and traditional CAT. Finally, we discuss the challenges FlexCAT still faces.

19.1 Introduction

In education, testing procedures can be lengthy. Especially for respondents who are unable to focus for long time periods, such as very young students or students in special needs education, standard educational tests pose a problem. When students get tired or distracted, they may resort to careless responding, or they may decide to stop the test procedure, possibly resulting in a biased test result or incorrect follow-

L. A. van der Ark (✉) · N. Smits

Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands

e-mail: L.A.vanderark@uva.nl; n.smits@uva.nl

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

L. A. van der Ark et al. (eds.), *Essays on Contemporary Psychometrics*,

Methodology of Educational Measurement and Assessment,

https://doi.org/10.1007/978-3-031-10370-4_19

up treatment. The test time can be reduced using computer adaptive testing (CAT; e.g., Magis et al., 2017; Wainer, 2000). However, CAT requires a large item bank, approximately unidimensional test data, and a latent trait with a known (typically a normal) distribution. Many tests, especially *typical performance tests*, do not allow for the construction of large item banks, as there are only a limited number of things one can ask a respondent. Also, many tests produce test data that are not approximately unidimensional. For example, there may be a dominant dimension and one or more nuisance dimensions. Finally, tests measuring certain phenomena typically produce a latent trait that has a skewed distribution (for some examples, see Molenaar et al., 2012). For such tests, traditional CAT may be suboptimal.

Consider the School Attitude Questionnaire Internet (SAQI, Vorst, 2006; also see, Psi Testuitgevers, n.d.), a test for students aged 9–16 years. The 160 trichotomous items measure motivation, well-being, and self-confidence with respect to going to school. The SAQI consists of ten scales. The SAQI provides scores at the scale level, aggregated scale level (i.e., motivation, well-being, and self-confidence), and at the overall level (a total score). The administration of 160 items may take more than 2 h, which can be strenuous for young students. Using a CAT could be helpful to reduce the response burden. However, the requirements of a CAT pose a problem. For constructs such as motivation, well-being, and self-confidence, it is infeasible to write enough items to fill a large item bank, as there is only a limited number of questions one can ask on these topics. Also, the SAQI aggregated-level scores “motivation,” “well-being,” and “self-confidence,” and the SAQI total score are the sum of multiple scale scores. As a result these scores are multidimensional. Also, even several SAQI scale-level scores are multidimensional. As traditional CAT assumes that the data are unidimensional, traditional CAT may produce biased estimates of the SAQI scores, and this bias may also be present in other typical-performance tests and possibly also in some maximum-performance tests.

In this chapter, we propose an alternative view on CAT, coined *FlexCAT*, that allows for the use of more flexible models than item response theory (IRT) models, which are traditionally used in CAT. First, we briefly describe the five building blocks of a traditional CAT. Second, we introduce FlexCAT using the same five building blocks. Third, using SAQI item scores, we compare the accuracy of FlexCAT and traditional CAT. Finally, we discuss the challenges of FlexCAT that must be resolved.

19.2 Traditional CAT

CAT procedures are iterative procedures. The algorithms for CAT have often been described as containing five building blocks (e.g., Wainer, 2000; Weiss & Kingsbury, 1984). Figure 19.1 shows a flow diagram of the five building blocks in an iterative CAT procedure that also fits FlexCAT. Building blocks “calibration” and “starting level” are grouped together in the *preliminary phase*, as the calibration and determining the starting level take place before the item administration. The

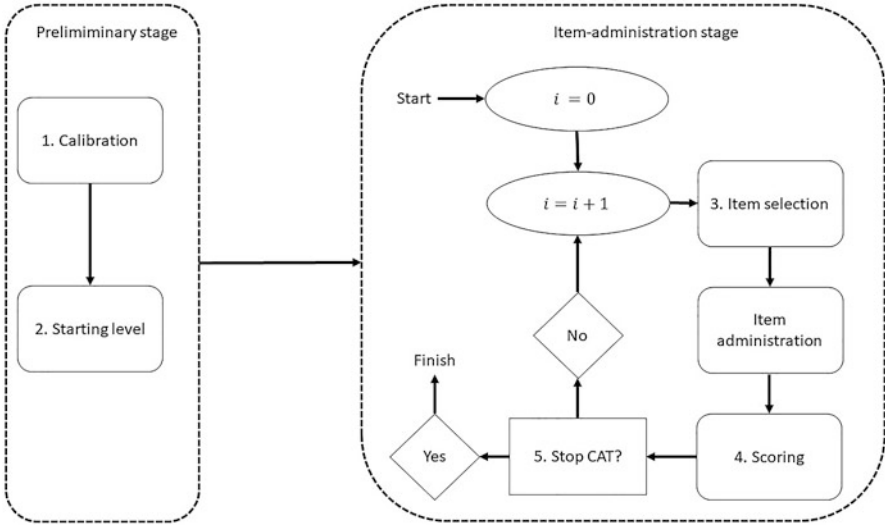


Fig. 19.1 Flow diagram of the five numbered building blocks of CAT in an iterative process (indicated by i): (1) calibration, (2) starting level, (3) item selection, (4) scoring, and (5) the decision whether to stop the CAT. Administering an item to a respondent (indicated by “Item administration”) is not part to the CAT algorithm and therefore not considered a building block

remaining building blocks are grouped together in the *item administration* phase, these building blocks are part of the measurement procedure of a single respondent.

Calibration First, the items of the complete test (the “item bank”) should be calibrated under an IRT model to obtain the item parameters that feed the CAT-algorithm. The selected IRT model should match the item format (e.g., Edelen & Reeve, 2007). Suppose the two-parameter logistic model is used to model dichotomously scored items. The probability that a randomly chosen respondent with latent trait score θ has a response X_j of 1 on item j is given by

$$P(X_j = 1|\theta) = \frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}}, \tag{19.1}$$

where α_j is the item’s slope parameter and δ_j is its location parameter.

Starting Level Usually, there is no information available about the respondent before the administration of the first item, and therefore some provisional estimate of the latent trait is required at the start of the CAT (Wainer, 2000). Most often, the average of the latent trait in the population is taken as a starting point, and the item that is most informative for that value is thus selected. Once the starting level has been determined, the item administration stage starts.

Item Selection Once an estimate of the respondent's latent trait has been obtained, a new item is selected that is most informative about this estimate. Let a prime denote the first derivative. Then, for item j , Fisher information

$$I_j(\theta) = \frac{[P'(X_j=1|\theta)]^2}{P(X_j=1|\theta)P(X_j=0|\theta)}, \quad (19.2)$$

may be used to quantify measurement quality as a function of the latent trait. From the items that have not yet been administered, the item with the highest information at the current estimate $\hat{\theta}$ is selected. The selected item is administered to the respondent, and the resulting item score is obtained.

Scoring After obtaining the item score, the CAT updates the estimate of the respondent's latent trait value. There are two popular latent trait estimation methods. Maximum likelihood (ML) estimates θ as the value with the highest likelihood of producing the observed responses (Thissen, 1991). By contrast, Bayesian estimation adds to this likelihood a prior distribution of the latent trait, such as the standard normal distribution (e.g., Embretson & Reise, 2000). Bayesian estimation can and ML estimation cannot provide an estimate for perfect response patterns. Let $f(\theta)$ denote the prior distribution of θ , and let $L(\theta)$ denote the likelihood function. One Bayesian method, expected a posteriori (EAP), takes the average of the posterior distribution of the latent trait, that is,

$$\hat{\theta}_{\text{EAP}} = \frac{\int \theta f(\theta)L(\theta)d\theta}{\int f(\theta)L(\theta)d\theta}. \quad (19.3)$$

Stopping Rule The CAT algorithm alternately administers items and updates the estimate of the respondent's latent trait score until the item pool is exhausted unless a termination criterion is specified, such as a pre-specified level of measurement precision. This criterion is met when the respondent's standard error of θ is small enough. The standard error when using EAP estimation is given by

$$SE(\hat{\theta}_{\text{EAP}}) = \sqrt{\frac{\int (\theta - \hat{\theta}_{\text{EAP}})^2 f(\theta)L(\theta)d\theta}{\int f(\theta)L(\theta)d\theta}}. \quad (19.4)$$

19.3 General Concept of FlexCAT

The main differences between FlexCAT and traditional CAT are in the building blocks calibration and starting level. The other building blocks also differ between FlexCAT and traditional CAT, but these differences are merely adaptations that are required because the building blocks calibration and starting level are rather different. Therefore, we discuss these two building blocks first.

19.3.1 Calibration

In FlexCAT, the calibration step entails the estimation of the *density of the item-score vectors* using a large sample. An item-score vector is a vector containing scores on all items. Suppose a test consists of J items, indexed by j ($j = 1, \dots, J$), and suppose that item j has $C_j + 1$ response categories, $0, \dots, c, \dots, C_j$. Then the number of possible item-score vectors equals $V = \prod_j (C_j + 1)$. For simplicity, but without loss of generalizability, we assume that all items have the same number of categories, that is, $C_j = C$ for all j . As a result, the number of possible item-score vectors equals

$$V = \prod_j (C + 1) = (C + 1)^J. \quad (19.5)$$

Let X_j denote the integer score on item j , with realization x_j ($x_j \in \{0, \dots, c, \dots, C\}$). Let $\mathbf{r}_v = (x_{v1}, \dots, x_{vJ})^T$ ($v = 1, \dots, V$) denote the v th item-score vector. The item-score vectors can be collected in a $V \times J$ matrix $\mathbf{R} = (\mathbf{r}_1^T, \dots, \mathbf{r}_V^T)$. The density of the item-score vectors, collected in the $V \times 1$ vector $\mathbf{p} = (P(\mathbf{r}_1), \dots, P(\mathbf{r}_V))$, plays a central role in the calibration step.

In traditional CAT, it is assumed that an IRT model generates \mathbf{p} . Using Eq. 19.1 and the property of local independence, it follows that

$$\begin{aligned} P(\mathbf{r}_v) &= P(X_1 = x_{v1}, \dots, X_J = x_{vJ}) = \int \prod_j P(X_j = x_{vj} | \theta) f(\theta) d\theta \\ &= \int \prod_j \left[\frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}} \right]^{x_{vj}} \left[1 - \frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}} \right]^{1 - x_{vj}} f(\theta) d\theta. \end{aligned} \quad (19.6)$$

Estimation of \mathbf{p} in traditional CAT (Eq. 19.6) thus requires estimating the item parameters α_j and δ_j and the distribution of the latent trait, $f(\theta)$.

The first notion of FlexCAT is that it has no assumptions on the process that may have generated \mathbf{p} , and the procedure is completely data driven. Vector \mathbf{p} can be estimated using any *convenient* model that provides a *good description* of the item-score vector density. In this chapter, “convenient” means that \mathbf{p} can be estimated directly from the test data, without the test constructor providing additional information (e.g., the number of dimensions or distributional assumptions). A “good description” is used pragmatically and means that the estimated item-score density, $\hat{\mathbf{p}}$, describes the associations in the test data so well that it provides a useful tool for measurement and prediction.

Hence, in FlexCAT the calibration stage consists of finding an estimate of \mathbf{p} with a model of choice. Besides IRT models, candidate models for estimating \mathbf{p} include the latent class model (LCM; e.g., Vermunt et al., 2008; Linzer, 2011; Van Buuren & Eggen, 2017), the divisive LCM (Van der Palm et al., 2016), kernel estimation methods (e.g., Li & Racine, 2003), and decision trees (e.g., Ho, 1995; Yan et al., 2004).

19.3.2 *Starting Level*

In traditional IRT, the estimated latent trait (e.g., $\hat{\theta}_{EAP}$, Eq. 19.4) is used as a score to communicate the measurement of a respondent. The starting level—when there is no information about the respondent yet—is the average latent trait level. The second notion of FlexCAT is that any score that can be derived from \mathbf{p} can be used to communicate measurement results. Hence, for FlexCAT, determining which score will be used in the CAT procedure is part of the building block “starting level.” Besides $\hat{\theta}_{EAP}$, a possible candidate is the *total score* (or equivalently, the mean item score) as most tests use the total score for measurement. Both the estimated latent trait and the total score can also be transformed to standard scores, percentile scores, or stanines to facilitate communication and interpretation. These adapted test scores can also be used as scores in FlexCAT. If the goal of the test is selection or prediction, a response variable could be a useful score. Examples of response variables include treatment (yes, no), placement (several nominal categories), or selection (selected, not selected). Note that when FlexCAT is used for prediction, the response variable (Y) must be included in the calibration model. For example, if a ten-item test should predict whether or treatment is effective ($Y = 1$) or not ($Y = 0$), then the v th item-score vector used for estimating \mathbf{p} should be $\mathbf{r}_v = (X_{v1} = x_{v1}, \dots, X_{v10} = x_{v10}, Y_v = y_v)$. LCMs and decision trees can easily incorporate response variables while calibrating items, but this is more difficult for standard IRT models.

19.3.3 *Item Selection, Scoring, and Stopping Rule*

In FlexCAT, the item-administration stage—item selection, scoring (or more accurately updating the score), and stopping rules—are essentially the same as for traditional CAT. However, based on the choices made during building blocks “calibration” and “starting level,” the building blocks in the item administration stage may have to be adapted. For example, when using the LCM for calibration and the total score for measurement, Fisher information (Eq. 19.2) is unavailable, and alternatives should be developed. Also, for the—discrete—total score, a stopping rule based on the modal value may be preferred over a stopping rule based on standard errors of the score (Eq. 19.4). As the building blocks in the item administration stage should be adapted depending on the choices made for calibration model and score, FlexCAT is more like an umbrella term for different types of CAT.

19.4 FlexCAT Using the Latent Class Model and the Total Score

19.4.1 Calibration

As a showcase, we show the estimation of item-score vector density \mathbf{p} using the LCM with W latent classes denoted LCM(W). Let Ξ denote the categorical latent variable having W categories (classes). The parameters of LCM(W) are the *class weights* $\pi_w \equiv P(\Xi = w)$ ($w = 1, \dots, W$) and the *conditional item score probabilities* $\pi_{j(c)|w} \equiv P(X_j = c | \Xi = w)$ ($j = 1, \dots, J; c = 0, \dots, C; w = 1, \dots, W$). Under the LCM(W)

$$P(X_j = x_j) = \sum_w \pi_w \pi_{j(x_j)|w}. \tag{19.7}$$

LCMs assume that item scores are locally independent given the score on Ξ , that is,

$$P(\mathbf{r}_v) = P(X_1 = x_1, \dots, X_J = c_J) = \sum_w \prod_j \pi_w \pi_{j(c_j)|w} \tag{19.8}$$

(cf. Eq. 19.6).

Table 19.1 shows a constructed small example with three dichotomous items. It is assumed that the estimated parameters of the LCM(2) provide a good description of the data. Hence, $\hat{\mathbf{p}}$ is derived from the parameters of the LCM(2) (see note in Table 19.1). For density estimation using the LCM, two issues are important.

Table 19.1 Example of LCM(2) parameter estimates for three dichotomous items, the matrix containing the $V = 8$ possible item-score vectors (\mathbf{R}), and the estimated density of the item-score vectors ($\hat{\mathbf{p}}$), which is derived from the latent class parameters (see note)

Latent class parameters			\mathbf{R}	$\hat{\mathbf{p}}$
$\hat{\pi}_w$	$\hat{\pi}_{j 1}$	$\hat{\pi}_{j 2}$		
$\begin{pmatrix} .2 \\ .8 \end{pmatrix}$	$\begin{pmatrix} .3 & .7 \\ .2 & .8 \\ .1 & .9 \end{pmatrix}$	$\begin{pmatrix} .6 & .4 \\ .9 & .1 \\ .6 & .4 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} .2604 \\ .1836 \\ .0336 \\ .0624 \\ .1756 \\ .1404 \\ .0304 \\ .1136 \end{pmatrix}$

Note. $\hat{p}_1 = \hat{\pi}_1 \times (\hat{\pi}_{1(0)|1} \times \hat{\pi}_{2(0)|1} \times \hat{\pi}_{3(0)|1}) + \hat{\pi}_2 \times (\hat{\pi}_{1(0)|2} \times \hat{\pi}_{2(0)|2} \times \hat{\pi}_{3(0)|2}) = .2 \times (.3 \times .2 \times .1) + .8 \times (.6 \times .9 \times .6) = .2604$, $\hat{p}_2 = \hat{\pi}_1 \times (\hat{\pi}_{1(0)|1} \times \hat{\pi}_{2(0)|1} \times \hat{\pi}_{3(1)|1}) + \hat{\pi}_2 \times (\hat{\pi}_{1(0)|2} \times \hat{\pi}_{2(0)|2} \times \hat{\pi}_{3(1)|2}) = .2 \times (.3 \times .2 \times .9) + .8 \times (.6 \times .9 \times .4) = .1836$, etc.

Goodness of Fit If the LCM is used as a density estimation tool, the interpretation of the latent classes is not particularly important (Vermunt et al., 2008; also, see Linzer, 2011). Therefore, issues that are important in traditional latent class analysis, such as local optima (e.g., McCutcheon, 2002), obtaining a modest number of latent classes to facilitate interpretation, and identifiability (e.g., Goodman, 1974), are not so important for the LCM as a density estimation tool, as long as the estimated density captures the higher-order interactions well. If the number of latent classes, W , is too small, the density is underfitted, which means that important associations or interactions are possibly ignored in the estimated density. If W is too large, the density may be overfitted; that is, the density estimate contains certain random fluctuations that are sample specific. Determining the correct W is typically done using information criteria, such as AIC (e.g., Bozdogan, 1987) or BIC (Schwarz, 1978). For increasing numbers of W (starting with $W = 1$), the information criterion is computed for LCM(W), and the LCM(W) that produces the lowest value of the information criterion is selected as a density estimator. It is well known that AIC tends to overestimate W , and BIC tends to underestimate W (e.g., Lukociene & Vermunt, 2010). Vermunt et al. (2008, p. 378) noted that overfitting is less problematic than underfitting, and for now, we advocate using AIC to determine W . However, this is an issue that should be investigated further, as there are many alternative information criteria and also indices for local fit (e.g., Nagelkerke et al., 2016).

Computational Feasibility The size of the $V \times 1$ vector \mathbf{p} can increase dramatically. For example, for the SAQI ($J = 160$ items, $C + 1 = 3$ categories), $V = 3^{160} \approx 2.18 \times 10^{76}$ (cf. Equation 19.5), which is computationally infeasible. As the number of free parameters in the LCM equals $W - 1 + W \times J \times C$, for the SAQI, $\hat{\mathbf{p}}$ is estimated using $W - 1 + W \times 480$ parameters. For $W = 200$, which is a large number of latent classes (e.g., see example in Vermunt et al., 2008), the number of parameters is less than 100,000, which is computationally feasible, although the density estimation procedure may be slow. Standard software (e.g., **poLCA**; Linzer & Lewis, 2011; or **Latent GOLD**, Vermunt & Magidson, 2013) can be used to estimate \mathbf{p} .

19.4.2 Starting Level

At the starting level, the density of the selected score is estimated. Here we use total score $X_+ = \sum_j X_j$. For J items, each having item scores $0, 1, \dots, C$, there are $H = JC + 1$ possible total scores, indexed by h ($h \in \{0, 1, \dots, H - 1\}$). Let $\mathbf{x}_+ = (0, \dots, H - 1)^T$ be an $H \times 1$ vector containing all possible total scores. The density of the total scores can be collected in an $H \times 1$ vector $\mathbf{p}_{X_+} = (P[X_+ = 0], \dots, P[X_+ = H - 1])^T$. Let \mathbf{Q} be a $V \times H$ design matrix that relates \mathbf{p} to \mathbf{p}_{X_+} , and let $\mathbf{r}_+ = (r_{+1}, \dots, r_{+v}, \dots, r_{+V})^T$ be a $V \times 1$ vector containing the total scores of the item-score vectors in \mathbf{R} ; that is, $\mathbf{r}_+ = \mathbf{R} \cdot \mathbf{1}$. For the elements of \mathbf{Q} , simple matrix algebra shows that $q_{v, h+1} = 1$ if $r_{+v} = h$ and $q_{v, h+1} = 0$ otherwise,

Table 19.2 Continuation of the example in Table 19.1 showing the relation between the estimated item-score vector density $\hat{\mathbf{p}}$ and total-score density $\hat{\mathbf{p}}_{X_+}$. Item-score vectors (\mathbf{R}) and their estimated density ($\hat{\mathbf{p}}$) are taken from Table 19.1. The total scores produced by the item-score vectors are in $\mathbf{r}_+ = \mathbf{R} \cdot \mathbf{1}$. Design matrix \mathbf{Q} is derived from \mathbf{r}_+ (see text). Vector \mathbf{x}_+ contains all possible total scores. Total-score-density equals $\hat{\mathbf{p}}_{X_+} = \mathbf{Q}^T \hat{\mathbf{p}}$

\mathbf{R}	$\hat{\mathbf{p}}$	\mathbf{r}_+	\mathbf{Q}	\mathbf{x}_+	$\hat{\mathbf{p}}_{X_+}$
$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} .2604 \\ .1836 \\ .0336 \\ .0624 \\ .1756 \\ .1404 \\ .0304 \\ .1136 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 1 \\ 2 \\ 2 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}$	$\begin{pmatrix} .2604 \\ .3928 \\ .2332 \\ .1136 \end{pmatrix}$

for $h = 0, \dots, H - 1$. It follows that $\mathbf{p}_{X_+} = \mathbf{Q}^T \mathbf{p}$ (and $\hat{\mathbf{p}}_{X_+} = \mathbf{Q}^T \hat{\mathbf{p}}$). Table 19.2 continues the example from Table 19.1 and illustrates $\hat{\mathbf{p}}_{X_+} = \mathbf{Q}^T \hat{\mathbf{p}}$.

19.4.3 Selecting the Next Item

Just as CAT, FlexCAT is an iterative process (Fig. 19.1). The starting level can be seen as iteration $i = 0$, where no item has yet been administered. Iteration i ($i = 1, 2, \dots$) starts with the selection of the i th item. As noted earlier, Fisher information (Equation 19.2) is unavailable here. A possible strategy for selecting the i th item for respondent n is searching for the item that provides as much information as possible on respondent n 's expected total score.

At the start of iteration i , there are $i - 1$ items that have already been administered to respondent n , whereas the remaining $G = J - i + 1$ items, indexed by g ($g = 1, \dots, G$), have not yet been administered to respondent n . Let $\mathbf{r}^{n,i-1}$ denote the item-score vector of respondent n at iteration $i - 1$; that is, $\mathbf{r}^{n,i-1}$ contains $i - 1$ observed item scores obtained in the previous iterations and G missing item scores. Similarly, let $\mathbf{r}_{X_g=c}^{n,i-1}$ denote the item-score vector of respondent n at iteration $i - 1$, assuming that respondent n will obtain score c on item g in iteration i . Let $P(X_g = c | \mathbf{r}^{n,i-1})$ denote the probability that respondent n will obtain score c on item g in iteration i , let $E(X_+ | \mathbf{r}^{n,i-1})$ denote the expected total score at iteration $i - 1$ for respondent n , and let $E(X_+ | \mathbf{r}_{X_g=c}^{n,i-1})$ denote the expected total score at iteration $i - 1$ for respondent n assuming that respondent n will obtain score c on item g in iteration i . A possible way to express the additional value of item g in iteration i on respondent n 's expected total score is

$$\Delta_g^{n,i} = \sum_c P(X_g = c | \mathbf{r}^{n,i-1}) \left| E(X_+ | \mathbf{r}_{X_g=c}^{n,i-1}) - E(X_+ | \mathbf{r}^{n,i-1}) \right|. \quad (19.9)$$

The absolute difference between $E(X_+ | \mathbf{r}_{X_g=c}^{n,i-1})$ and $E(X_+ | \mathbf{r}^{n,i-1})$ in Equation 19.9 is the effect of having $X_g = c$ on the expected total score; this effect is weighed by the probability that $X_g = c$ actually occurs. $\Delta_g^{n,i}$ is then the sum of these weighed effects over all response categories of item g . The item that produces the highest value $\Delta_g^{n,i}$ is selected as the next item to be administrated. $\Delta_g^{n,i}$ can be computed relatively easily. Let $\mathbf{a}^{i-1,n}$ be an indicator vector of length V , with $a_v^{i-1,n} = 1$ if the v th item-score vector in \mathbf{R} is still admissible given respondent n 's responses in the previous $i - 1$ iterations, and $a_v^{i-1,n} = 0$, otherwise. Similarly, let $\mathbf{a}_{X_g=c}^{i-1,n}$ be an indicator vector of length V , with $a_v^{i-1,n} = 1$ if the v th item-score vector in \mathbf{R} is still admissible given respondent n 's responses in the previous $i - 1$ iterations and given that respondent n would obtain item score $X_g = c$ if item g were to be administered in iteration i ; and $a_v^{i-1,n} = 0$, otherwise. Table 19.3 shows an example to illustrate $\mathbf{a}^{i-1,n}$ and $\mathbf{a}_{X_g=c}^{i-1,n}$.

Let $\mathbf{x} \circ \mathbf{y}$ denote the Hadamard or elementwise product of vectors \mathbf{x} and \mathbf{y} , and let $\left[\frac{\mathbf{x}}{\mathbf{y}}\right]$ be a vector that consist of the elementwise division of \mathbf{x} by \mathbf{y} . For example, for $\mathbf{x} = [3, 2]$ and for $\mathbf{y} = [1, 2]$, then $\mathbf{x} \circ \mathbf{y} = [3, 4]$, and $\left[\frac{\mathbf{x}}{\mathbf{y}}\right] = [3, 1]$. The terms in Eq. 19.9 can be expressed as

$$P(X_g = c | \mathbf{r}^{i-1,n}) = \frac{\mathbf{1}^T [\mathbf{a}_{X_g=c}^{i-1,n} \circ \mathbf{p}]}{\mathbf{1}^T [\mathbf{a}^{i-1,n} \circ \mathbf{p}]}, \tag{19.10}$$

$$E(X_+ | \mathbf{r}^{i-1,n}) = \mathbf{x}_+^T \mathbf{Q}^T \left[\frac{\mathbf{a}^{i-1,n} \circ \mathbf{p}}{\mathbf{11}^T (\mathbf{a}^{i-1,n} \circ \mathbf{p})} \right], \tag{19.11}$$

Table 19.3 Example showing design vectors $\mathbf{a}^{i-1,n}$ and $\mathbf{a}_{X_g=c}^{i-1,n}$ for respondent n in iteration $i = 2$, who has endorsed item 3 ($X_3 = 1$) in iteration 1

\mathbf{R}	$\mathbf{a}^{1,n}$	$\mathbf{a}_{X_1=0}^{1,n}$	$\mathbf{a}_{X_1=1}^{1,n}$	$\mathbf{a}_{X_2=0}^{1,n}$	$\mathbf{a}_{X_2=1}^{1,n}$
$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$

Note: As respondent n has endorsed item 3 in iteration 1, all item-score vectors in \mathbf{R} containing $X_3 = 0$ are inadmissible in iteration 2; hence, the corresponding elements in $\mathbf{a}^{1,n}$ are zeroes. In $\mathbf{a}_{X_1=0}^{1,n}$, the additional constraint is that $X_1 = 0$, leaving only two admissible item-score vectors. A similar logic applies to $\mathbf{a}_{X_1=1}^{1,n}$, $\mathbf{a}_{X_2=0}^{1,n}$, and $\mathbf{a}_{X_2=1}^{1,n}$.

and

$$E \left(X_+ | \mathbf{r}_{X_g=c}^{i-1,n} \right) = \mathbf{x}_+^T \mathbf{Q}^T \left[\frac{\mathbf{a}_{X_g=c}^{i-1,n} \circ \mathbf{p}}{\mathbf{11}^T (\mathbf{a}_{X_g=c}^{i-1,n} \circ \mathbf{p})} \right]. \quad (19.12)$$

We have provided Eqs. 19.10, 19.11, and 19.12 in matrix notation, so they are consistent with our computer code in the vector-based programming language R (R Core Team, 2021). In the Appendix, we elaborate on these equations. As \mathbf{x}_+^T and \mathbf{Q} are fixed design matrices, and \mathbf{p} has been estimated in the preliminary stage (Fig. 19.1) and remains fixed in the item administration stage, Eqs. 19.10, 19.11, and 19.12 show that only design vectors $\mathbf{a}^{i-1,n}$ and $\mathbf{a}_{X_g=c}^{i-1,n}$ require modification for computing $\Delta_g^{n,i}$ (Eq. 19.9). In the running example, at iteration 1 (no items have been administered), Eq. 19.9 results in $\Delta_1^{n,1} = 0.612$, $\Delta_2^{n,1} = 0.544$, and $\Delta_3^{n,1} = 0.660$. Hence, item 3 would be selected as the first item to be administered to all respondents.

19.4.4 Scoring

After a new item has been selected, the item is administered to the respondent (Fig. 19.1). Once the respondent has provided the score to the selected item, the estimated score density has to be updated from $\hat{\mathbf{p}}_{X_+}^{i-1,n}$ to $\hat{\mathbf{p}}_{X_+}^{i,n}$. Suppose that respondent n has obtained score c on item g in iteration i , then $\mathbf{a}^{i,n}$ is an $V \times 1$ indicator vector, with $a_v^{i,n} = 1$ if the v th item-score vector in \mathbf{R} is still admissible given respondent n 's responses to the previously administered i items, and $a_v^{i,n} = 0$, otherwise.

Vector $\mathbf{a}^{i,n}$ can be updated from $\mathbf{a}^{i-1,n}$ by setting the elements in $\mathbf{a}^{i-1,n}$ that correspond to response patterns in which $X_g = c$ to 0. The item-score vector density and total-score density are updated using

$$\hat{\mathbf{p}}^{i,n} = \left[\frac{\mathbf{a}^{i-1,n} \circ \mathbf{p}}{\mathbf{11}^T (\mathbf{a}^{i-1,n} \circ \mathbf{p})} \right] \quad (19.13)$$

and

$$\hat{\mathbf{p}}_{X_+}^{(i,n)} = \mathbf{Q}^T \hat{\mathbf{p}}^{(i,n)} \quad (19.14)$$

19.4.5 Stopping Rule

As a possible stopping rule, FlexCAT may be terminated if the modal value of $\hat{\mathbf{p}}_{X_+}^{i,n} > c$; that is, $\max(\hat{\mathbf{p}}_{X_+}^{i,n}) > c$, where $0 \leq c \leq 1$. If $c < \max(\hat{\mathbf{p}}_{X_+}^{0,n})$, FlexCAT stops before any item has been administered. If $c = 1$, all items will be administered. For all remaining values of c , it holds that if c becomes larger, the precision of the score estimate increases, but the expected number of administered items increases as well. We stress that alternative stopping rules may be used as well. For example, one may compute the expected sum score $E(X_+ | \mathbf{r}^{i,n})$ (cf. Equation 19.9) and use its standard deviation as a measure of precision.

19.4.6 Small Example

Table 19.4 shows the iterative procedure for the running example based on the LCM(2) in Table 19.1, using the stopping rule $\max(\hat{\mathbf{p}}_{X_+}^{i,n}) > .9$. At iteration $i = 0$, Table 19.4 shows the matrix of item-score vectors (\mathbf{R} ; taken from Table 19.1), design matrix $\mathbf{a}^{0,n}$, estimated item-score vector density $\hat{\mathbf{p}}$ (taken from Table 19.1), the best estimate of the item-score vector density for respondent n at iteration 0 ($\hat{\mathbf{p}}^{0,n}$), transformation matrix \mathbf{Q} (taken from Table 19.2), and the estimated score density ($\hat{\mathbf{p}}_{X_+}^{0,n}$). Note that $\mathbf{a}^{0,n} = \mathbf{1}$ shows that all item-score vectors are still admissible. Also note that $\hat{\mathbf{p}}^{0,n} = \hat{\mathbf{p}}$ as there is no information yet on respondent n in iteration 0. As $\max(\hat{\mathbf{p}}_{X_+}^{0,n}) = .3929 < .9$, FlexCAT continues.

At iteration $i = 1$, $\Delta^{1,n} = (\Delta_1^{1,n}, \Delta_2^{1,n}, \Delta_3^{1,n})^T$ (Equation 19.9) has the highest value at $\Delta_3^{1,n}$; hence, item 3 is selected as the new item and presented to respondent n . Respondent n obtains item score $X_3 = 1$. Hence design matrix $\mathbf{a}^{1,n}$ has all elements that pertain to item-score vectors for which $X_3 = 0$ set to zero, resulting in updates of the item-score vector density ($\hat{\mathbf{p}}^{1,n}$) and total-score density ($\hat{\mathbf{p}}_{X_+}^{1,n}$). As $\max(\hat{\mathbf{p}}_{X_+}^{1,n}) = .4056 < .9$, the CAT continues. At iteration 2, item 1 is selected, and respondent n obtains item score $X_1 = 1$. As $\max(\hat{\mathbf{p}}_{X_+}^{2,n}) = .5527 < .9$, the CAT continues. At iteration 3, all items have been administered, necessarily leading to $\max(\hat{\mathbf{p}}_{X_+}^{3,n}) = 1 > .9$, so FlexCAT terminates, and the expected (and real) score equals 2.

Table 19.4 Iterative procedure for the running example based on the LCM(2) in Table 19.1. For details see text

i	$\Delta^{i,n}$	g	X_g	\mathbf{R}	$\mathbf{a}^{i,n}$	\mathbf{p}	$\hat{\mathbf{p}}^{i,n}$	\mathbf{IQ}	$\hat{\mathbf{p}}_{X_+}^{i,n}$	S
0				$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} .2604 \\ .1836 \\ .0336 \\ .0624 \\ .1756 \\ .1404 \\ .0304 \\ .1136 \end{pmatrix}$	$\begin{pmatrix} .2604 \\ .1836 \\ .0336 \\ .0624 \\ .1756 \\ .1404 \\ .0304 \\ .1136 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} .2604 \\ .3928 \\ .2332 \\ .1136 \end{pmatrix}$	N
1	$\begin{pmatrix} .6120 \\ .5440 \\ .6600 \end{pmatrix}$	3	1		$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$		$\begin{pmatrix} 0 \\ .3672 \\ 0 \\ .1248 \\ 0 \\ .2808 \\ 0 \\ .2272 \end{pmatrix}$	$\begin{pmatrix} 0 \\ .3672 \\ .4056 \\ .2272 \end{pmatrix}$	N	
2	$\begin{pmatrix} .5966 \\ .5530 \\ - \end{pmatrix}$	1	1		$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$		$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ .5528 \\ 0 \\ .4472 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ .5527 \\ .4472 \end{pmatrix}$	N	
3	$\begin{pmatrix} - \\ .4944 \\ - \end{pmatrix}$	2	0		$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$		$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$	Y	

Note: i = iteration; $\Delta^{i,n}$ = vector of deltas (see text); g = selected item; X_g = respondent's n score on the selected item; for \mathbf{R} , $\mathbf{a}^{i,n}$, $\hat{\mathbf{p}}$, $\hat{\mathbf{p}}^{i,n}$, \mathbf{Q} , and $\hat{\mathbf{p}}_{X_+}^{i,n}$, see text; S = Stop?; N = No (i.e., $\max(\hat{\mathbf{p}}_{X_+}^{i,n}) \leq .9$); Y = Yes (i.e., $\max(\hat{\mathbf{p}}_{X_+}^{i,n}) > .9$)

19.5 Comparing FlexCAT and Traditional CAT

We compared the outcomes of traditional CAT to the outcomes of FlexCAT using LCM and the total score as described above. The study serves as an illustration of how different types of CAT can be compared.

19.5.1 Method

Data We used the scores from 4211 Belgian students aged between 9 and 19 (53% women) to the 16 items of the SAQI scale *Leertaakgerichtheid* (Orientation Towards Learning Task). The data had been deidentified, and respondents with missing item scores had been removed before we obtained the data. As dichotomous item scores were easiest to handle in a traditional CAT, we coded the response “that is true” to item score 1, and responses “that is sometimes true” and “that is not true” to item score 0. We investigated the dichotomous item scores using the Mokken scale analysis (e.g., Sijtsma & Van der Ark, 2017) and found no violations of unidimensionality, local independence, or monotonicity. The scalability coefficient for the entire scale was $H = .427$ ($SE = .007$), suggesting a “medium scale” using Mokken’s (1971) benchmarks.

Simulation Design The original data were split randomly in a training set (80% of the item-score vectors, $N = 3369$) used for calibration and a validation set (20% of the item score vectors, $N = 842$). The two types of CAT were applied to each of the 842 item-score vectors in the validation set. The response to an item in the CAT equaled a respondent’s actual response in the data. As a result, the data obtained from each CAT procedure was a 842×16 matrix. Items administered in the CAT had scores equal to the item scores in the data, and items not administered in the CAT had missing values.

For FlexCAT, we used the settings as described in this chapter. As a stopping rule, we used $\max(\mathbf{p}_{X_+}^{i,n}) > c$, using the following values of c : .90, .85, .80, and .75. The LCM was estimated using the R-package **poLCA** (Linzer & Lewis, 2011). For the remainder, we used our own computer code. For the calibration of traditional CAT, we used the two-parameter logistic model. In the traditional CAT, the average percentage of administered items was set approximately equal to the average percentage of administered items of FlexCAT by finetuning the required standard error in traditional CAT’s stopping rule. This allowed us to compare the quality of the measurement under an equal level of response burden. Both the calibration and the iterative item administration of the CAT were conducted using the R-package **mirt** (Chalmers, 2012) for calibration, and the R-package **mirtCAT** (Chalmers, 2016) for running the traditional CAT with default settings.

Dependent Variables As the scores used in traditional CAT (estimated latent trait value) and FlexCAT (estimated total score) were incomparable, we used the stanines of the respective scores to compare the two types of CAT: More specifically, we reported the percentage of respondents for which the stanine estimated using CAT was equal to the stanine computed from the complete data, the percentage of respondents for which the difference between the two stanines was 1, and the percentage of the respondents for which the difference between the two stanines was greater than 1. In addition we compared computed the correlation between a respondent's estimated and real score.

19.5.2 Results and Discussion

Depending on the stopping rule, for FlexCAT, the median percentage of administered items ranged between 75% (12 items) and 87.5% (14 items) and for traditional CAT between 75% (12 items) and 100% (16 items). For FlexCAT, the distribution of the number of administered was approximately symmetric (Table 19.5, upper panel) and skewed to the left for traditional CAT (Table 19.5, lower panel). These skewed distributions indicate that, compared to FlexCAT, a large proportion of the respondents in the traditional CAT required relatively few items, and a large proportion of the respondents require all items. FlexCAT showed smaller differences between the actual stanine and the expected stanine than traditional CAT (Table 19.6, middle columns), whereas the correlation between the actual scores and estimated scores were very high for both types of CAT (Table 19.6, last column).

For this example, results showed that FlexCAT and traditional CAT are both doing well, and although FlexCAT performed a bit better, the differences were not overwhelming. This can be expected as we found no violations of unidimensionality, local independence, and monotonicity for this scale, which suggests that a two-parameter logistic model can estimate the item-score vector density rather well. The percentage of items that were administered was less than typically expected in CAT,

Table 19.5 The percentage items of administered in FlexCAT and the corresponding percentage of items administered in traditional CAT, for the SAQI scale Leertaakgerichtheid (Orientation Towards Learning Task)

CAT	<i>c</i>	Min (%)	First quartile (%)	Second quartile (%)	Third quartile (%)	Max (%)
FlexCAT	.90	75.0	81.2	87.5	93.8	100.0
	.85	62.5	75.0	81.2	87.5	100.0
	.80	56.2	75.0	81.2	87.5	100.0
	.75	50.0	68.8	75.0	81.2	100.0
Trad. CAT	.90	56.3	68.8	100.0	100.0	100.0
	.85	50.0	62.5	93.8	100.0	100.0
	.80	50.0	56.3	81.2	100.0	100.0
	.75	43.8	50.0	75.0	100.0	100.0

Table 19.6 Difference between actual stanine and estimated stanine for FlexCAT and traditional CAT for the SAQI scale Leertaakgerichtheid (Orientation Towards Learning Task) and the correlation between the estimated and actual score

CAT	<i>c</i>	Difference			Correlation
		0	1	>1	
FlexCAT	.90	98%	2%	0%	.998
	.85	96%	4%	0%	.997
	.80	93%	7%	0%	.995
	.75	91%	9%	0%	.993
Trad. CAT	.90	92%	8%	0%	.996
	.85	93%	7%	0%	.994
	.80	89%	11%	0%	.992
	.75	87%	13%	0%	.991

which may be due to the rather strict stopping rules. Finally, it may be noted that FlexCAT was rather slow: In the simulations, processing a single respondent took approximately 40 s, compared to less than 1 s for a traditional CAT. As the number of items increase, computation time increases too, so for larger data sets, FlexCAT may be too slow.

19.6 Discussion

We proposed a generalization of CAT, coined FlexCAT, and we conjecture that FlexCAT will be useful for tests and questionnaires that do not meet the requirements of IRT models, tests, and questionnaires that are used for both measurement and prediction, and tests and questionnaires that have different measurement levels and items with different numbers of response categories. In a first example concerning the SAQI scale “Orientation Towards Learning Task,” we used the LCM to estimate the density of the item-score vectors (\mathbf{p}), and we used the total score for measurement, finding slightly better results for FlexCAT. The similarity could explain the quality of the scale, which showed no violations of the IRT-model assumptions. However, when multiple scales of SAQI should be administered, then FlexCAT has the additional advantage over traditional CAT: Item scores from scales that already have been administered may help predict the total score of a scale that still has to be administered, and thus reducing the response burden. The percentage of administered items was higher than expected, which suggests that stopping rules and other settings of FlexCAT should be thoroughly investigated. This chapter is merely the start of FlexCAT, and many things need to be investigated before FlexCAT can be used.

The LCM is an attractive candidate to estimate \mathbf{p} . We are not the first ones to apply the LCM to CAT. Cheng (2009) and Wang et al. (2012) used the LCM for a CAT for cognitive diagnostic models, which can be conceived as an LCM with 2^Q latent classes, where Q is the number of attributes required to make a test. From a FlexCAT perspective, these authors estimated \mathbf{p} using the LCM(2^Q) and used the same 2^Q classes weights as measurement scores. Similarly, Van Buuren and Eggen

(2017) estimated \mathbf{p} using the LCM with a small number of latent classes and used expected class membership as the measurement score. Our use of LCMs in the SAQI example was different, in the sense that we used the LCM as a convenient device to obtain an accurate estimate of \mathbf{p} , and we were not interested in the number of latent classes, class weights, or parameter identifiability. We were not the first to use the LCM as a density estimation method either. Van der Palm et al. (2016) used the divisive LCM to estimate discrete densities.

Before the LCM can be used as an off-the-shelf density estimator for FlexCAT, the following problems need to be resolved. First is the *curse of dimensionality* problem. As the number of items increases, the order of vector \mathbf{p} , which is C^J , increases exponentially. For, example, for $J = 130$ items having $C = 5$ ordered response categories, $C^J \approx 7.3 \times 10^{90}$. As 7.3×10^{90} is more than 1 billion times the commonly accepted number of particles in the observed universe, these numbers are beyond the computational limits that are physically possible (cf. Lloyd, 2000). Estimating \mathbf{p} for this test using the LCM(200) requires $(W - 1) + W \times J \times (C - 1) = 199 + 200 \times 130 \times 4 = 104,199$ free parameters. This is not a computational problem, even for a regular laptop, but the huge model makes FlexCAT very slow, possibly too slow for a sound administration. The administration of a ten-item CAT required 40 seconds, and the computation time increases as the number of items increases. This is one of the main issues that must be investigated. In addition, local optima (e.g., Shireman et al., 2016) may have a large effect on the estimates. Second, choices for goodness of fit criteria, item selection rules, and stopping rules need to be investigated.

Other models can also be used to estimate \mathbf{p} . From a FlexCAT perspective, Yan et al. (2004) used decision trees to estimate \mathbf{p} and the total score for measurement. Recently, Gonzalez (2021) provided machine-learning techniques for individual diagnostic assessment. Implementation of other models requires an adaptation of “building blocks” 2, 3, 4, and 5, which will lead to new challenges. Various choices of density-estimation models and scores for FlexCAT must be compared using both simulated and real-life data, to learn which choices tend to work well.

Probably the biggest challenge is the application of FlexCAT in real-life CAT administrations. In addition to a well-working FlexCAT, in which optimal choices have been made with respect to the density estimation, the score, item-selection rules, and stopping rules, it requires fast and user-friendly software and training programs for test administrators.

A.1 Appendix

Consider Eq. 19.10:

$$P(X_g = c | \mathbf{r}^{i-1, n}) = \frac{\mathbf{1}^T [\mathbf{a}_{X_g=c}^{i-1, n} \circ \mathbf{p}]}{\mathbf{1}^T [\mathbf{a}^{i-1, n} \circ \mathbf{p}]} \quad (19.10)$$

In the numerator, $\mathbf{a}_{X_g=c}^{i-1,n} \circ \mathbf{p}$ produces a $V \times 1$ vector \mathbf{p}^{**} where $p_v^{**} = p_v$ if $a_{X_g=c}^{i-1,n} = 1$ and $p_v^{**} = 0$ otherwise; that is, those probabilities from \mathbf{p}^T are selected that pertain to item-score vectors that are still admissible for respondent n , given the $i - 1$ previous item scores and given that score c on item g has been obtained in iteration i . Pre-multiplying \mathbf{p}^{**} with a unit vector sums up the admissible probabilities producing $P(\mathbf{r}^{n,i-1}, X_g = c)$. Analogously, in the denominator, $\mathbf{a}^{i-1,n} \circ \mathbf{p}$ produces a $V \times 1$ vector \mathbf{p}^* where $p_v^* = p_v$ if $a^{i-1,n} = 1$ and $p_v^* = 0$ otherwise; that is, those probabilities from \mathbf{p}^T are selected that pertain to item-score vectors that are still admissible for respondent n , given the $i - 1$ previous item scores. Pre-multiplying \mathbf{p}^* with a unit vector sums up the admissible probabilities producing $P(\mathbf{r}^{n,i-1})$. The ratio of $P(\mathbf{r}^{n,i-1}, X_g = c)$ and $P(\mathbf{r}^{n,i-1})$ equals $P(X_g = c | \mathbf{r}^{i-1,n})$.

In Eq. 19.11,

$$E \left(X_+ | \mathbf{r}^{i-1,n} \right) = \mathbf{x}_+^T \mathbf{Q}^T \left[\frac{\mathbf{a}^{i-1,n} \circ \mathbf{p}}{\mathbf{11}^T (\mathbf{a}^{i-1,n} \circ \mathbf{p})} \right], \tag{19.11}$$

the numerator of the last term results in vector \mathbf{p}^* (cf. denominator of Eq. 19.10), whereas the denominator equals $\mathbf{11}^T \mathbf{p}^*$, which is a $V \times 1$ vector with each element equal to $\sum_v p_v^*$. Hence the last term of Eq. 19.11 is the $V \times 1$ vector of rescaled probabilities of admissible item-score vectors $\left[\frac{p_1^*}{\sum_v p_v^*}, \frac{p_2^*}{\sum_v p_v^*}, \dots, \frac{p_V^*}{\sum_v p_v^*} \right]^T = \mathbf{p}^{n,i-1}$ (e.g., Table 19.4), Hence, Eq. 19.11 reduces to

$$E \left(X_+ | \mathbf{r}^{n,i-1} \right) = \mathbf{x}_+^T \cdot \mathbf{Q}^T \cdot \mathbf{p}^{n,i-1} = \mathbf{x}_+^T \cdot \mathbf{p}_{X_+}^{n,i-1}, \tag{A.1}$$

where $\mathbf{p}_{X_+}^{n,i-1}$ is the density of the total scores given the admissible item-score vectors. Because $\mathbf{x}_+^T \cdot \mathbf{p}_{X_+}^{n,i-1} = \sum_{h=0}^{H-1} h P(X_+ = h | \mathbf{r}^{n,i-1}) = E(X_+ | \mathbf{r}^{n,i-1})$, Eq. 19.11 is true. Equation 19.12 follows a very similar logic.

References

Bozdogan, H. (1987). Model selection and Akaike’s Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345–370. <https://doi.org/10.1007/BF02294361>

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(1), 1–29. <https://doi.org/10.18637/jss.v048.i06>

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, 71, 1–38. <https://doi.org/10.18637/jss.v071.i05>

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74(4), 619–632. <https://doi.org/10.1007/s11336-009-9123-2>

- Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(5), 5–18. <https://doi.org/10.1007/s11136-007-9198-0>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for Psychologists*. Psychology Press. <https://doi.org/10.4324/9781410605269>
- Gonzalez, O. (2021). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychological Methods*, 26(2), 236–254. <https://doi.org/10.1037/met0000317>
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2), 215–231. <https://doi.org/10.2307/2334349>
- Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282). <https://doi.org/10.1109/ICDAR.1995.598994>
- Li, Q., & Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2), 266–292. [https://doi.org/10.1016/S0047-259X\(02\)00025-8](https://doi.org/10.1016/S0047-259X(02)00025-8)
- Linzer, D. A. (2011). Reliable inference in highly stratified contingency tables: Using latent class models as density estimators. *Political Analysis*, 19(2), 173–187. <https://doi.org/10.1093/pan/mpm006>
- Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of statistical software*, 42(1), 1–29. <https://doi.org/10.18637/jss.v042.i10>
- Lloyd, S. (2000). Ultimate physical limits to computation. *Nature*, 406(6799), 1047–1054. <https://doi.org/10.1038/35023282>
- Lukociene, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 241–250). Springer. https://doi.org/10.1007/978-3-642-01044-6_22
- Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catr and mstr*. Springer. <https://doi.org/10.1007/978-3-319-69218-0>
- McCutcheon, A. L. (2002). Basic concepts and procedures in single- and multiple-group latent class analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 56–88). Cambridge University Press. <https://doi.org/10.1017/CBO9780511499531>
- Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. De Gruyter.
- Molenaar, D., Dolan, C. V., & De Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika*, 77(3), 455–478. <https://doi.org/10.1007/S11336-012-9273-5>
- Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2016). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology*, 46(1), 252–282. <https://doi.org/10.1177/0081175015581379>
- Psi testuitgevers. (n.d.). *SAQI vertaald* [SAQI translated]. Retrieved August 19, 2021, from https://www.psitestuitgevers.nl/producten/saqi_svl/vertaald/
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. <https://www.jstor.org/stable/2958889>
- Shireman, E. M., Steinley, D., & Brusco, M. J. (2016). Local optima in mixture modeling. *Multivariate Behavioral Research*, 51(4), 466–481. <https://doi.org/10.1080/00273171.2016.1160359>
- Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, 70(1), 137–158. <https://doi.org/10.1111/bmsp.12078>
- Thissen, D. (1991). *MULTILOG user's guide* [Computer software]. Scientific Software.

- Van Buuren, N., & Eggen, T. H. (2017). Latent-class-based item selection for computerized adaptive progress tests. *Journal of Computerized Adaptive Testing*, 5(2). <https://doi.org/10.7333/jcat.v5i2.62>
- Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2016). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification*, 33(1), 52–72. <https://doi.org/10.1007/s00357-016-9195-5>
- Vermunt, J. K., & Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax*. Statistical Innovations Inc. <https://www.statisticalinnovations.com/wp-content/uploads/LGtechnical.pdf>
- Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of categorical data using latent class analysis. *Sociological Methodology*, 38(1), 369–397. <https://doi.org/10.1111/j.1467-9531.2008.00202.x>
- Vorst, H. C. M. (2006). *School attitude questionnaire – Internet (SAQI)*. Libbe Mulder. <https://hdl.handle.net/11245/1.272122>
- Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Erlbaum.
- Wang, C., Chang, H.-H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, 44(1), 95–109. <https://doi.org/10.3758/s13428-011-0143-3>
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4), 361–375. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>
- Yan, D., Lewis, C., & Stocking, M. (2004). Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics*, 29(3), 293–316. <https://doi.org/10.3102/10769986029003293>

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 licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence 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

On the Relationship Between Unidimensional Item Response Theory and Higher-Order Cognitive Diagnosis Models



Jimmy de la Torre and Kevin Carl Santos

Abstract Cognitive diagnosis models (CDMs) have gained popularity in recent years due to the diagnostic feedback they could provide. For this reason, its emergence brought about a shift in the psychometric paradigm—from merely determining the subjects’ locations on a latent continuum to generating the subjects’ multidimensional profiles for a given set of fine-grained attributes. Although based on disparate underlying assumptions, it is not usual for many researchers to fit unidimensional item response theory (IRT) models and CDMs to the same educational or psychological assessment data. This chapter aims to explore the conditions under which such a practice can be deemed acceptable. By imposing certain conditions, the higher-order generalized deterministic input, noisy, “and” gate (HO-G-DINA) model is reformulated to express the success probability on an item as a function of the higher-order latent trait. Based on this model reformulation, this study provides a framework for relating the two classes of psychometric models, as well as boundaries within which this can be done. The correspondence between unidimensional IRT and the HO-G-DINA models is further examined using simulated and real data.

20.1 Introduction

At present, many existing large-scale educational assessments are developed and analyzed using unidimensional item response theory (IRT) models, which assume that the success probability on an item is a function of a single latent trait θ .

J. de la Torre (✉)
The University of Hong Kong, Hong Kong S.A.R., China
e-mail: j.delatorre@hku.hk

K. C. Santos
University of the Philippines - Diliman, Quezon City, Philippines
e-mail: kpsantos1@up.edu.ph

Although scores derived from these models are useful for scaling and ordering purposes, they are typically of limited value when it comes to pinpointing the students' specific strengths and weaknesses. As such, these scores do not provide diagnostic and prescriptive information that can facilitate instruction and learning.

In this regard, the advent of cognitive diagnosis models (CDMs), which are psychometric models that can be used to support instruction and learning, has sparked a vast interest among researchers and practitioners. These models are developed specifically to determine a student's mastery or nonmastery of multiple fine-grained skills (e.g., de la Torre 2009). To maximize the benefits of CDMs, they should be used in conjunction with cognitively diagnostic assessments (CDAs), which are assessments deliberately and thoughtfully designed to measure the different components required for someone to be deemed proficient in a particular domain of interest (de la Torre & Minchen 2014).

As of yet, the potential advantages of using CDMs to generate richer diagnostic feedback have not been fully realized as the rapid methodological CDM advancement has outpaced their applications in the educational settings. To date, only few diagnostic assessments have been developed within the CDM framework. For instance, Tjoe and de la Torre (2014) developed a proportional reasoning (PR) test for middle school students, whereas Bradshaw et al. (2014) constructed a multidimensional test examining middle grade teachers' understanding of fraction multiplication and division. Due to the dearth of such assessments, researchers have employed CDMs on assessments anchored in unidimensional IRT framework in the hope of extracting more diagnostic information. This approach is referred to as *retrofitting* as CDMs are fitted to the data *post hoc* (De la Torre & Karelitz 2009).

Retrofitted applications abound and include the analyses of the Trends in International Mathematics and Science Study data (Birenbaum et al. 2005; Choi et al. 2015; Lee et al. 2011; Tatsuoka et al. 2004), the 2003 Florida Comprehensive Assessment Test (FCAT) data (Lee et al. 2012), the 2003 National Assessment of Educational Progress data (de la Torre 2006), and the Graduate Record Examinations (GRE) data (Gorin & Embretson 2006) as examples. Chen and de la Torre (2014) laid out a procedure on how to diagnostically model extant large-scale assessment data by demonstrating it using the reading assessment of the Programme for International Student Assessment (PISA) 2000. More recently, Liu et al. (2018) proposed a step-by-step retrofitting framework and illustrated it by using a mock version of the Test of English as Foreign Language listening test.

Although a number of studies (e.g., De la Torre & Karelitz 2009; Lee et al. 2012) have compared CDMs and IRT models, it remains unclear to date the extent to which CDMs and unidimensional IRT models can be simultaneously used to analyze the same assessment data. To address this issue, the current study examines a unifying framework for relating the two psychometric frameworks, as well as boundaries within which the relationship can be expected to hold.

20.2 Background

20.2.1 Unidimensional Item Response Theory Models

IRT is a model-based measurement, where item responses are expressed as a function of the examinees' proficiency levels and item characteristics. Unidimensional IRT models provide single overall scores reflecting the proficiency levels of the examinees. Many new and revised assessments were developed based on IRT principles, including the Armed Services Vocational Aptitude Battery, the Scholastic Aptitude Test (SAT), and the GRE (Embretson & Reise 2000).

Four of the commonly used unidimensional IRT models are the four-parameter logistic (4PL; Barton & Lord 1981), the three-parameter logistic (3PL; Birnbaum 1968), the two-parameter logistic (2PL; Birnbaum 1968), and the Rasch (Rasch 1960), sometimes referred to as the one-parameter logistic (1PL), models. Let X_j be the binary response to item j . It is equal to 1 if item j is answered correctly and 0 otherwise. The item response function (IRF) of the 4PL IRT model can be expressed as

$$P(X_j = 1) = \gamma_j + (v_j - \gamma_j) \frac{1}{1 + e^{-\alpha_j(\theta - \delta_j)}}, \quad (20.1)$$

where α_j and δ_j are the discrimination and difficulty parameters and γ_j and v_j are the lower and the upper asymptotes, respectively, for item j and θ is the proficiency parameter. When $v_j = 1$, Eq. 20.1 reduces to the 3PL model; furthermore, if $\gamma_j = 0$, the 3PL simplifies to the 2PL model. Additionally, when common slope for all items (i.e., $\alpha_j = \alpha$ for all j) is assumed, the 2PL reduces to the 1PL.

20.2.2 Cognitive Diagnosis Models

CDMs, also referred to as diagnostic classification models, are restricted latent class models that can generate a multivariate binary vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k, \dots, \alpha_K)^T$, where $\alpha_k = 1$ or $\alpha_k = 0$ indicates mastery or nonmastery of attribute k . An important component of CDMs is the \mathbf{Q} -matrix (Tatsuoka, 1983). It is a $J \times K$ binary matrix that specifies the required skills to answer each item correctly, where J and K represent the number of test items and the number of attributes, respectively. The (j, k) th element of the \mathbf{Q} -matrix, denoted by q_{jk} , is equal to 1 if the k th attribute is required to answer item j correctly and is equal to 0 otherwise.

Although one of the most studied CDMs, the conjunctiveness condensation function assumed by the deterministic input, noisy, "and" gate (DINA; Haertel 1989; Junker & Sijtsma 2001) is deemed too restrictive. To relax the conjunctive assumption, several general CDMs for dichotomous responses have been proposed

in the literature, and one of these is the generalized DINA (G-DINA; de la Torre 2011) model. Without loss of generality, let the first K_j^* attributes be required for item j and α_{ij}^* be the l th reduced attribute pattern whose elements are the required attributes for item j . The IRF of the G-DINA model is given by

$$P(\alpha_{ij}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk} \alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'} \alpha_{lk} \alpha_{lk'} + \dots + \delta_{j12\dots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \tag{20.2}$$

where δ_{j0} is the baseline probability, δ_{jk} s are the main effects, $\delta_{jkk'}$ s are the two-way interaction effects, and $\delta_{j12\dots K_j^*}$ is the highest-order interaction effect.

Each reduced attribute pattern corresponds to a latent group. As a general CDM, aside from according each latent group its own success probability, the G-DINA model subsumes several reduced CDMs. The G-DINA model reduces to the DINA model by setting all the parameters, except δ_{0j} and $\delta_{j12\dots K_j^*}$, to zero; it reduces to the deterministic input, noisy, “or” gate (DINO; Templin & Henson 2006) model with the constraints

$$\delta_{jk} = -\delta_{jk'k''} = \dots = (-1)^{K_j^*+1} \delta_{j12\dots K_j^*}, \tag{20.3}$$

for $k = 1, \dots, K_j^* - 1$, and $k'' > k', \dots, K_j^*$; finally, it reduces to the additive CDM (A-CDM; de la Torre 2011) when all interaction effects are set to zero. Additive CDMs in other link functions can be derived in the same manner.

20.2.3 Relating IRT and CDMs

De la Torre and Karelitz (2009) systematically examined the relationship of unidimensional IRT models and CDMs, particularly, between the 2PL IRT model and the DINA model with a hierarchical attribute structure assumed. To allow for the data from the two psychometric frameworks to be comparable, the 2PL item parameters were transformed into the DINA model’s slip and guessing parameters, denoted by s_j and g_j , respectively, using the logistic-to-step transformation (LST), which employs the group-level expected misclassification indices to convert the 2PL model parameters into the DINA model parameters. Their simulation study revealed that, when highly diagnostic IRT-based data are retrofitted with CDM, and vice versa, comparable results can be obtained. They also found that the 2PL analysis of the IRT-based data resulted in small biases compared with CDM-based data analyzed using the DINA model, whereas the 2PL analysis of the CDM-based data yielded relatively large biases relative to the CDM analysis of the IRT-based data. Furthermore, in terms of item parameter estimation, large inaccuracies were found in retrofitting CDM data with the 2PL model. This could be attributed to the

differences between the IRFs of two the models—the 2PL model has 0 and 1 as its lower and upper asymptotes, respectively, whereas the DINA model has g_j and $1 - s_j$, which are typically greater than 0 and less than 1, respectively.

Meanwhile, Lee et al. (2012) investigated the relationships between CDM and IRT, as well as classical test theory (CTT) indices using empirical data, specifically the FCAT data. Their results found that items with low CTT-based discrimination index, and, to some extent, high IRT-based guessing parameter, can be expected to have low CDM-based discrimination index. Moreover, it was found that items deemed diagnostic in the CDM sense are slightly less difficult but highly discriminating in the CTT sense or more moderately difficult with lower guessing and higher discrimination parameters in the IRT sense.

Finally, de la Torre and Douglas (2004) fitted the higher-order version of the DINA model and 2PL model to fraction-subtraction data. The proficiencies estimated from the two models had a correlation of 0.96. This result points to the relationship that may exist between unidimensional IRT and CDM.

As a first foray, these studies provided interesting insights regarding the relationship between IRT models and CDMs. However, it is not clear that these findings have sufficient generalizability given the specific models considered, assumptions made, and data analyzed. For this reason, a more rigorous investigation, which includes establishing the mathematical relationship between unidimensional IRT models and CDMs, is needed.

20.3 Equivalence Between Unidimensional Item Response Theory and Higher-Order Cognitive Diagnosis Models

As with many approaches to latent variable modeling, a distinction between the measurement and structural components can be made in cognitive diagnosis modeling. The measurement component or the IRF, $p(x_j|\alpha_l)$, $j = 1, \dots, J$ and $l = 1, \dots, L$, where J and L are the test length and the number of attribute patterns, respectively, is represented by CDMs, whereas the structural component is represented by the joint distribution of the attributes, $p(\alpha_l)$. In the CDM specification, various formulations have been used to specify the joint distribution of the attributes. One formulation involves the use of a unidimensional higher-order latent trait θ , and an example of such a formulation is the higher-order DINA (HO-DINA; de la Torre & Douglas 2004) model. In this formulation, the elements of α_l are assumed to be conditionally independent given θ . Specifically, the joint distribution of α_l conditional on θ can be written as

$$p(\alpha_l|\theta) = \prod_{k=1}^K p_k(\alpha_{lk}|\theta), \quad (20.4)$$

where $p_k(\alpha_{lk}|\theta)$ is called the *attribute mastery function* (AMF). Taken together, the marginal probability of x_j can be written as

$$p(x_j) = \sum_{l=1}^L p(x_j|\alpha_l)p(\alpha_l) = \sum_{l=1}^L \int_{\theta} p(x_j|\alpha_l)p(\alpha_l|\theta)p(\theta)\partial\theta. \tag{20.5}$$

As previously stated, the probability of success on item j based on the G-DINA model can be expressed as a function of the reduced attribute vector, as in, $p(x_j|\alpha_{lj}) = p(x_j|\alpha_{lj}^*)$. Lemma 20.1 states that the property in Eq. 20.4 can be extended to α_{lj}^* .

Lemma 20.1 *Let α_{lj}^* be the reduced attributed vector for item j and θ be the proficiency parameter. The joint distribution of α_{lj}^* conditional on θ can be written as*

$$p(\alpha_{lj}^*|\theta) = \sum_{\alpha_{l(Kj^*+1)}=0}^1 \cdots \sum_{\alpha_{lK}=0}^1 p(\alpha_l|\theta) = \prod_{k=1}^{K_j^*} p_k(\alpha_{lk}|\theta), \tag{20.6}$$

where the summations are taken across the attributes that are not required for the item.

Proof By definition of a marginal probability, we have

$$p(\alpha_{lj}^*|\theta) = \sum_{\alpha_{l(Kj^*+1)}=0}^1 \cdots \sum_{\alpha_{lK}=0}^1 p(\alpha_l|\theta). \tag{20.7}$$

Using Eq. 20.4 yields the following:

$$\begin{aligned} p(\alpha_{lj}^*|\theta) &= \sum_{\alpha_{l(Kj^*+1)}=0}^1 \cdots \sum_{\alpha_{lK}=0}^1 \prod_{k=1}^K p_k(\alpha_{lk}|\theta) \\ &= \prod_{k=1}^{K_j^*} p_k(\alpha_{lk}|\theta) \sum_{\alpha_{l(Kj^*+1)}=0}^1 \cdots \sum_{\alpha_{lK}=0}^1 \prod_{k=K_j^*+1}^K p_k(\alpha_{lk}|\theta) \\ &= \prod_{k=1}^{K_j^*} p_k(\alpha_{lk}|\theta) \sum_{\alpha_{l(Kj^*+1)}=0}^1 \cdots \sum_{\alpha_{lK}=0}^1 \prod_{k=K_j^*+1}^{K-1} p_k(\alpha_{lk}|\theta) p_K(\alpha_{lK}|\theta) \\ &= \prod_{k=1}^{K_j^*} p_k(\alpha_{lk}|\theta) \end{aligned}$$

$$\begin{aligned} & \times \sum_{\alpha_{I(K_j^*+1)}=0}^1 \cdots \sum_{\alpha_{I(K-1)}=0}^1 \left(\prod_{k=K_j^*+1}^{K-1} p_k(\alpha_{Ik}|\theta) p_K(0|\theta) \right. \\ & \left. + \prod_{k=K_j^*+1}^{K-1} p_k(\alpha_{Ik}|\theta) p_K(1|\theta) \right). \end{aligned} \tag{20.8}$$

By factoring $\prod_{k=K_j^*+1}^{K-1} p_k(\alpha_{Ik}|\theta)$ out, and using the fact that $p_K(0|\theta) + p_K(1|\theta) = 1$, we get

$$\begin{aligned} p(\alpha_{I_j^*}^*|\theta) &= \prod_{k=1}^{K_j^*} p_k(\alpha_{Ik}|\theta) \sum_{\alpha_{I(K_j^*+1)}=0}^1 \cdots \sum_{\alpha_{I(K-1)}=0}^1 \prod_{k=K_j^*+1}^{K-1} p_k(\alpha_{Ik}|\theta) [p_K(0|\theta) + p_K(1|\theta)] \\ &= \prod_{k=1}^{K_j^*} p_k(\alpha_{Ik}|\theta) \sum_{\alpha_{I(K_j^*+1)}=0}^1 \cdots \sum_{\alpha_{I(K-1)}=0}^1 \prod_{k=K_j^*+1}^{K-1} p_k(\alpha_{Ik}|\theta). \\ &\quad \vdots \\ &= \prod_{k=1}^{K_j^*} p_k(\alpha_{Ik}|\theta). \end{aligned} \tag{20.9}$$

To compare unidimensional IRT models and CDMs, it would be necessary to express the CDM success probability on item j as a function of θ . This probability, $p(x_j|\theta)$, is simply

$$p(x_j|\theta) = \sum_{l=1}^{2^K} p(x_j, \alpha_l|\theta) = \sum_{l=1}^{2^K} p(x_j|\alpha_l) p(\alpha_l|\theta) = \sum_{l=1}^{2^{K_j^*}} p(x_j|\alpha_{I_j^*}^*) p(\alpha_{I_j^*}^*|\theta). \tag{20.10}$$

We can use Lemma 20.1 to breakdown $p(\alpha_{I_j^*}^*|\theta)$ into the marginal distributions of α_{Iks} and use the higher-order formulation for the CDMs. For greater generality, the G-DINA model is employed for $p(x_j|\alpha_{I_j^*}^*)$.

To understand the properties of $p(x_j|\theta)$, it would be helpful to re-express Eq. 20.10. For notational convenience, we can write $p_k(1|\theta)$ and $p_k(0|\theta)$ as p_k and $1 - p_k$, respectively. In addition, when there is no confusion, the item subscript j

can also be omitted. Specifically, we want to show that

$$p(x_j|\theta) = \delta_0 + \sum_{k=1}^{K_j^*} \delta_k p_k + \sum_{k=1}^{K_j^*-1} \sum_{k'=k+1}^{K_j^*} \delta_{kk'} p_k p_{k'} + \cdots + \delta_{1\dots K_j^*} \prod_{k=1}^{K_j^*} p_k. \quad (20.11)$$

When only one attribute is required, Eq. 20.11 simplifies to

$$\begin{aligned} p(x|\theta) &= \sum_{a_1=0}^1 p(x|a_1)p_1(a_1|\theta) = p(x|0)p_1(0|\theta) + p(x|1)p_1(1|\theta) \\ &= \delta_0 q_1 + (\delta_0 + \delta_1)p_1 = \delta_0 + \delta_1 p_1. \end{aligned} \quad (20.12)$$

Now, when two attributes are required, Eq. 20.11 can be written as

$$\begin{aligned} p(x|\theta) &= \sum_{a_1=0}^1 \sum_{a_2=0}^1 p(x|a_1, a_2)p(a_1, a_2|\theta) \\ &= \sum_{a_1=0}^1 \sum_{a_2=0}^1 p(x|a_1, a_2)p_1(a_1|\theta)p_2(a_2|\theta) \\ &= p(x|0, 0)p_1(0|\theta)p_2(0|\theta) + p(x|1, 0)p_1(1|\theta)p_2(0|\theta) \\ &\quad + p(x|0, 1)p_1(0|\theta)p_2(1|\theta) + p(x|1, 1)p_1(1|\theta)p_2(1|\theta). \end{aligned} \quad (20.13)$$

Using the fact that $P(x|0, 0) = \delta_0$, $P(x|1, 0) = \delta_0 + \delta_1$, $P(x|0, 1) = \delta_0 + \delta_2$, and $P(x|1, 1) = \delta_0 + \delta_1 + \delta_2 + \delta_{12}$ and after simplifying, we have

$$\begin{aligned} p(x|\theta) &= \delta_0 q_1 q_2 + (\delta_0 + \delta_1)p_1 q_2 + (\delta_0 + \delta_2)q_1 p_2 + (\delta_0 + \delta_1 + \delta_2 + \delta_{12})p_1 p_2 \\ &= \delta_0 + \delta_1 p_1 + \delta_2 p_2 + \delta_{12} p_1 p_2. \end{aligned} \quad (20.14)$$

To generalize this to K_j^* , we first define the following:

$$\begin{aligned} \mathcal{A} &= \{a_1, \dots, a_{K_j^*}\} \\ \mathcal{A}^{(-k)} &= \mathcal{A} \setminus \{a_k\} = \{a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_{K_j^*}\} \\ \mathcal{A}^{(-k, -k')} &= \mathcal{A} \setminus \{a_k, a_{k'}\} = \mathcal{A}^{(-k)} \cap \mathcal{A}^{(-k')}, \\ \mathcal{P} &= \{p_1, \dots, p_{K_j^*}\} \\ \mathcal{P}^{(-k)} &= \mathcal{P} \setminus \{p_k\} = \{p_1, \dots, p_{k-1}, p_{k+1}, \dots, p_{K_j^*}\} \\ \mathcal{P}^{(-k, -k')} &= \mathcal{P} \setminus \{p_k, p_{k'}\} = \mathcal{P}^{(-k)} \cap \mathcal{P}^{(-k')}, \end{aligned}$$

and $\delta_{(-k)} = \delta_{1 \dots (k-1)(k+1) \dots K_j^*}$. When K_j^* attributes are required for item j , we can write the general expression for $p(x_j|\theta)$ as

$$\begin{aligned}
 p(x_j|\theta) &= \sum_{a_1=0}^1 \cdots \sum_{a_{K_j^*}=0}^1 p(x_j|a_1, \dots, a_{K_j^*}) p(a_1, \dots, a_{K_j^*}|\theta) \\
 &= \delta_0 \sum_{\mathcal{A}} \prod_{\mathcal{A}, \mathcal{P}} p_k^{a_k} q_k^{1-a_k} + \sum_{k=1}^{K_j^*} \delta_k p_k \sum_{\mathcal{A}^{-k}} \prod_{\mathcal{A}^{-k}, \mathcal{P}^{-k}} p_{k'}^{a_{k'}} q_{k'}^{1-a_{k'}} \\
 &\quad + \sum_{k=1}^{K_j^*-1} \sum_{k'=k+1}^{K_j^*} \delta_{kk'} p_k p_{k'} \sum_{\mathcal{A}^{-k, -k'}} \prod_{\mathcal{A}^{-k, -k'}, \mathcal{P}^{-k, -k'}} p_{k''}^{a_{k''}} q_{k''}^{1-a_{k''}} + \cdots \\
 &\quad + \sum_{k=1}^{K_j^*} \delta_{(-k)} \prod_{k'=1}^{K_j^*} p_{k'} (p_k + q_k) / p_k + \delta_{1 \dots K_j^*} \prod_{k=1}^{K_j^*} p_k \\
 &= \delta_0 + \sum_{k=1}^{K_j^*} \delta_k p_k + \sum_{k=1}^{K_j^*-1} \sum_{k'=k+1}^{K_j^*} \delta_{kk'} p_k p_{k'} + \cdots + \delta_{1 \dots K_j^*} \prod_{k=1}^{K_j^*} p_k.
 \end{aligned} \tag{20.15}$$

We refer to Eq. 20.15 as the reformulated higher-order G-DINA (RHO-G-DINA) model. Note that for the RHO-G-DINA model to be a valid IRF, it should be monotonically nondecreasing as a function of θ .

20.3.1 Sufficient Conditions for a Monotonically Nondecreasing $p(x|\theta)$

For $p(x|\theta)$ to be monotonically nondecreasing, the following sufficient conditions need to be met.

1. The AMF of $p_k, k = 1, \dots, K$, is of the form

$$p_k = \frac{\exp[\zeta_k(\theta - \varphi_k)]}{1 + \exp[\zeta_k(\theta - \varphi_k)]}, \tag{20.16}$$

where ζ_k and φ_k represent the higher-order discrimination and difficulty parameters with respect to attribute k , respectively.

2. Monotonicity property should be satisfied. That is, $p(x|\alpha_l^*) \leq p(x|\alpha_{l'}^*)$ whenever $\alpha_l^* \preceq \alpha_{l'}^*$.

As defined by de la Torre (2011), $\alpha_l^* \preceq \alpha_{l'}^*$ means $\alpha_{lk} \leq \alpha_{l'k}$ for $k = 1, \dots, K_j^*$. For $K_j^* = 1$, $p(x|\alpha_l^*) \leq p(x|\alpha_{l'}^*)$ implies that $\delta_1 \geq 0$; for $K_j^* = 2$, it implies that $\delta_1, \delta_2 \geq 0$ and $\delta_{12} \geq \max(-\delta_1, -\delta_2)$; and so forth.

For notational convenience, we can assume that all the K attributes are required; however, the results are equally applicable to K_j^* . We also define $q_k = 1 - p_k$. To prove Theorem 20.1, we need Lemma 20.2, which states that the function g of p_1, \dots, p_K can simply be expressed as a product of the q_k s.

Lemma 20.2 For any positive integer K ,

$$\begin{aligned}
 g(p_1, \dots, p_K) &= 1 - \sum_{k=1}^K p_k + \sum_{k=1}^{K-1} \sum_{k'>k}^K p_k p_{k'} + \dots \\
 &\quad + (-1)^{K-1} \sum_{k=1}^K \prod_{k'=1}^K p_{k'}/p_k + (-1)^K \prod_{k=1}^K p_k \\
 &= \prod_{k=1}^K q_k.
 \end{aligned} \tag{20.17}$$

Proof When $K = 1$, $1 - p_1 = q_1$. For $K = 2$,

$$1 - p_1 - p_2 + p_1 p_2 = (1 - p_1)(1 - p_2) = q_1 q_2.$$

Thus, Lemma 20.2 holds for $K = 1$ and 2. Now, assume that this is true for $K - 1$. That is,

$$\begin{aligned}
 g(p_1, \dots, p_{K-1}) &= 1 - \sum_{k=1}^{K-1} p_k + \sum_{k=1}^{K-2} \sum_{k'>k}^{K-1} p_k p_{k'} + \dots \\
 &\quad + (-1)^{K-2} \sum_{k=1}^{K-1} \prod_{k'=1}^{K-1} p_{k'}/p_k + (-1)^{K-1} \prod_{k=1}^{K-1} p_k \\
 &= \prod_{k=1}^{K-1} q_k.
 \end{aligned} \tag{20.18}$$

To show that it is true for K , we have

$$\begin{aligned}
 g(p_1, \dots, p_K) &= 1 - \sum_{k=1}^K p_k + \sum_{k=1}^{K-1} \sum_{k'>k}^K p_k p_{k'} + \dots \\
 &\quad + (-1)^{K-1} \sum_{k=1}^K \prod_{k'=1}^K p_{k'}/p_k + (-1)^K \prod_{k=1}^K p_k \\
 &= 1 - \left(\sum_{k=1}^{K-1} p_k + p_K \right) + \left(\sum_{k=1}^{K-2} \sum_{k'>k}^{K-1} p_k p_{k'} + p_K \sum_{k=1}^{K-1} p_k \right) \\
 &\quad + \dots + \left((-1)^{K-1} \prod_{k=1}^{K-1} p_k + (-1)^{K-1} p_K \sum_{k=1}^{K-1} \prod_{k'=1}^{K-1} p_{k'}/p_k \right) \\
 &\quad + (-1)^K p_K \prod_{k=1}^{K-1} p_k.
 \end{aligned}
 \tag{20.19}$$

Grouping together similar terms yields the following expression:

$$\begin{aligned}
 g(p_1, \dots, p_K) &= \left(1 - \sum_{k=1}^K p_k + \sum_{k=1}^{K-1} \sum_{k'>k}^K p_k p_{k'} + \dots + (-1)^K \prod_{k=1}^K p_k \right) \\
 &\quad - p_K \left(1 - \sum_{k=1}^K p_k + \sum_{k=1}^{K-1} \sum_{k'>k}^K p_k p_{k'} + \dots + (-1)^K \prod_{k=1}^K p_k \right).
 \end{aligned}
 \tag{20.20}$$

Using the assumption in Eq. 20.18 will produce the desired result, as in,

$$g(p_1, \dots, p_K) = g(p_1, \dots, p_{K-1}) - p_K g(p_1, \dots, p_{K-1}) = q_K \prod_{k=1}^{K-1} q_k = \prod_{k=1}^K q_k.
 \tag{20.21}$$

Therefore, we have shown that Lemma 20.2 holds true for any K .

To show that $p(x_j|\theta)$ is monotonically nondecreasing, we need to show that $\partial p(x_j|\theta)/\partial \theta \geq 0$. Note that when $K_j^* = 1$, the derivative is equal to

$$\frac{\partial p(x_j|\theta)}{\partial \theta} = \frac{\partial (\delta_0 + \delta_1 p_1)}{\partial \theta} = \delta_1 \zeta_1 p_1 q_1 \geq 0,
 \tag{20.22}$$

which is always nonnegative because $\delta_1 \geq 0, \zeta_1 > 0$, and $0 \leq p_1 \leq 1$. For $K_j^* = 2$, we assume that $\delta = \delta_1 \leq \delta_2$. Then, the derivative is equal to

$$\begin{aligned} \frac{\partial p(x_j|\theta)}{\partial \theta} &= \frac{\partial(\delta_0 + \delta_1 p_1 + \delta_2 p_2 + \delta_{12} p_1 p_2)}{\partial \theta} \\ &= \delta_1 \zeta_1 p_1 q_1 + \delta_2 \zeta_2 p_2 q_2 + \delta_{12} [p_1 p_2 (\zeta_1 q_1 + \zeta_2 q_2)]. \end{aligned} \tag{20.23}$$

Without loss of generality, we assume that $\delta = \delta_1 \leq \delta_2$ so that $\max(-\delta_1, -\delta_2) = -\delta$ implies that $\delta_{12} \geq -\delta$. Hence,

$$\begin{aligned} \frac{\partial p(x_j|\theta)}{\partial \theta} &\geq \delta [\zeta_1 p_1 q_1 + \zeta_2 p_2 q_2 - (p_1 p_2 (\zeta_1 q_1 + \zeta_2 q_2))] \\ &= \delta [q_1 q_2 (\zeta_1 p_1 + \zeta_2 p_2)] \geq 0. \end{aligned} \tag{20.24}$$

Again, without loss of generality, we can assume that $\zeta_k = \zeta$ for all k . Because ζ and δ are always assumed to be positive, they can be dropped from Eq. 20.24. Therefore, in general, showing $\partial p(x_j|\theta)/\partial \theta \geq 0$ is equivalent to showing that Theorem 20.1 is true.

Theorem 20.1 *For any K ,*

$$\begin{aligned} f(p_1, \dots, p_K) &= \sum_{k=1}^K p_k q_k - \sum_{k=1}^{K-1} \sum_{k'>k}^K p_k p_{k'} (q_k + q_{k'}) \\ &\quad + \sum_{k=1}^{K-2} \sum_{k'>k}^{K-1} \sum_{k''>k'}^K p_k p_{k'} p_{k''} (q_k + q_{k'} + q_{k''}) + \dots \\ &\quad + (-1)^{K+1} \prod_{k=1}^K p_k \sum_k^K q_k \\ &= \sum_{k=1}^K p_k \prod_{k=1}^K q_k \geq 0. \end{aligned} \tag{20.25}$$

Theorem 20.1 can be proved using mathematical induction. Previously, we have verified that the theorem holds for $K = 1$ and 2 . We now assume that Theorem 20.1 holds true for $K - 1$. That is,

$$\begin{aligned} f(p_1, \dots, p_{K-1}) &= \sum_{k=1}^{K-1} p_k q_k - \sum_{k=1}^{K-2} \sum_{k'>k}^{K-1} p_k p_{k'} (q_k + q_{k'}) \\ &\quad + \sum_{k=1}^{K-3} \sum_{k'>k}^{K-2} \sum_{k''>k'}^{K-1} p_k p_{k'} p_{k''} (q_k + q_{k'} + q_{k''}) + \dots \end{aligned}$$

$$\begin{aligned}
& + (-1)^K \prod_{k=1}^{K-1} p_k \sum_{k=1}^{K-1} q_k \\
& = \sum_{k=1}^{K-1} p_k \prod_{k=1}^{K-1} q_k \geq 0.
\end{aligned} \tag{20.26}$$

We can then rewrite the $f(p_1, \dots, p_K)$ in Eq. 20.25 as

$$\begin{aligned}
& \left(\sum_{k=1}^{K-1} p_k q_k + p_K q_K \right) - \left(\sum_{k=1}^{K-2} \sum_{k'=k+1}^{K-1} p_k p_{k'} (q_k + q_{k'}) + \sum_{k=1}^{K-1} p_k p_K (q_k + q_K) \right) \\
& + \left(\sum_{k=1}^{K-3} \sum_{k'=1}^{K-2} \sum_{k''=k+1}^{K-1} p_k p_{k'} p_{k''} (q_k + q_{k'} + q_{k''}) + \sum_{k=1}^{K-2} \sum_{k'=k+1}^{K-1} p_k p_{k'} p_K (q_k + q_{k'} + q_K) \right) \\
& + \dots + (-1)^K \left[\prod_{k=1}^{K-1} p_k \sum_{k=1}^{K-1} q_k + \sum_{k=1}^{K-1} \prod_{k'}^K p_{k'1} / p_k \left(\sum_{k'=1}^K q_{k'} - q_k \right) \right] \\
& + (-1)^{K+1} \prod_{k=1}^{K-1} p_k p_K \left(\sum_{k=1}^{K-1} q_k + q_K \right).
\end{aligned} \tag{20.27}$$

We can group the terms of Eq. 20.27 as follows:

$$\begin{aligned}
& \left[\sum_{k=1}^{K-1} p_k q_k - \sum_{k=1}^{K-2} \sum_{k'=k+1}^{K-1} p_k p_{k'} (q_k + q_{k'}) \right. \\
& + \sum_{k=1}^{K-3} \sum_{k'=1}^{K-2} \sum_{k''=k+1}^{K-1} p_k p_{k'} p_{k''} (q_k + q_{k'} + q_{k''}) + \dots + (-1)^K \prod_{k=1}^{K-1} p_k \sum_{k=1}^{K-1} q_k \left. \right] \\
& + \left[p_K q_K - \sum_{k=1}^{K-1} p_k p_K (q_k + q_K) + \sum_{k=1}^{K-2} \sum_{k'=k+1}^{K-1} p_k p_{k'} p_K (q_k + q_{k'} + q_K) + \dots \right. \\
& + (-1)^K \sum_{k=1}^{K-1} \prod_{k'=1}^K p_{k'1} / p_k \left(\sum_{k'=1}^K q_{k'} - q_k \right) + (-1)^{K+1} \prod_{k=1}^{K-1} p_k p_K \left(\sum_{k=1}^{K-1} q_k + q_K \right) \left. \right].
\end{aligned} \tag{20.28}$$

Using the definition of $f(p_1, \dots, p_{K-1})$ in Eq. 20.26, we can simplify Eq. 20.28 as

$$\begin{aligned}
 & f(p_1, \dots, p_{K-1}) + p_K q_K - \left(p_K \sum_{k=1}^{K-1} p_k q_k - p_K q_K \sum_{k=1}^{K-1} p_k \right) \\
 & + \left(p_K \sum_{k=1}^{K-2} \sum_{k'=k+1}^{K-1} p_k p_{k'} (q_k + q_{k'}) + p_K q_K \sum_{k=1}^{K-2} \sum_{k'=k+1}^{K-1} p_k p_{k'} \right) + \dots \\
 & + (-1)^K \left[p_K \sum_{k=1}^{K-1} \prod_{k'=1}^{K-1} p_{k'1}/p_k \left(\sum_{k'=1}^{K-1} q_{k'} - q_k \right) + p_K q_K \sum_{k=1}^{K-1} \prod_{k'=1}^{K-1} p_{k'1}/p_k \right] \\
 & + (-1)^{K+1} \left(p_K \prod_{k=1}^{K-1} p_k \sum_{k=1}^{K-1} q_k + p_K q_K \prod_{k=1}^{K-1} p_k \right). \tag{20.29}
 \end{aligned}$$

This can be further simplified as

$$\begin{aligned}
 & f(p_1, \dots, p_{K-1}) - p_K \left[\sum_{k=1}^{K-1} p_k q_k - \sum_{k=1}^{K-2} \sum_{k'=k+1}^{K-1} p_k p_{k'} (q_k + q_{k'}) + \dots \right. \\
 & \left. + (-1)^{K-1} \sum_{k=1}^{K-1} \prod_{k'=1}^{K-1} p_{k'1}/p_k \left(\sum_{k'=1}^{K-1} q_{k'} - q_k \right) + (-1)^K \prod_{k=1}^{K-1} p_k \sum_{k=1}^{K-1} q_k \right] \\
 & + p_K q_K \left(1 - \sum_{k=1}^{K-1} p_k + \sum_{k=1}^{K-2} \sum_{k'=k+1}^{K-1} p_k p_{k'} + \dots \right. \\
 & \left. + (-1)^K \sum_{k=1}^{K-1} \prod_{k'=1}^{K-1} p_{k'1}/p_k + (-1)^{K+1} \prod_{k=1}^{K-1} p_k \right) \tag{20.30}
 \end{aligned}$$

Using the definitions of $g(p_1, \dots, p_{K-1})$ in Eq. 20.18 based on Lemma 20.2 and $f(p_1, \dots, p_{K-1})$ in Eq. 20.26, Eq. 20.30 can be written as

$$\begin{aligned}
 f(p_1, \dots, p_K) &= [f(p_1, \dots, p_{K-1}) - p_K f(p_1, \dots, p_{K-1})] + p_K q_K g(p_1, \dots, p_{K-1}) \\
 &= (1 - p_K) f(p_1, \dots, p_{K-1}) + p_K q_K g(p_1, \dots, p_{K-1}) \\
 &= q_K [f(p_1, \dots, p_{K-1}) + p_K g(p_1, \dots, p_{K-1})] \\
 &= q_K \left[\sum_{k=1}^{K-1} p_k \prod_{k=1}^{K-1} q_k + p_K \prod_{k=1}^{K-1} q_k \right]
 \end{aligned}$$

$$\begin{aligned}
 &= q_K \prod_{k=1}^{K-1} q_k \left(\sum_{k=1}^{K-1} p_k + p_K \right) \\
 &= \sum_{k=1}^K p_k \prod_{k=1}^K q_k.
 \end{aligned} \tag{20.31}$$

Because p_k s and q_k s for all k are probabilities, then $f(p_1, \dots, p_K)$ is nonnegative. This completes the proof of Theorem 20.1. Note that Theorem 20.1 states that, as long as the AMF of p_k is of the form Eq. 20.16 and the monotonicity property is satisfied, $p(x_j|\theta)$, based on RHO-G-DINA model, is a monotonically nondecreasing function of θ .

20.3.2 Special Cases of $p(x|\theta)$

In this subsection, we examine some special cases of $p(x_j|\theta)$ based on the RHO-G-DINA model. As noted earlier, when $K_j^* = 1$, Eq. 20.11 reduces to

$$p(x|\theta) = \delta_0 + \delta_1 p_1, \tag{20.32}$$

which is equivalent to the 4PL IRT model in Eq. 20.1 with the upper asymptote $v = \delta_0 + \delta_1$ and the guessing parameter $\gamma = \delta_0$. It reduces to the 3PL IRT model when $v = \delta_0 + \delta_1 = 1$ with $\gamma = \delta_0$. When $\delta_1 = 1$ and $\delta_0 = 0$, it simplifies to the 2PL or 1PL IRT model, depending on the values of ζ_{ks} . Because the RHO-G-DINA model is saturated when $K_j^* = 1$, $p(x|\theta)$ is already in its simplest form (i.e., no other specific CDM can be considered). Note that this equivalence is not surprising because there is only one required attribute (i.e., unidimensional case).

Now, we examine the case when $K_j^* \geq 2$. Again, it has been shown that when $K_j^* = 2$, Eq. 20.11 is equal to

$$p(x_j|\theta) = \delta_0 + \delta_1 p_1 + \delta_2 p_2 + \delta_{12} p_1 p_2. \tag{20.33}$$

Similarly, when $K_j^* = 3$, $p(x_j|\theta)$ reduces to

$$\begin{aligned}
 p(x_j|\theta) &= \delta_0 + \delta_1 p_1 + \delta_2 p_2 + \delta_3 p_3 \\
 &\quad + \delta_{12} p_1 p_2 + \delta_{13} p_1 p_3 + \delta_{23} p_2 p_3 + \delta_{123} p_1 p_2 p_3.
 \end{aligned} \tag{20.34}$$

To illustrate how $p(x_j|\theta)$ behaves for $K_j^* = 2$ and 3, p_1 , p_2 , and p_3 are plotted together with $p(x_j|\theta)$ in Fig. 20.1. The curve of each AMF, which follows a logistic function, is to be expected because of its assumed form. However, because the IRF of the RHO-G-DINA model is a function of multiple AMFs, it resembles but does not strictly follow a logistic function. Thus, provided that the sufficient conditions are met, the RHO-G-DINA model IRF may be approximated by logistic function.

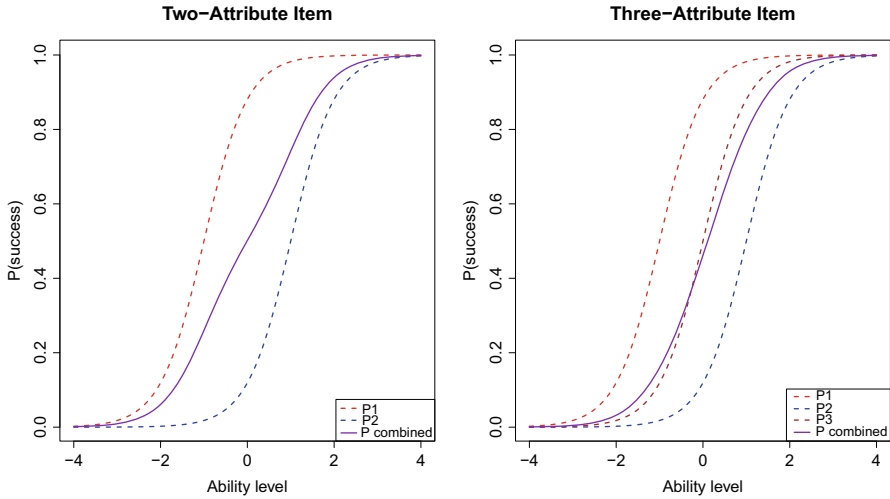


Fig. 20.1 Attribute mastery probabilities p_k s and item success probability $p(x_j|\theta)$ conditional on ability level θ for $K_j^* = 2$ and 3

However, the exact conditions under which an IRT model can approximate the RHO-G-DINA IRF needs to be explored. Additionally, the extent to which the estimated θ obtained when unidimensional IRT models are fitted to CDM-based data will correspond to the RHO-G-DINA-generated θ remains to be seen.

20.4 A Simulation Study

In this section, a small simulation study was conducted to further examine the relationship of the two psychometric frameworks. The primary objective of this simulation study is to investigate the impact of a number of factors on the quality of the approximation of the RHO-G-DINA model IRF and the correlation between the true and estimated proficiencies.

20.4.1 Design and Analysis

In this simulation study, the item responses were generated from the RHO-G-DINA model, with the test length fixed to $J = 50$, the number of attributes to $K = 10$, and the sample size to $N = 10,000$. For the slope, ζ_k were set to either 0.5 or 2.5, for all k , to represent low or high attribute discrimination parameters, respectively. The intercepts φ_k s were selected from the intervals $(-2.5, 2.5)$ or $(-3.5, 3.5)$ with

equal increments; for the item quality, the lowest and one minus the highest success probabilities $P(\mathbf{0})$ and $1 - P(\mathbf{1})$ were set to 0.0, 0.1, or 0.2. Finally, the proficiency parameter θ was generated from the standard normal distribution. Each attribute was measured by 11 test items, and the test contained 10, 20 and 20 one-, two-, and three-attribute items, respectively.

The IRF of the RHO-G-DINA model was approximated using the 2PL, 3PL, or 4PL IRT model. For each IRT model, the item parameters were chosen by minimizing the squared weighted Euclidean distance between the IRFs of the IRT and RHO-G-DINA models, where the weights were obtained from the standard normal density function. The mean of the weighted Euclidean distances across the 50 items were then calculated. In addition to the IRF comparison, the three unidimensional IRT models were fitted to the HO-G-DINA-generated data to examine the correspondence between the IRT-estimated θ and the true as well as estimated higher-order θ . The `GDINA` (Ma & de la Torre 2020) and `mirr` (Chalmers 2012) R packages were used in the simulation study.

20.4.2 Results

Table 20.1 presents the mean of the weighted Euclidean distance between the fitted IRT model and the true RHO-G-DINA model IRFs. The simulation results indicated that even the simplest IRT model under consideration (i.e., the 2PL) can provide a good approximation to the RHO-G-DINA model when $P(\mathbf{0})$ and $1 - P(\mathbf{1})$ are

Table 20.1 Mean of the weighted Euclidean distance between the fitted IRT and the RHO-G-DINA model IRFs

Discrimination	Location	$P(\mathbf{0}), 1 - P(\mathbf{1})$	Fitted model		
			2PL	3PL	4PL
High	Narrow	0.0	0.005	0.005	0.005
		0.1	0.049	0.035	0.003
		0.2	0.085	0.074	0.002
	Wide	0.0	0.007	0.007	0.007
		0.1	0.050	0.037	0.005
		0.2	0.084	0.070	0.004
Low	Narrow	0.0	0.002	0.001	0.001
		0.1	0.012	0.007	0.001
		0.2	0.015	0.009	0.001
	Wide	0.0	0.002	0.002	0.001
		0.1	0.011	0.006	0.000
		0.2	0.013	0.008	0.001

Notes: The discrimination parameter was set to either 2.5 (high) or 0.5 (low); the location parameter was obtained from either $(-2.5, 2.5)$ (narrow) or $(-3.5, 3.5)$ (wide); $P(\mathbf{0})$ and $1-P(\mathbf{1})$ are the lowest and highest success probabilities

both zero. However, as expected, when the success probabilities were increased to 0.1 and 0.2, the 2PL approximation got poorer due to its inability to approximate the lower and upper asymptotes of the RHO-G-DINA model IRF. As can be seen from Table 20.1, better approximations were obtained when more complex IRT models (i.e., 3PL and 4PL) were used. It was not surprising that the lower and upper asymptotes of the 4PL IRT model allowed it to provide the best approximation of the RHO-G-DINA model IRFs. It can be noted that although a smaller value of ζ_k led to better results, the impact of the range of φ_k was not very clear, at least when $\zeta_k = 2.5$.

Figure 20.2 displays the 4PL IRT model approximations of the RHO-G-DINA model IRFs for $K_j^* = 1, 2,$ and 3 when $\zeta_k = 2.5$, φ_k ranged from -2.5 to 2.5 , and $P(\mathbf{0})$ and $1 - P(\mathbf{1})$ were both 0.2 . For these specific conditions, the CDM and IRT curves are virtually indistinguishable, which is an indication that the IRF of the RHO-G-DINA model can be well approximated by an IRT model.

The range of the intercept did not have a substantial impact on the correlation between the true and estimated proficiencies; hence, only results for $\varphi_k \in (-2.5, 2.5)$ are presented. Table 20.2 gives the correlations between the true (i.e., RHO-G-DINA-based) and the estimated proficiencies. As a baseline, the correlation between θ and $\hat{\theta}$ using the RHO-G-DINA model was obtained, and the correlation was at least 0.89 when the attribute discrimination parameter was high (i.e., $\zeta_k = 2.5$). When the unidimensional IRT models were fitted to the data, the $\hat{\theta}$ estimates had correlations that were only slightly lower than those from the RHO-G-DINA estimates. However, with low attribute distribution parameter (i.e., $\zeta_k = 0.5$), the correlations between true and RHO-G-DINA estimated proficiencies dropped dramatically to as low as 0.46 . The corresponding correlations between true and IRT estimated proficiencies were also much lower.

To examine the correspondence between IRT and CDM, the correlation between unidimensional IRT and RHO-G-DINA proficiency estimates is given in Table 20.3. The two estimates were highly correlated. In particular, the correlation was at least 0.97 when $\zeta\varphi_k = 2.5$. Thus, when the attributes are discriminating, unidimensional

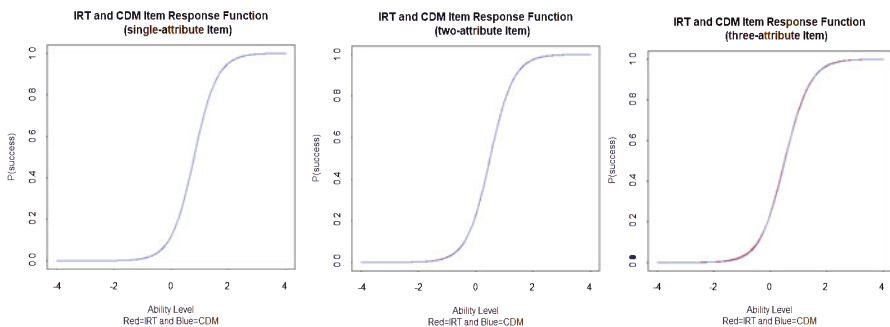


Fig. 20.2 Exact CDM (reformulated higher-order G-DINA model) and approximate IRT (four-parameter logistic model) item success probabilities conditional on ability level θ for $K_j^* = 1, 2,$ and 3 , $\zeta_k = 2.5$, and $P(\mathbf{0}) = 1 - P(\mathbf{1}) = 0.2$

Table 20.2 Correlation between θ and $\hat{\theta}$ with difficulty parameter $\varphi_k \in (-2.5, 2.5)$

Disc.	$P(\mathbf{0}), 1 - P(\mathbf{1})$	True model	Fitted model		
		RHO-G-DINA	2PL	3PL	4PL
$\zeta = 2.5$	0.0	0.92	0.91	0.91	0.91
	0.1	0.91	0.90	0.89	0.89
	0.2	0.89	0.87	0.87	0.88
$\zeta = 0.5$	0.0	0.53	0.44	0.45	0.42
	0.1	0.51	0.45	0.45	0.42
	0.2	0.46	0.42	0.43	0.40

Note: Disc. = discrimination parameter

Table 20.3 Correlation between RHO-G-DINA and IRT $\hat{\theta}$ s with difficulty parameter $\varphi_k \in (-2.5, 2.5)$

Disc.	$P(\mathbf{0}), 1 - P(\mathbf{1})$	Fitted model		
		2PL	3PL	4PL
$\zeta = 2.5$	0.0	0.98	0.98	0.98
	0.1	0.98	0.97	0.98
	0.2	0.97	0.98	0.98
$\zeta = 0.5$	0.0	0.82	0.85	0.79
	0.1	0.89	0.89	0.82
	0.2	0.90	0.93	0.86

Note: Disc. = discrimination parameter

IRT models provide a good approximation of RHO-G-DINA $\hat{\theta}$. This has a practical implication when K is large—as the number of attributes increases, it becomes more computationally challenging to fit a HO-CDM because the number of latent classes grows exponentially; however, this is not the case with unidimensional IRT models; thus, it can provide a good approximation when the primary interest is to estimate the proficiency levels of the examinees.

20.5 Real Data Example

Apart from a simulation study, we also conducted an IRT and CDM analysis of a PR dataset. The goal of this analysis is to explore the relationship between the two psychometric frameworks using real educational assessment data.

20.5.1 Proportional Reasoning Test

20.5.1.1 Data

Tjoe and de la Torre (2014) developed a PR test to measure the proportional reasoning skills of middle school students. The test measures eight attributes,

namely, a_1 , prerequisite skills and concepts (e.g., basic arithmetic operations, finding the greatest common factors, reducing fractions); a_2 , comparing two fractions; a_3 , ordering three or more fractions; a_4 , constructing ratios; a_5 , constructing proportions; a_6 , identifying a multiplicative relationship between sets of values; a_7 , differentiating a proportional relationship from a non-proportional relationship; and a_8 , applying algorithms in solving PR problems (e.g., cross-multiplication algorithm, building-up/down strategy). The responses of 807 middle school students from two different schools in the United States to 31 PR items were analyzed in this chapter to illustrate the relationship between the two psychometric frameworks.

20.5.1.2 Method

In a recent analysis, Ma et al. (2020) fitted several CDMs and the 3PL model to the PR data. In this analysis, we fitted three CDMs (i.e., saturated and 1PL and 2PL higher-order G-DINA models) and four unidimensional IRT models (i.e., 4PL, 3PL, 2PL, and 1PL) to the same data. The Akaike information criterion (AIC; Akaike 1974) and Bayesian information criterion (BIC; Schwarz 1978) were used to evaluate the relative fit of the aforementioned models. Moreover, the correlations among the different $\hat{\theta}$ s obtained were calculated to determine the extent of the correspondence between the different models. To further compare and contrast the IRT and CDM estimates, the number of mastered attributes was also plotted against the proficiency estimates. Again, the GDINA (Ma & de la Torre 2020) and `mirt` (Chalmers 2012) R packages were used in the analysis of the PR data.

20.5.1.3 Results

Table 20.4 shows the AIC and BIC of the fitted IRT and CDM models. Among the IRT models, the 3PL and 2PL models were preferred because they obtained the lowest AIC and BIC, respectively; for the CDMs, the 2PL-G-DINA model was preferred based on AIC and BIC, followed by the 1PL-G-DINA model. Table 20.5 displays the correlations between the HO-G-DINA and IRT proficiency estimates. For the models preferred, the correlations between the two sets of proficiency

Table 20.4 Model fit evaluation for the PR test data

Framework	Model	AIC	BIC
IRT	4PL	27,865.39	28,447.36
	3PL	27,834.06	28,270.54
	2PL	27,836.10	28,127.08
	1PL	28,278.97	28,429.16
CDM	Sat. G-DINA	28,246.16	30,419.17
	2PL-G-DINA	27,988.05	29,039.35
	1PL-G-DINA	28,060.15	29,078.60

Table 20.5 Correlation between the HO-G-DINA-based $\hat{\theta}$ and the IRT-based $\hat{\theta}$

IRT	CDM	
	2PL-G-DINA	1PL-G-DINA
4PL	0.96	0.96
3PL	0.96	0.96
2PL	0.96	0.96
1PL	0.94	0.95

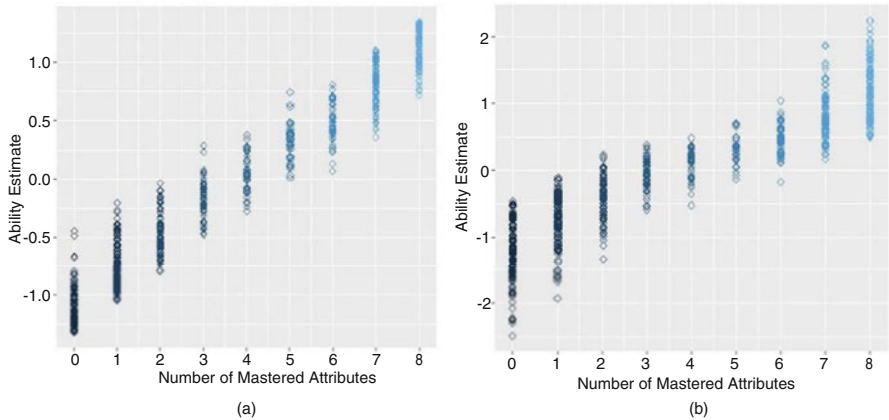


Fig. 20.3 (a) Plot of 2PL-G-DINA $\hat{\theta}$ versus number of attributes mastered. (b) Plot of 2PL $\hat{\theta}$ versus number of attributes mastered

estimates were very high (i.e., 0.96). It should be noted that this correlation is consistent with what de la Torre and Douglas (2004) found in analyzing a fraction-subtraction dataset.

Figure 20.3 plots the $\hat{\theta}$ based on the 2PL-G-DINA and 2PL IRT models against the number of attributes mastered. The two plots show the same general trend—students with higher proficiencies also had mastered more attributes. However, it can also be seen that a fixed number of attributes mastered can correspond to a wide range of (and overlapping) proficiency estimates. That is, some individuals with higher proficiencies had fewer number of attributes mastered. This finding suggests that targeted remediation cannot be solely based on proficiency estimates. The overlaps also indicate that for a fixed proficiency level, students can have different numbers of attributes mastered. For instance, students with $\hat{\theta} = 0.0$ based on the 2PL-G-DINA model can master three to five attributes; similarly, students with a relatively high estimated proficiency (i.e., $\hat{\theta} > 1.0$) still can have a deficiency. Finally, although the two estimates had a high correlation (i.e., 0.96), it can be noted that their ranges differ—estimates based on the 2PL have a larger variability. This suggests that additional adjustments may be needed to put the two estimates on the same scale.

20.6 Discussion

CDMs can provide diagnostic feedback that can inform instruction and learning, and their potential can be maximized when used in conjunction with CDAs. However, developing CDAs from scratch requires substantial time and resources. For this reason, it is not uncommon for unidimensional IRT models and CDMs to be treated interchangeably, as in, they are fitted to the same assessment data. However, without the necessary framework to relate the two classes of models, the validity of inferences from such analyses can be called into question.

In this work, certain conditions were imposed to reformulate the HO-G-DINA model and express its success probability as a function of a higher-order latent trait θ . Based on this reformulation, a framework for relating the two classes of psychometric models, as well as boundaries when this can be done, is provided. It has been shown that when the attributes follow a higher-order structure and the AMF slope is sufficiently large, a correspondence between unidimensional IRT models and CDMs can be established. Under these conditions, estimating both finer-grained attributes and an overall proficiency from the same data is deemed reasonable. Results from analyzing simulated and real data indicate that IRT and CDM can provide highly correlated estimates of the latent trait. However, estimating the latent trait estimate alone would be insufficient to provide the finer-grained information necessary to differentiate individuals based on their attribute profiles.

This chapter also suggests that a higher-order attribute structure alone may not be sufficient to establish a correspondence between unidimensional IRT models and CDMs. Specifically, when the AMF slope is low, the resulting data may be too multidimensional for a unidimensional IRT model to adequately fit. Thus, not all data can be fitted IRT models and CDMs simultaneously. The opposite problem—the data may be too unidimensional—is an equally important issue worth discussing. It remains to be seen whether additional information can be gained by fitting CDMs to highly unidimensional data. If anything can be gleaned from a number of existing examples, retrofitting CDMs to unidimensional IRT-based assessment data may not lead to appreciable diagnostic insights. For data to be simultaneously appropriate for both the unidimensional IRT models and the multidimensional CDMs, they need to follow the Goldilocks principle—they have to have just the right dimensionality, not too unidimensional, yet not too multidimensional.

The framework discussed in this chapter solely focuses on relating CDMs to parametric IRT; thus, exploring the relationship between CDMs and nonparametric IRT is also of interest (see Chap. 10 of this book). The seminal work on this topic was done by Junker and Sijtsma (2001). However, their work included only two constrained CDMs—the DINA and noisy input, deterministic, “and” gate models. In addition, as Sijtsma and Van der Ark (2021) noted, a number of outstanding issues remain in this area (e.g., deriving the stochastic ordering of the latent trait by means of the sum scores property for the DINA model). Aside from considering a wider class of CDMs, future research should examine the extent to which using a higher-order rather than a saturated attribute distribution can facilitate the understanding of how CDMs are related to nonparametric IRT.

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>
- Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1), 1–8. <https://doi.org/10.1002/j.2333-8504.1981.tb01255.x>
- Birenbaum, M., Tatsuoka, C., & Xin, T. (2005). Large-scale diagnostic assessment: Comparison of eighth graders' mathematics performance in the United States, Singapore and Israel. *Assessment in Education Principles Policy and Practice*, 12(2), 167–181. <https://doi.org/10.1080/09695940500143852>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's proficiency. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley Publishing.
- Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, 33(1), 2–14. <https://doi.org/10.1111/emip.12020>
- Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Chen, J. & de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: The case of the programme for international student assessment in reading. *Psychology*, 5(18), 1967–1978. <https://doi.org/10.4236/psych.2014.518200>
- Choi, K. M., Lee, Y.-S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science, & Technology Education*, 11(6), 1563–1577. <https://doi.org/10.12973/eurasia.2015.1421a>
- de la Torre, J. (2006). *Skills profile comparisons at the state level: An application and extension of cognitive diagnosis modeling in NAEP*. Presentation at the International Meeting of the Psychometric Society, Montreal, Canada.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130. <https://doi.org/10.3102/1076998607309474>
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179–199. <https://doi.org/10.1007/s11336-011-9207-7>
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333–353. <https://doi.org/10.1007/BF02295640>
- de la Torre, J. & Karelitz, T. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement*, 46(4), 450–469. <https://doi.org/10.1111/j.1745-3984.2009.00092.x>
- de la Torre, J. & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educativa*, 20(2), 89–97. <https://doi.org/10.1016/j.pse.2014.11.001>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.
- Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30(5), 394–411. <https://doi.org/10.1177/0146621606288554>
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26(4), 301–323. <https://doi.org/10.1111/j.1745-3984.1989.tb00336.x>

- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272. <https://doi.org/10.1177/01466210122032064>
- Lee, Y.-S., de la Torre, J. & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: an empirical investigation. *Asia Pacific Education Review*, 13(2), 333–345. <https://doi.org/10.1007/s12564-011-9196-3>
- Lee, Y.-S., Park, Y. S. & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, 11(2), 144–177. <https://doi.org/10.1080/15305058.2010.534571>
- Liu, R. Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, 78(3), 357–383. <https://doi.org/10.1177/0013164416685599>
- Ma W. & de la Torre J (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93(14), 1–26. <https://doi.org/10.18637/jss.v093.i14>
- Ma, W., Minchen, N. & de la Torre, J. (2020). Choosing between CDM and unidimensional IRT: The proportional reasoning test case, *Measurement: Interdisciplinary Research and Perspectives*, 18(2), 87–96. <https://doi.org/10.1080/15366367.2019.1697122>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Sijtsma, K., & Van der Ark, L. A. (2021). *Measurement models for psychological attributes*. CRC Press. <https://doi.org/10.1201/9780429112447>
- Tatsuoka, K. K., Corter, J., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, 41(4), 901–906. <https://doi.org/10.3102/00028312041004901>
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11(3), 287–305. <https://doi.org/10.1037/1082-989X.11.3.287>
- Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, 26(2), 237–255. <https://doi.org/10.1007/s13394-013-0090-7>

Chapter 21

A Sparse Latent Class Model for Polytomous Attributes in Cognitive Diagnostic Assessments



Siqi He, Steven Andrew Culpepper, and Jeff Douglas

Abstract Diagnostic models (DMs) have been widely applied to binary response data. However, in the field of educational and psychological measurement, a wealth of ordinal data are collected to measure latent structures where the traditional binary attributes may not adequately describe the complex response patterns. Considering that, we propose an extension of the sparse latent class model (SLCM) with ordinal attributes, with the purpose of fully exploring the relationships between attributes and response patterns. Furthermore, we discuss the strict and generic identifiability conditions for the ordinal SLCMs. We apply the model to the Short Dark Triad data and revisit the underlying personality structure. Evidence supports that SLCMs have better model fit to this real data than the exploratory factor models. We also confirm the efficiency of a Gibbs algorithm in recovering the empirical item parameters via a Monte Carlo simulation study. This study discusses a way of constructing DMs with ordinal attributes which helps broaden its applicability to personality assessment.

21.1 Introduction

Cognitive diagnostic assessments, aimed at providing fine-grained information about respondents' mastery of latent attributes, have gained increasing research attention in recent decades. The considerable expansion of cognitive diagnostic models (CDMs) has heightened the need for an inclusive and comprehensive modeling approach, where the sparse latent class models (SLCM; Chen et al. 2020) have served this purpose to fit most existing CDMs in an exploratory fashion.

The SLCM was originally proposed with binary attributes, including the deterministic input, noisy, "and" gate model (DINA; De La Torre 2009; Junker & Sijtsma 2001); the deterministic input, noisy, "or" gate model (DINO; J. L. Templin & Henson, 2006); the reduced non-compensatory reparameterized unified model

S. He · S. A. Culpepper · J. Douglas (✉)
University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: siqihe2@illinois.edu; sculpepp@illinois.edu; jeffdoug@illinois.edu

(NC-RUM; DiBello et al. 1995; Templin et al. 2010) and the compensatory-RUM model (C-RUM; Hagenaaers 1993; Maris 1999); and the family of general diagnostic models (De La Torre 2011; Henson et al. 2009; von Davier 2005). With binary representations of latent attributes, examinees' attribute patterns are composed of either "mastery" or "non-mastery." However, binary attributes are sometimes not accurate enough to describe the level of mastery as respondents can theoretically possess a specific attribute to different extents. Bolt and Kim (2018) provided empirical evidence that attributes derived from the fraction subtraction test (Tatsuoka 1987) are oversimplified if defined as binary. In addition, previous research also found that allowing attributes to have multiple levels improved the model-data fit (Haberman et al. 2008; von Davier 2018). These justifications support that assuming multiple levels of attributes are sometimes more desirable than binary levels. Therefore, developing CDMs with polytomous attributes would maximize the understanding of response patterns in binary or even polytomous data.

Many existing CDMs have been developed to measure polytomous responses, such as the Ordered Category Attribute Coding DINA model (OCAC-DINA; Karelitz 2004), the reduced reparameterized unified models (R-RUM; Templin 2004), the log-linear cognitive diagnostic model (LCDM ; Templin & Bradshaw(2013), the general diagnostic models (GDM; von Davier 2005), and the pG-DINA model (Chen & Culpepper 2020). Based on whether the interactions between attributes are considered, these models can be further specified as the main-effect cognitive diagnostic models (i.e., the OCAC-DINA, R-RUM, LCDM, and GDM) and the all-effect cognitive diagnostic models (i.e., the pG-DINA). The former involves only the main effects of the required attributes, whereas the latter involves both the main effects and interaction effects. With the between-attribute interaction effects being considered, we are able to discover all types of attribute relationships and how they can affect the observed responses. A fully saturated model is the most general parameterization of the joint attribute distribution, where all the main effects and interaction effects are taken into consideration. However, the challenge is, when the attributes or attribute levels increase, the item parameters increase exponentially. This risk of the overparameterization makes its application restricted to the confirmatory settings. The fact is for confirmatory models, accurate scoring requires the correct specification of item-attribute relationship. Otherwise, misspecification of the item structure could result in inaccurate classification. To this end, an exploratory model has been instrumental in promoting our understanding of the item-attribute structures when they are not prespecified.

With the intention of constructing an exploratory SLCM model, the key concern is to determine whether polytomous attributes are necessary and how many intermediate levels are appropriate. With data-defined polytomous attributes, the levels of attributes and their interpretations can be derived from the model-fitting process. With expert-defined polytomous attributes, the justifications of attribute levels can be provided by experts in the related areas, especially in the area of educational testing and psychopathology. In this study, we take on the first approach to explore the attribute dimensions and levels. Once an adequate model size is determined, we can move to the formal estimation of model parameters.

This chapter is organized in the following sections. The 2. *A Sparse Latent Class Model with Polytomous Attributes* section introduces the model parameterization and discusses the identifiability conditions for the model. The 3. *Gibbs Sampling* section provides a Gibbs algorithm for Markov chain Monte Carlo estimation with the potential to enforce the identifiability and monotonicity constraints. The 4. *An Empirical Application* section illustrates how well the model fits the Short Dark Triad (SD3) data compared to an exploratory item factor model and performs a Monte Carlo simulation study to assess the statistical properties and feasibility of our model. The 5. *Discussion* section provides a summary of this study and some potential directions for future research.

21.2 A Sparse Latent Class Model with Polytomous Attributes

21.2.1 Model Configurations

21.2.1.1 Unstructured Mixture Model

Consider a scale that consists of J items and K underlying attributes. We denote the vector of response probabilities for a latent class c on item j as $\theta_{cj} = (\theta_{cj0}, \dots, \theta_{cj,P-1})'$, where the element θ_{cjp} represents the probability of observing the response category p on item j by the latent class c . Note that the scale can be either dichotomous ($P = 2$) or polytomous ($P > 2$). Given the vector of response probability θ_{cj} , the probability of observing an ordinal response y_j is written as

$$P(y_j | \theta_{cj}, \eta' \mathbf{v} = c) = \sum_{p=0}^{P-1} \theta_{cjp} \mathcal{I}(y_j = p), \quad (21.1)$$

where $y_j \in \{0, \dots, P-1\}$ and \mathcal{I} is an indicator function. To describe the observed response patterns $\prod_{j=1}^J P_j$, we introduce a collection of K ordinal attributes with M levels. In this setting, a total of M^K latent classes can be created. Let the latent class index be $c \in \{0, \dots, M^K - 1\}$; each latent class has a K -vector of latent ordinal attributes $\eta \in \{0, \dots, M-1\}^K$ that can be mapped to an integer index c via bijection $\eta' \mathbf{v} = c$ with $\mathbf{v} = (M^{K-1}, M^{K-2}, \dots, 1)'$. Next, we assume the membership in class c follows a multinomial distribution with structural parameters $\pi_c \in \{\pi_0, \dots, \pi_{M^K-1}\}$ where π_c denotes the probability of membership in latent class c and $\sum_{c=0}^{M^K-1} \pi_c = 1$. By integrating out the latent class variable

c , we can write the likelihood of observing a random vector $\mathbf{y} = (y_1, \dots, y_J)$ as

$$p(\mathbf{y} \mid \Theta, \boldsymbol{\pi}) = \sum_{c=0}^{M^K-1} \pi_c \prod_{j=1}^J P(y_j \mid \theta_{cj}, \boldsymbol{\eta}'\mathbf{v} = c), \tag{21.2}$$

where $\Theta \in \mathbb{R}^{J \times M^K \times P}$ denotes the response probability array and $\boldsymbol{\pi}$ denotes the structural parameter vector. In a sample of N respondents, we denote the ordinal responses for respondent i as $y_i, i = 1, \dots, N$. The likelihood of observing this sample is

$$p(\mathbf{Y} \mid \Theta, \boldsymbol{\pi}) = \prod_{i=1}^n p(\mathbf{y}_i \mid \Theta, \boldsymbol{\pi}), \tag{21.3}$$

where \mathbf{Y} denoted the $N \times J$ response matrix.

In addition, a cumulative link function $\Psi(\cdot)$ is proposed to define the ordinal responses (Culpepper 2019). Specifically, we compute the probability of response category p by taking the difference in two adjacent cumulative probabilities. The response probability for latent class c on item j is written as

$$\theta_{cjp} = \Psi(\tau_{j,p+1} - \mu_{cj}) - \Psi(\tau_{j,p} - \mu_{cj}) \tag{21.4}$$

$$\Psi(\tau_{j,p} - \mu_{cj}) = P(y_j \leq p \mid \mu_{cj}, \tau_{j,p}), \tag{21.5}$$

where $\boldsymbol{\tau} \in \{\tau_{j,0}, \dots, \tau_{j,P}\}$. We define $\tau_{j,0} = -\infty, \tau_{j,P} = \infty$ which result in $\Psi(\tau_{j,0} - \mu_{cj}) = 0, \Psi(\tau_{j,P} - \mu_{cj}) = 1$. Here, μ_{cj} is the latent class mean parameter discussed in the next section. $\Psi(\tau_{j,p} - \mu_{cj})$ denotes the cumulative probability of a response at level p or less.

The assumption of local independence implies the response distribution \mathbf{Y} given \mathbf{B} and $\boldsymbol{\pi}$ can be presented as

$$p(\mathbf{Y} \mid \mathbf{B}, \boldsymbol{\pi}) = \prod_{j=1}^J p(\mathbf{Y}_j \mid \boldsymbol{\beta}_j, \boldsymbol{\pi}) = \prod_{i=1}^N \sum_{c=0}^{M^K-1} \pi_c \prod_{j=1}^J p(y_{ij} \mid \boldsymbol{\beta}_j, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c), \tag{21.6}$$

where y_{ij} refers to individual i 's response on item j and $\boldsymbol{\alpha}_i$ denotes the attribute profile vector for individual i .

21.2.1.2 Structured Mixture Model

A challenge with unstructured mixture model is that, as K or M increases, the number of parameters per item M^K grows exponentially. To reduce the number

of parameters, we impose a specific structure on μ_{cj} by representing μ_{cj} as $\mu_{cj} := \alpha'_c \beta_j$ where α_c is attribute profile design vector and β_j is an item parameter vector of the same length as α_c . A saturated model is the most general exploratory model which includes all the main- and higher-order interaction terms of predictors. Within the saturated model framework, K attributes with M levels lead to a total number of M^K predictors. We let $\mathbf{A} = (\alpha_0, \dots, \alpha_{M^K-1})$ be a $M^K \times M^K$ design matrix which comprises the attribute profile vector α for each latent class. Moreover, we let $\mathbf{B} = (\beta_1, \dots, \beta_J)$ be a $J \times M^K$ item coefficient matrix. In this way, since we assume a sparse pattern on \mathbf{B} explained later, the number of effective parameters is greatly reduced.

We code the attribute profile η as the design vector α using a cumulative coding strategy. Specifically, for attribute k , we define

$$\mathbf{a}_k = (1, \mathcal{I}(\eta_k \geq 1), \dots, \mathcal{I}(\eta_k \geq M - 1))', \quad (21.7)$$

so that \mathbf{a}_k incorporates an intercept for the first element and main effects for the exceeding different attribute levels. Therefore, the attribute design vector for an arbitrary class can be written as

$$\alpha = (\mathbf{a}'_1 \otimes \dots \otimes \mathbf{a}'_K)', \quad (21.8)$$

where \otimes is the Kronecker product sign which frames all possible cross-level interactions between K attributes. Below we illustrate how the coding system works with a specific example.

Table 21.1 displays the attribute profile matrix \mathbf{A} for a saturated SLCM with $K = 2$ attributes and $M = 4$ levels per attribute, where the first column prints the latent class integer c and the second column prints the latent class label in the way that each digit represents to which level latent classes master the attributes. For instance, latent class “12” in the column name implies the possession of the first attribute to the first level, and the remaining columns in the table refer to the predictors which compose the attribute profile vector α as Eq. 21.8 describes. The design vector α contains intercept component “00”; main-effect components “01,” “02,” “03,” “10,” “20,” and “30”; and two-way interaction components “11,” “12,” “13,” “21,” “22,” “23,” “31,” “32,” and “33.” For instance, label “11” in the row name corresponds to the cross-attribute effect between the first level of attribute 1 and the first level of attribute 2. For latent class “12” ($\eta_1 = 1, \eta_2 = 2$), component “11” ($a_1 = 1, a_2 = 1$) is active given the coding rule $\eta_1 \geq 1$ and $\eta_2 \geq 1$. Specifically, for latent class “12” (i.e., $\eta_1 = 1, \eta_2 = 2$), the active components refer to “00,” “01,” “02,” “10,” “11,” and “12” columns.

21.2.1.3 Model Identifiability

Model identifiability issues have received considerable attention in the study of CDMs. Traditionally, parameter constraints derived for model identifiability

Table 21.1 Example design matrix A of latent classes by attribute predictors for $K = 2, M = 4$

α																	
c	η	00	01	02	03	10	11	12	13	20	21	22	23	30	31	32	33
0	00	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	01	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	02	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	03	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
4	10	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	11	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0
6	12	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0
7	13	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
8	20	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
9	21	1	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0
10	22	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	0
11	23	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
12	30	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
13	31	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
14	32	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0
15	33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

are often imposed on a design Q-matrix. The Q-matrix specifies item-attribute relationships by showing which attributes each item measures but not showing how attributes interact to affect response probabilities. Conversely, in SLCM, a sparsity matrix $\Delta_{J \times M^K}$ substitutes the role of the traditional Q-matrix but takes the inter-attribute relationships into consideration. The sparsity matrix $\Delta_{J \times M^K}$ depicts the underlying pattern of the item parameter matrix \mathbf{B} , where J denotes the total amount of items and M^K denotes the total amount of predictors. An element $\delta = 1$ suggests its corresponding β is active, whereas an element $\delta = 0$ suggests its corresponding β is 0.

Definition 2.1 presents a classic notion of likelihood identifiability where a different set of parameter values leads to different values of likelihood. To this end, a model must be identifiable to elicit consistent parameter estimates. Based on the work established by Culpepper (2019) for the SLCM with binary attributes and ordinal responses, and the work by Chen et al. (2020) with binary attributes and dichotomous responses, we propose identifiability conditions for the SLCM with ordinal responses and ordinal attributes.

Definition 2.1 A parameter set $\Omega(\boldsymbol{\pi}, \mathbf{B})$ is identifiable, if there are two sets of parameters $(\boldsymbol{\pi}, \mathbf{B})$ and $(\bar{\boldsymbol{\pi}}, \bar{\mathbf{B}})$ such that $\mathbb{P}(\boldsymbol{\pi}, \mathbf{B}) = \mathbb{P}(\bar{\boldsymbol{\pi}}, \bar{\mathbf{B}})$ implies $\boldsymbol{\pi} = \bar{\boldsymbol{\pi}}, \mathbf{B} = \bar{\mathbf{B}}$.

We define a parameter space $\Omega(\boldsymbol{\pi}, \mathbf{B})$ for the latent class proportion parameter $\boldsymbol{\pi}$ and the item coefficient parameter \mathbf{B} as

$$\Omega(\boldsymbol{\pi}, \mathbf{B}) = \{(\boldsymbol{\pi}, \mathbf{B}) : \boldsymbol{\pi} \in \Omega_1(\boldsymbol{\pi}), \mathbf{B} \in \Omega_2(\mathbf{B})\}, \quad (21.9)$$

where $\Omega_1(\boldsymbol{\pi}) = \{\boldsymbol{\pi} \in \mathbb{R}^{M^K} : \sum_{c=0}^{M^K-1} \pi_c = 1, \pi_c > 0\}$ and $\Omega_2(\mathbf{B}) = \{\mathbf{B} \in \mathbb{R}^{J \times M^K}\}$.

Conditioned on the attribute profile vector $\boldsymbol{\alpha}$, the joint distribution of \mathbf{Y} is a product of multinomial distributions represented by a P^J vector:

$$\mathbb{P}_{\boldsymbol{\alpha}}(\mathbf{B}) = \bigotimes_{j=1}^J \boldsymbol{\theta}_{cj}, \quad (21.10)$$

where \bigotimes refers to the Kronecker product and $\boldsymbol{\theta}_{cj} = (\theta_{cj0}, \dots, \theta_{cj, P-1})'$. Further, the marginal distribution of \mathbf{Y} over the proportion parameter $\boldsymbol{\pi}$ is

$$\mathbb{P}(\boldsymbol{\pi}, \mathbf{B}) = \sum_{\boldsymbol{\alpha}} \pi_{\boldsymbol{\alpha}} \mathbb{P}_{\boldsymbol{\alpha}}(\mathbf{B}). \quad (21.11)$$

21.2.1.4 Monotonicity Constraints

Imposing monotonicity constraints ensures that mastering any irrelevant skills to an item will not increase the endorsing probability. Xu (2017) and Xu and Shang (2018) proposed the monotonicity constraints as follows:

$$\min_{c \in \mathbb{S}_0} \mu_{cj} \geq \mu_{c_0j}, \quad (21.12)$$

$$\max_{c \in \mathbb{S}_0} \mu_{cj} < \min_{c \in \mathbb{S}_1} \mu_{cj} = \max_{c \in \mathbb{S}_1} \mu_{cj}, \quad (21.13)$$

where c_0 represents the latent class that does not own any relevant attribute and μ_{c_0j} denotes its latent class mean parameters. \mathbb{S}_0 denotes a set of latent classes that own at least one but not all relevant attributes, and \mathbb{S}_1 denotes a set of latent classes that own all the relevant attributes. Given $\mu_{cj} = \boldsymbol{\beta}'_j \boldsymbol{\alpha}_c$, we can derive a lower bound condition L_{cj} for each item parameters β_{cj} that if β_{cj} is lower bounded, the constraints above are satisfied. The derivation details can be found in Chen et al. (2020).

21.2.1.5 Strict Identifiability

Theorem 2.2 *The parameter space $\Omega(\boldsymbol{\pi}, \mathbf{B})$ is strictly identifiable if conditions (S1) and (S2) are satisfied.*

(S1) When Δ matrix takes the form of $\Delta = \begin{pmatrix} D_1 \\ D_2 \\ D^* \end{pmatrix}$ after row swapping where D_1 ,

$$D_2 \in \mathbb{D}_s, \mathcal{D}_s = \left\{ D \in \{0, 1\}^{K \times M^K} : D = \begin{bmatrix} 0 & \mathbf{1}'_{M-1} & 0 & \dots & 0 & \dots & 0 \\ 0 & 0 & \mathbf{1}'_{M-1} & \dots & 0 & \dots & 0 \\ \vdots & 0 & 0 & \ddots & 0 & \dots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{1}'_{M-1} & \dots & 0 \end{bmatrix} \right\}.$$

Note: $\mathbf{1}_{M-1}$ is a vector of 1 of length $M - 1$ which represents the activeness of an attribute on all $M - 1$ levels.

(S2) In D^* , for any attribute $k = 1, 2, \dots, K$, there exists an item $j > 2K$, such that all the main-effect components regarding this attribute are active ($\delta_{j,k0}, \delta_{j,k1}, \dots, \delta_{j,k(M-1)} = 1$).

The proof details are provided in Appendix ‘‘Strict Identifiability Proof’’.

21.2.1.6 Generic Identifiability

In this section, we propose the generic identifiability condition in (G1) and (G2) in Theorem 2.4. The generic condition is less stringent than the strict conditions (S1) and (S2) given in 2.2. Generic identifiability allows part of the model parameters to be non-identifiable such that these exceptional values are of measure zero in the parameter space.

Definition 2.3 A parameter set $\Omega_\Delta(\boldsymbol{\pi}, \mathbf{B})$ is generically identifiable if the Lebesgue measure of the unidentifiable space C_Δ with respect to $\Omega_\Delta(\boldsymbol{\pi}, \mathbf{B})$ is zero.

Theorem 2.4 *The parameter space $\Omega_\Delta(\boldsymbol{\pi}, \mathbf{B})$ is generically identifiable if condition (G1) and (G2) are satisfied.*

(G1) When Δ matrix takes the form of $\Delta = \begin{pmatrix} D_1 \\ D_2 \\ D^* \end{pmatrix}$ after row swapping where D_1 ,

$$D_2 \in \mathbb{D}_g, \mathbb{D}_g = \left\{ D \in \{0, 1\}^{K \times M^K} : D = \begin{bmatrix} * & \mathbf{1}'_{M-1} & * & \dots & * & \dots & * \\ * & * & \mathbf{1}'_{M-1} & \dots & * & \dots & * \\ \vdots & * & * & \ddots & * & \dots & \vdots \\ * & * & * & \dots & \mathbf{1}'_{M-1} & \dots & * \end{bmatrix} \right\}$$

Note: $\mathbf{1}_{M-1}$ is a vector of 1 of length $M - 1$ which represents the activeness of a specific attribute on all levels.

- (G2) In D^* , for any attribute $k = 1, 2, \dots, K$, there exists an item $j > 2K$, such that all the main-effect components regarding this attribute are active ($\delta_{j,k0}, \delta_{j,k1}, \dots, \delta_{j,k,M-1} = 1$).

The proof details are provided in Appendix ‘‘Generic Identifiability Proof’’.

21.3 Gibbs Sampling

Following the Bayesian model formulation displayed in Sect. 21.2.1, this section outlines an MCMC approach for the proposed SLCM models. First, we introduce a deterministic relationship between the observed ordinal response Y_{ij} and a continuous augmented latent variable Y_{ij}^* as Eqs. 21.14 and 21.15 show. The augmented variable Y_{ij}^* is generated from a normal distribution conditioned on the latent class mean parameter $\mu_{ij} = \alpha_i' \beta_j$. If Y_{ij}^* falls into the range $[\tau_{jp}, \tau_{j,p+1})$, the random variable Y_{ij} takes the value of p .

$$Y_{ij} = \sum_{p=0}^P p \mathcal{I}(\tau_{jp} \leq Y_{ij}^* < \tau_{j,p+1}) \quad (21.14)$$

$$Y_{ij}^* | \alpha_i, \beta_j \sim N(\alpha_i' \beta_j, 1) \quad (21.15)$$

We consider a multinomial prior for latent attribute variable α_i as $\alpha_i \sim \text{Multinomial}(\boldsymbol{\pi})$. The latent class structural probability $\boldsymbol{\pi}$ follows a conjugate Dirichlet distribution $\boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{d}_0)$ with hyperparameter $\mathbf{d}_0 = \mathbf{1}'_{MK}$.

In addition, we adopt a spike and slab prior for \mathbf{B} as Culpepper (2019) described. For each single β parameter, we formulate the Bayesian model as follows:

$$\beta_{jc} | \delta_{jc} \sim \begin{cases} \mathcal{N}(0, \sigma_\beta^2) \mathcal{I}(\beta_{jc} > L_{jc}) & \delta_{jc} = 1 \\ \mathcal{I}(\beta_{jc} = 0) & \delta_{jc} = 0 \end{cases}$$

$$\delta_{jc} | \omega \sim \text{Bernoulli}(\omega)$$

$$\omega \sim \text{Beta}(w_0, w_1),$$

where $(\sigma_\beta^2, w_0, w_1)$ are user-specified hyperparameters and L_{jc} refers to the lower bound for satisfying the monotonicity constraints mentioned in Sect. 21.2.1.4. Noted that the intercepts in $\boldsymbol{\Delta}$ are always set to be active with $\delta_{j0} = 1$.

As a binary variable, we sample δ_{jc} from the following conditional distribution:

$$\delta_{jc} \mid \mathbf{y}_j^*, \mathbf{A}, \boldsymbol{\beta}_{j(c)}, \omega, \sigma_\beta^2 \sim \text{Bernoulli}(\tilde{\omega}_{jc}), \tag{21.16}$$

where $A = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_{M^k})$ refers to the attribute profile matrix, \mathbf{y}_j^* is the augmented responses vector, and $\boldsymbol{\beta}_{j(c)}$ is the coefficient vector $\boldsymbol{\beta}_j$ that discards the p -th element. Once δ_{jc} is updated, we update $\beta_{j(c)}$ given the full conditional distribution:

$$\beta_{jc} \mid \mathbf{y}_j^*, \mathbf{A}, \boldsymbol{\beta}_{j(c)}, \omega, \sigma_\beta^2, \delta_{jc} \sim \mathcal{N} \left(\tilde{\mu}_{jc}, \tilde{\sigma}_c^2 \right) [\mathcal{I}(\beta_{jc} > L_{jc})]^{\delta_{jc}} [\mathcal{I}(\beta_{jc} = 0)]^{1-\delta_{jc}}. \tag{21.17}$$

Given Eqs. 21.16 and 21.17, the Bernoulli parameter $\tilde{\omega}_{jc}$ is derived as

$$w_{jc} = \frac{w \Phi \left(\frac{-L_{jc}}{\sigma_\beta} \right)^{-1} \left(\frac{\tilde{\sigma}_c}{\sigma_\beta} \right) \Phi \left(\frac{\tilde{\mu}_{jc} - L_{jc}}{\tilde{\sigma}_c} \right) \exp \left(\frac{\tilde{\mu}_{jc}^2}{2\tilde{\sigma}_c^2} \right)}{w \Phi \left(\frac{-L_{jc}}{\sigma_\beta} \right)^{-1} \left(\frac{\tilde{\sigma}_c}{\sigma_\beta} \right) \Phi \left(\frac{\tilde{\mu}_{jc} - L_{jc}}{\tilde{\sigma}_c} \right) \exp \left(\frac{\tilde{\mu}_{jc}^2}{2\tilde{\sigma}_c^2} \right) + 1 - w} \tag{21.18}$$

$$\tilde{\mu}_{jc} = \tilde{\sigma}_c^2 A'_c (y_{jc}^* - A_{(c)} \boldsymbol{\beta}_{j(c)}) \tag{21.19}$$

$$\tilde{\sigma}_c^2 = (A'_c A_c + \sigma_\beta^{-2})^{-1} \tag{21.20}$$

where A_c refers to the c -th column in the design matrix A . Note the derivation details can be found in Chen et al. (2020). The full MCMC sampling process is summarized in Algorithm 1, whereas Algorithm 2 presents the detailed sampling steps of the parameter matrix \mathbf{B} and $\boldsymbol{\Delta}$.

21.4 An Empirical Application

21.4.1 Short Dark Triad

In the past decade, a great interest has been directed to measure the dark pattern of behaviors, goals, and characters. The Dark Triad (DT; Paulhus & Williams 2002) is one of the most popularly studied personality constructs, which encompasses three substantive dimensions: Machiavellianism, narcissism, and psychopathy. However, different studies have made contrasting conclusions on the construct of the DT (Persson et al. 2019). For instance, Furnham et al. (2013) have argued that psychopathy sometimes subsumes Machiavellianism and narcissism inadvertently. Others have declared that Machiavellianism and psychopathy have the same core and should be deemed as the same measure (Garcia & Rosenberg 2016;

Glenn & Sellbom 2015; McHoskey et al. 1998). Traditionally, assessment of the DT often requires distinct measures for each dimension. To simplify the process of data collection, a measure, namely, Short Dark Triad (SD3; Jones & Paulhus 2014), was created with 27 items selected in a 5-point Likert-type scale (1, “Strongly disagree”; 2, “Slightly disagree”; 3, “Neutral”; 4, “Slightly agree”; and 5, “Strongly agree”). The SD3 items are shown in Appendix “Short Dark Triad”.

In this study, we investigate the latent construct underlying the SD3 scale through exploratory SLCMs. The dataset are available on the Open Psychometrics website (<https://openpsychometrics.org/tests/SD3/>), where we select a random sample of $N = 5000$ observations. Original item affiliations to the three dimensions can be inferred from the item index in Appendix “Short Dark Triad” (“M” refers to “Machiavellianism”; “N” refers to “narcissism”; “P” refers to “psychopathy”).

21.4.2 Model Comparisons

Traditionally, exploratory factor analysis (EFA; Furnham et al. 2014) has been the most popular tool to excavate latent constructs underlying manifest variables in a self-reported questionnaire. Unlike EFA models which treat personality traits as continuous variables, the SLCM allows us to explore the potential for interpreting the personality trait as discrete variables. The purpose of this section is to (1) fit exploratory SLCM with different attributes ($K = 2, 3, 4$) and attribute levels ($M = 2, 3, 4$) and (2) compare the exploratory SLCM to EFA models with ($K = 2, 3, 4$) factors.

Using a Bayesian approach, we apply a tenfold cross-validation approach to estimate out-of-sample predictive accuracy using within-sample estimates. We choose the k-fold cross-validation approach due to its simplicity compared to the leave-one-out method. Considering that an increasing number of folds help reduce the bias term (Vehtari & Lampinen 2002) caused by data split, we partition the DT3 dataset ($N = 5000$) into *ten* subsets $\{y_k \mid k = 1, \dots, 10\}$. One fold is used as testing data ($N = 500$), and the remaining folds are used as training data ($N = 4500$). For fold k , the testing and training data are denoted as y_k and $y_{(-k)}$, respectively. For each training data $y_{(-k)}$, we employ the algorithm discussed in Sect. 21.3 to obtain posterior draws of the exploratory SLCM and computed posterior means as the point estimates for \mathbf{B} and $\mathbf{\Delta}$. Thresholds τ are fixed to be $\tau = \{-\infty, 0, 2, 4, \infty\}$. We run 10 Markov chain Monte Carlo (MCMC) chains, and for each MCMC chain, a total of 80,000 iterations are generated. Specifically, within each chain, the first 20,000 iterations are discarded as burn-in samples, and the left 60,000 iterations are retained as posterior samples. Finally, one chain is chosen out of the *ten* MCMC chains as it generated the highest marginal likelihood. Note here the

point estimates are computed by taking element-wise means for the item parameter matrix \mathbf{B} and the structural parameter vector $\boldsymbol{\pi}$ over posterior distributions. With the posterior distributions estimated from the training data, we then evaluate model fit in the testing data using log point-wise predictive density (lpd) which is defined as

$$\begin{aligned} \text{lpd} &= \sum_{k=1}^{10} \sum_{i \in y_k} \log p(\mathbf{y}_i | \mathbf{y}_{(-k)}) \\ &= \sum_{k=1}^{10} \sum_{i \in y_k} \log \int p(\mathbf{y}_i | \boldsymbol{\Omega}_k) p(\boldsymbol{\Omega}_k | \mathbf{y}_{(-k)}) d\boldsymbol{\Omega}_k, \quad i \in y_k, \end{aligned} \quad (21.21)$$

where $\boldsymbol{\Omega}_k = \{\mathbf{B}_k, \boldsymbol{\pi}_k\}$ represents the item parameters estimated from the training data $\mathbf{y}_{(-k)}$. To compute lpd in practice, we evaluate the integration of $\boldsymbol{\Omega}_k$ using MCMC posterior draws, and the log point-wise predictive density for data points in the testing data \mathbf{y}_k is written as

$$\widehat{\text{lpd}}_k = \sum_{i \in y_k} \log \frac{1}{S} \sum_{s=1}^S f_m(\mathbf{y}_i | \mathbf{B}^{k,s}, \boldsymbol{\pi}), \quad (21.22)$$

where $\mathbf{B}^{k,s}$ are the s_{th} draws from the posterior distributions given the training data $\mathbf{y}_{(-k)}$ and i is the index of individuals. The complete log predictive density can be calculated by summing all observations over the 10-folds as $\widehat{\text{lpd}} = \sum_{k=1}^{10} \widehat{\text{lpd}}_k$. Furthermore, the marginal likelihood of response \mathbf{y}_i for the s_{th} draws is written as

$$f_m(\mathbf{y}_i | \mathbf{B}^{k,s}, \boldsymbol{\pi}) = \sum_{\boldsymbol{\alpha}_i^{k,s}} f_c(\mathbf{y}_i | \mathbf{B}^{k,s}, \boldsymbol{\alpha}_i^{k,s}) g(\boldsymbol{\alpha}_i^{k,s} | \boldsymbol{\pi}), \quad (21.23)$$

where the latent variable $\boldsymbol{\alpha}_i^{k,s}$ is integrated out with the hyperparameter $\boldsymbol{\pi}$. Since data y_{ij} are independent response data conditioned on the model parameter $\boldsymbol{\beta}_j$ and the attribute pattern $\boldsymbol{\alpha}_i$, we have

$$f_c(\mathbf{y}_i | \mathbf{B}^{k,s}, \boldsymbol{\alpha}_i^{k,s}) = \prod_{j=1}^J \sum_{p=0}^{P-1} \mathbb{1}(y_{ij} = p) P(y_{ij} = p | \boldsymbol{\beta}_j^{k,s}, \boldsymbol{\alpha}_i^{k,s}) \quad (21.24)$$

For EFA models, we apply the function “MCMCordfactanal” in the R package **MCMCpack** (Martin et al. 2011) to perform Bayesian estimation for posterior inference. We use its default setting of 10,000 burn-in and 20,000 mcmc iterations, and for each training sample $\mathbf{y}_{(-k)}$, we take 500 posterior draws to compute predictive accuracy in testing sample \mathbf{y}_k . Similar to SLCM,

we calculate the lpd criterion for the EFA models wherein the integration in Eq.(21.22) is obtained via the function “hcubature” in the R package **cubature** (Narasimhan et al. 2020). A higher lpd value indicates a model with higher prediction accuracy. Note that, considering the variety of latent variables assumed by SLCM and EFA, we use the marginal likelihoods f_m instead of the conditional likelihoods f_c in the predictive likelihood function $f_m(y_i | \mathbf{B}^{k,s}, \boldsymbol{\pi})$.

As seen in Table 21.2, inclusion of more factors leads to a superior fit of EFA models for the reason that more nuisance variance is considered. In the two-dimensional case of $K = 2$, the SLCM ($lpd \geq -183595.3$) performs better than the EFA ($lpd = -186914.5$) if the attribute is ordinal $M \geq 3$. Furthermore, in the three-/four-dimensional case of $K \geq 3$, the SLCM consistently outperform the EFA regardless of the choice of M . Overall, the lpd of SLCM shows a slight rise when the dimension K or attribute level M increases, suggesting that increasing either the dimension or attribute level could improve the model fit. At this point, the best fitting model is the SLCM with $K = 2$ attributes and $M = 4$ attribute levels ($lpd = -183,443.2$).

Given the information concerning the relative model fit, we estimated an exploratory SLCM with $K = 2$ and $M = 4$ in the same SD3 dataset ($N = 5000$). We ran 10 chains of length 80,000 with a burn-in of 20,000 iterations and keep the chain with the highest marginal likelihood. Figure 21.1 displays the estimated sparse structure of \mathbf{B} , where we can summarize the two attributes as (1) narcissism and (2) Machiavellianism. The x-axis of Fig. 21.1 presents the predictors, where 01, 02, and 03 refer to the main-effect predictors of Machiavellianism; 10, 20, and 30 refer to the main-effect predictors of narcissism; and sparsity of the matrix is reflected on the fact that loadings manifest on the main-effect predictors. To this end, we can obtain a rough conclusion on the item and attribute relationships. In particular, items M1–M9, P1, P3, P5, P6, and P9 load mostly on Machiavellianism; items N1–N4, N6–N8, P2, and P7 load on narcissism; and items N5, N9, P4, and P8 load equally on the two dimensions.

Table 21.2 Model comparison results in lpd, SD3 ($N = 5000$)

Dimension	EFA		SLCM
$K = 2$	-186,914.5	$M = 2$	-189,320.7
		$M = 3$	-183,595.3
		$M = 4$	-183,443.2
$K = 3$	-185,898.2	$M = 2$	-183,929.6
		$M = 3$	-183,916.3
		$M = 4$	-183,902.9
$K = 4$	-184,743.8	$M = 2$	-183,577.2
		$M = 3$	-183,552.5
		$M = 4$	-183,532.1

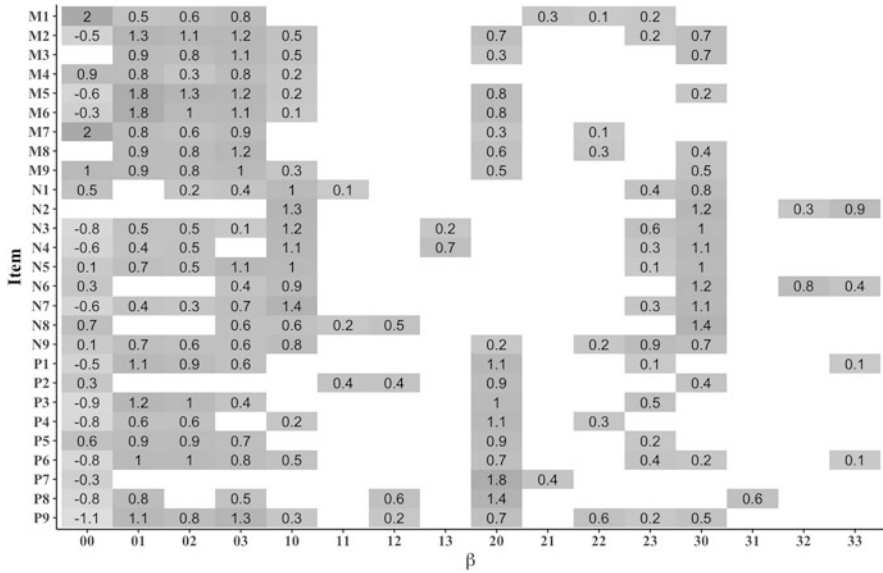


Fig. 21.1 Estimated B matrix for DT3 data under SLCM with $K = 2, M = 4$

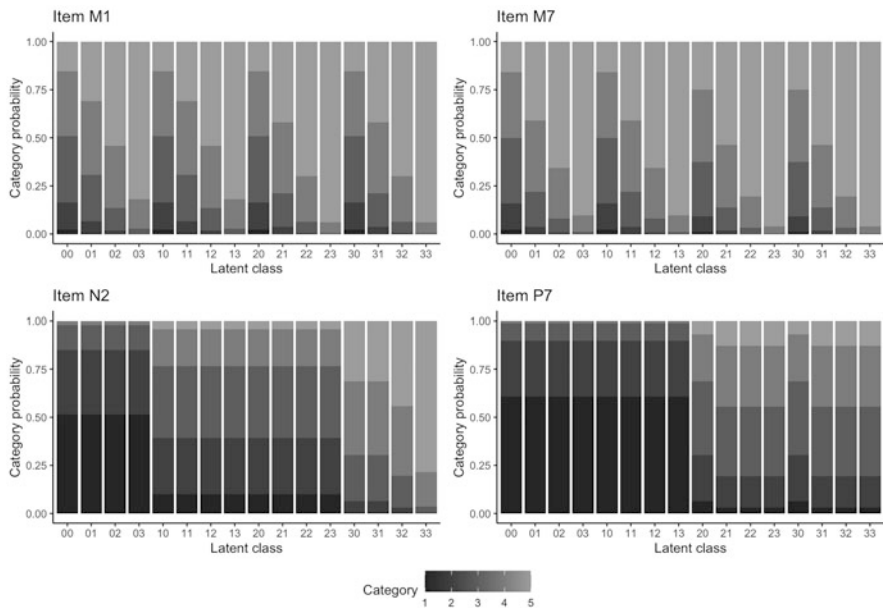


Fig. 21.2 Estimated item category response function by latent class and category probability

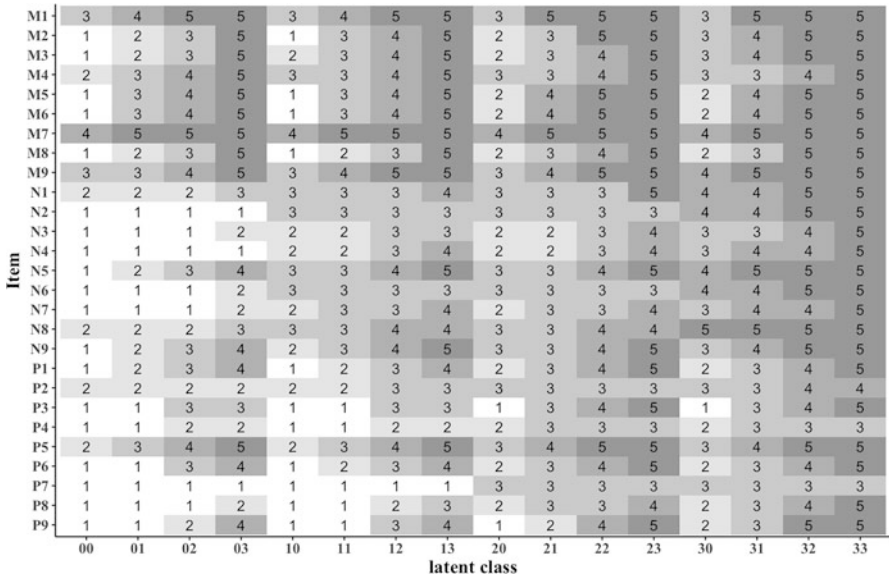


Fig. 21.3 Dominant responses by latent classes and items

As a fully saturated model, interactions can take place between any level of any two attributes. Figure 21.2 are stack barplots that clearly depict how increases in attributes correspond to changes in category response probabilities. The x- and y-axes indicate the latent classes and response probabilities; the stacked bars represent the five response categories. The barplots selectively present the item category response function for items M1, M7, N2, and P7. Item M1 has active coefficients $\beta_{00} = 2.0$, $\beta_{01} = 0.5$, $\beta_{02} = 0.6$, $\beta_{03} = 0.8$, $\beta_{21} = 0.3$, $\beta_{22} = 0.1$, and $\beta_{23} = 0.2$. We can see that all main-effect terms regarding Machiavellianism manifest, while narcissism is active only on the interaction terms. In Fig. 21.3, latent classes 10, 20, and 30 that represents the group are less likely to endorse category 5 (5 = “Strongly agree) compared to other latent classes.

The active coefficients of item N2 are $\beta_{10} = 1.3$, $\beta_{30} = 1.2$, $\beta_{32} = 0.3$, and $\beta_{33} = 0.9$. We can tell that narcissism is more significant than Machiavellianism. In Fig. 21.3, latent classes 01, 02, and 03 which reflects the mastery of three levels on narcissism are most likely to endorse category 1 (5 = “Strongly disagree) compared to other latent classes.

Moreover, Fig. 21.3 shows the dominant response category for each latent class on the 27 items. Given a specific item, the value on the table represents the response category which a latent class has the highest probabilities to endorse over the other response categories.

21.4.3 Monte Carlo Simulation Study

This section presents a simulation study to evaluate the parameter recovery rate for SLCM with different sample sizes. We use the previous estimates of item parameter \mathbf{B} and the structural parameter $\boldsymbol{\pi}$ (see Appendix “Empirical Item Parameters”) as the population parameters. Specifically, we have $J = 27$ items and assume the underlying dimension is $K = 2$ and attribute level is $M = 4$. The sample size is set to be $n = 1000, 2000, 3000$, with each sample size condition replicated for 100 times. For each replicated dataset, we run 10 chains where each single chain has a total of 60,000 iterations with a burn-in sample of 20,000 inside. The chain with the highest likelihood is chosen to perform posterior inference and compute the recovery accuracy. Note here we generate different attribute profiles $\boldsymbol{\alpha}$ and responses \mathbf{Y} per replication.

The estimation accuracy of \mathbf{B} is evaluated in terms of the mean absolute deviation (MAD) for each single β . For each replication, we record the posterior mean of each single parameter in \mathbf{B} as the estimates. Next, we compute the absolute deviation between the estimates and the true parameters. Then, we take the average of the absolute deviation over replications. Table 21.3 enunciates the MAD of β s with its activeness into consideration. In specific, the first row “ \mathbf{B}_Δ ” refers to the MAD averaged over the entire matrix \mathbf{B} . The second row “ $\mathbf{B}_{\Delta=1}$ ” and the third row “ $\mathbf{B}_{\Delta=0}$ ” refer to the MAD averaged over the locations where $\delta_s = 1$ and $\delta_s = 0$, respectively. Likewise, we also compute and record the mean absolute deviation (MAD) for each π . Table 21.4 presents the recovery rate of Δ in terms of the proportion of entries that are correctly recovered. The first row “ Δ ” refers to the proportion of correctly recovered δ s over the whole matrix. The second row “ $\Delta = 1$ ” and the third row “ $\Delta = 0$ ” refer to the proportion of 1’s and 0’s in the population Δ matrix that are correctly recovered.

The result in Table 21.3 shows that the average EAD for \mathbf{B} is 0.100, 0.069 and 0.052 for sample sizes of $n = 1000, 2000$, and 3000. The average EAD for $\boldsymbol{\pi}$ is 0.009, 0.008, and 0.007 corresponding to sample sizes of $n = 1000, 2000$, and 3000. Additionally, Table 21.4 shows the recovery rate for \mathbf{D} is 0.883, 0.913, and 0.976 for sample sizes of $n = 1000, 2000$, and 3000.

Table 21.3 Mean absolute deviation (MAD) of \mathbf{B} and $\boldsymbol{\pi}$

	$n = 1000$	$n = 2000$	$n = 3000$
\mathbf{B}_Δ	0.100	0.069	0.052
$\mathbf{B}_{\Delta=1}$	0.237	0.166	0.128
$\mathbf{B}_{\Delta=0}$	0.046	0.031	0.022
$\boldsymbol{\pi}$	0.009	0.008	0.007

Table 21.4 Recovery accuracy of Δ

	$n = 1000$	$n = 2000$	$n = 3000$
Δ	0.883	0.913	0.931
$\Delta = 1$	0.702	0.773	0.814
$\Delta = 0$	0.954	0.968	0.976

and 0.932 for sample sizes of $n = 1000, 2000$, and 3000 . As seen, for both of them, the estimation accuracy rises as the sample size increases. Furthermore, in Table 21.3, we observe that the active entries in \mathbf{B} have a larger bias than the inactive entries. Similarly, we have at least 0.954 of the inactive entries in \mathbf{D} that are correctly estimated as “0” and at least 0.702 of the active entries in \mathbf{D} that are correctly estimated as “1.” The simulation results support that the model can be mostly recovered by our Gibbs algorithm.

21.4.4 Model Convergence

To evaluate the convergence of the Gibbs sampler, we generate *three* chains for the SLCM with $K = 2$ and $M = 4$ under the most computationally intensive condition $N = 3000$. For each of the latent class mean parameter μ and the structural parameter π , the Gelman-Rubin proportional scale reduction factor (PSRF), also known as \hat{R} , is calculated. A \hat{R} value of below 1.2 indicates the acceptable convergence. In our simulation, the maximum \hat{R} is found to be 0.97 for π and 1.06 for μ , with the 80,000 iterations and 20,000 burn-in samples inside. Therefore, we conclude the MCMC chains have reached a steady state.

21.5 Discussion

In this study, we propose a strict and generic model identifiability condition for SLCM with polytomous attributes, which expands the work of SLCM with binary attributes by Chen et al. (2020) and Culpepper (2019). We develop a Gibbs sampling algorithm with the design of enforcing the identifiability and monotonicity constraints. Specifically, with strict identifiability conditions imposed, we notice that the MCMC chains are often trapped and have a slow move forward. A possible explanation is that the strict conditions are too restrictive for the chains to search the right parameter space. Without explicitly enforcing the strict identifiability constraints, the models convergence in 80,000 draws with estimates satisfying the proved generic identifiability conditions. The simulation results demonstrate that the algorithm is efficient in recovering the parameters in different sample sizes.

Overall, our study is innovative in the following aspects. First, we provide a successful case study of applying SLCM to a personality scale. Personality have traditionally been viewed as continuous traits instead of discrete categories, and factor analysis (FA) approach which assumes continuous latent variables is often used in personality measurement. However, with the estimated person scores, it

is always a critical issue to identify cut-offs and classify individuals into different classes. To this end, if variables can be viewed more or less categorical, we can rely on the models with discrete latent variables to provide fine-grained information on the individual differences. Another advantage with discrete latent variables lies in its greatly reduced parameter space, with a potential of facilitating sampling. Overall, our study provides new insights into interpreting personality traits as a discrete measure. Plus, the model also has the potential of being applied to educational settings and contributes to better measurement for educational intervention (Chen & Culpepper 2020).

Second, our study, for the first time, compares SLCM and EFA models from an exploratory perspective. We found the SLCM fit significantly better than the EFA models with a higher prediction accuracy in several configurations. This finding has important implications for promoting applications of CDMs to the areas outside of educational measurement, where another early example is by Cho (2016) who has explored the construct validity of emotional intelligence in situational judgmental tests. In addition, our analysis of the item-attribute structures underlying SD3 supports the previous finding by Persson et al. (2019) that Machiavellianism and psychopathy are subsumable constructs. Moreover, they found the subscale composite scores for the three constructs contain relatively little specific variance, with an implication that reporting the total scores is more appropriate for SD3 than reporting the subscale scores. In our result, most items do not follow a simple structure pattern, which further support this statement that dimensions of SD3 are somewhat inseparable.

Third, from a methodology perspective, our paper addresses the model identifiability concerns of SLCM with polytomous attributes. The strict identifiability condition is way too restrictive in practice. For instance, when the number of attributes is relatively large compared to the items, (e.g., close to half the number of items), enforcing strict identifiability is equivalent to presuming a simple structure on all items. For personality assessment, a simple item structure is often unrealistic to achieve. For this reason, the generic identifiability that loosens some constraints broadens the model applicability.

There are still several recommendations for future study. First, although the MCMC chains successfully converge to the posterior distributions, the Gibbs samplers are still not efficient enough in exploring the parameter space. We have to run several chains and conduct a likelihood selection to find the one with best mixing. The difficulty of mixing could be due to the complexity of the saturated model wherein we have 16 parameters per item. To solve the mixing issue, future work is required to develop more flexible moves in the algorithm that can break local traps or jump between difference spaces.

Second, instead of framing the SLCM in an unstructured way, it is interesting to include a higher-order factor model (Culpepper & Chen 2019; De La Torre & Douglas 2004; Henson et al. 2009) or a multivariate normal distribution with

a vector of thresholds and a polychoric correlation matrix (Chen & Culpepper 2020; Henson et al. 2009; Templin et al. 2008) in the latent class structure. There is also abundant room for future progress in selecting the competing structures for π .

Third, it is also possible to prespecify the number of attributes with a more established approach. For instance, Chen et al. (2021) present a crimp sampling algorithm to jointly infer the number of attribute for DINA model. In our study, the choice of attribute level is limited by the study design. Future work with focus on the selection of attribute level is greatly suggested.

Algorithm 1 Full Gibbs sampling algorithm

Data: $Y_{N \times J}$; π ; $\alpha_{1:N}$; $B_{J \times M^K}$; τ ; $A_{M^K \times M^K}$; chain length T

Result: Y^* ; π ; α

```

for  $t$  in  $(1, \dots, T)$  do
  for  $j$  in  $(1, \dots, J)$  do
    for  $c \in (0, \dots, M^K - 1)$  do
      for  $y_{ij} \in \{0, \dots, P - 1\}$  do
         $\theta_{jc, y_{ij}} = \Phi(\tau_{y_{ij+1}} - \alpha'_c \beta_j) - \Phi(\tau_{y_{ij}} - \alpha'_c \beta_j)$ ;
      end
    end
  end
  for  $i$  in  $(1, \dots, N)$  do
    Sample  $\alpha_i$  from multinomial distribution;
     $P(\alpha_i = \alpha_c \mid \pi, y_i) \propto \frac{\pi_c \prod_{j=1}^J \theta_{c, y_{ij}}}{\sum_{c=0}^{M^K-1} \pi_c \prod_{j=1}^J \theta_{c, y_{ij}}}$ ;
  end
  for  $c \in (0, \dots, M^K - 1)$  do
    Sample  $\pi$  from Dirichlet distribution;
     $P(\pi \mid A) \propto \prod_{c=0}^{M^K-1} \pi_c^{\sum_{i=1}^N I(\alpha_i = \alpha_c) + d_{oc}}$ 
  end
  for  $j$  in  $(1, \dots, J)$  do
    for  $i$  in  $(1, \dots, N)$  do
      Sample  $y_{ij}^*$  from truncated normal distribution;
       $P(y_{ij}^* \mid \alpha_i, \beta_j) \propto N(\alpha'_i \beta_j, 1) I(\tau_{j, y_{ij}} < y_{ij}^* < \tau_{j, y_{ij+1}})$ 
    end
  end
  end
  Sample  $B$  and  $\Delta$  from Algorithm 2
end

```

Algorithm 2 Full Gibbs sampling algorithm: $\mathbf{B}; \Delta$

Data: Hyperparameters $\sigma_\beta, w, w_0, w_1; Y^0$ from Algorithm 1

Result: \mathbf{B}, Δ

```

for  $t$  in  $(1, \dots, T)$  do
  for  $c$  in  $(1, \dots, M^K - 1)$  do
     $\tilde{\sigma}_c^2 = \frac{1}{A'_c A_c + \sigma_\beta^{-2}}$ ;
    for  $j$  in  $1, \dots, J$  do
       $\tilde{\mu}_{jc} = \tilde{\sigma}_c^2 A'_c (Y_j^0 - A_{(c)} \beta_{j(c)})$ ;
       $L \leftarrow \max \left\{ \max_{\alpha \in \mathbb{L}_1} -\gamma'_\alpha, \max_{\alpha \in \mathbb{L}_0} \gamma'_\alpha - \gamma'_{q_j} \right\}$ ;
      if  $(L \leq 0)$  and the identifiability condition is satisfied then
        
$$w_{jc} = \frac{w \Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \left(\frac{\tilde{\sigma}_c^2}{\sigma_\beta^2}\right)^{\frac{1}{2}} \Phi\left(\frac{\tilde{\mu}_{jc} - L}{\tilde{\sigma}_c}\right) \exp\left(\frac{\tilde{\mu}_{jc}^2}{2\tilde{\sigma}_c^2}\right)}{w \Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \left(\frac{\tilde{\sigma}_c^2}{\sigma_\beta^2}\right)^{\frac{1}{2}} \Phi\left(\frac{\tilde{\mu}_{jc} - L}{\tilde{\sigma}_c}\right) \exp\left(\frac{\tilde{\mu}_{jc}^2}{2\tilde{\sigma}_c^2}\right) + 1 - w}$$
;
        Sample  $\delta_{jc}$  from Bernoulli( $w_{jc}$ )
      end
      if  $\delta_{jc} = 1$  then
        Sample  $\beta_{jc}$  from a truncated normal distribution;
         $P(\beta_{jc} \mid \tilde{\mu}_{jc}, \tilde{\sigma}_c, \delta_{jc} = 1) \propto N(\tilde{\mu}_{jc}, \tilde{\sigma}_c^2) I(\beta_{jc} > L)$ 
      else
         $\beta_{jc} = 0$ 
      end
    end
  end
  Sample  $w$  from Beta( $\sum_{j,c} (1 - \delta_{jc}) + w_0, \sum_{j,c} \delta_{jc} + w_1$ )
end

```

Appendices

Short Dark Triad

See Table 21.5.

Table 21.5 Short dark triad items in the original item affiliation

	Statements
M1	It's not wise to tell your secrets.
M2	I like to use clever manipulation to get my way.
M3	Whatever it takes, you must get the important people on your side.
M4	Avoid direct conflict with others because they may be useful in the future.
M5	It's wise to keep track of information that you can use against people later.
M6	You should wait for the right time to get back at people.
M7	There are things you should hide from other people because they don't need to know.
M8	Make sure your plans benefit you, not others.
M9	Most people can be manipulated.
N1	People see me as a natural leader.
N2	I hate being the center of attention.
N3	Many group activities tend to be dull without me.
N4	I know that I am special because everyone keeps telling me so.
N5	I like to get acquainted with important people.
N6	I feel embarrassed if someone compliments me.
N7	I have been compared to famous people.
N8	I am an average person.
N9	I insist on getting the respect I deserve.
P1	I like to get revenge on authorities.
P2	I avoid dangerous situations.
P3	Payback needs to be quick and nasty.
P4	People often say I'm out of control.
P5	It's true that I can be mean to others.
P6	People who mess with me always regret it.
P7	I have never gotten into trouble with the law.
P8	I enjoy having sex with people I hardly know.
P9	I'll say anything to get what I want.

Empirical Item Parameters

See Table 21.6.

Table 21.6 DT3 element-wise means for \mathbf{B} and $\boldsymbol{\pi}$ for a random sample of $N = 5000$ respondents

	00	01	02	03	10	11	12	13	20	21	22	23	30	31	32	33
M1	2.0	0.5	0.6	0.8						0.3	0.1	0.2				
M2	-0.5	1.3	1.1	1.2	0.5				0.7			0.2	0.7			
M3	0.0	0.9	0.8	1.1	0.5				0.3				0.7			
M4	0.9	0.8	0.3	0.8	0.2											
M5	-0.6	1.8	1.3	1.2	0.2				0.8				0.2			
M6	-0.3	1.8	1.0	1.1	0.1				0.8							
M7	2.0	0.8	0.6	0.9					0.3		0.1					
M8	0.0	0.9	0.8	1.2					0.6		0.3		0.4			
M9	1.0	0.9	0.8	1.0	0.3				0.5				0.5			
N1	0.5		0.2	0.4	1.0	0.1						0.4	0.8			
N2	0.0				1.3								1.2		0.3	0.9
N3	-0.8	0.5	0.5	0.1	1.2			0.2				0.6	1.0			
N4	-0.6	0.4	0.5		1.1		0.7					0.3	1.1			
N5	0.1	0.7	0.5	1.1	1.0							0.1	1.0			
N6	0.3			0.4	0.9								1.2		0.8	0.4
N7	-0.6	0.4	0.3	0.7	1.4							0.3	1.1			
N8	0.7			0.6	0.6	0.2	0.5						1.4			
N9	0.1	0.7	0.6	0.6	0.8				0.2		0.2	0.9	0.7			
P1	-0.5	1.1	0.9	0.6					1.1			0.1				0.1
P2	0.3					0.4	0.4		0.9				0.4			

(continued)

Table 21.6 (continued)

	00	01	02	03	10	11	12	13	20	21	22	23	30	31	32	33
P3	-0.9	1.2	1.0	0.4					1.0			0.5				
P4	-0.8	0.6	0.6		0.2				1.1		0.3					
P5	0.6	0.9	0.9	0.7					0.9			0.2				
P6	-0.8	1.0	1.0	0.8	0.5				0.7			0.4	0.2			0.1
P7	-0.3								1.8	0.4						
P8	-0.8	0.8		0.5			0.6		1.4					0.6		
P9	-1.1	1.1	0.8	1.3	0.3		0.2		0.7		0.6	0.2	0.5			
π	0.03	0.03	0.06	0.03	0.03	0.09	0.13	0.06	0.07	0.15	0.09	0.03	0.03	0.09	0.06	0.02

Strict Identifiability Proof

The proof is mainly based on Kruskal’s theorem (Kruskal 1976, 1977) and the tripartition strategy proposed by Allman et al. (2009). We first introduce the probability matrix $\mathbf{H}(\mathbf{\Delta}, \mathbf{B})$ and its Kruskal rank in Definitions 1 and 2.

Definition 1 The class-response matrix $\mathbf{H}(\mathbf{\Delta}, \mathbf{B})$ is a matrix of size $M^K \times P^J$, wherein the rows denote attribute patterns and the columns denote response patterns. An arbitrary element (α_c, \mathbf{y}) in $\mathbf{H}(\mathbf{\Delta}, \mathbf{B})$ presents the probability of observing a response pattern \mathbf{y} from the latent class with attribute profile α_c :

$$H_{c,j}(\mathbf{\Delta}, \mathbf{B}) = P(\mathbf{Y} = \mathbf{y} \mid \beta_j, \alpha_c) = \prod_{j=1}^J \sum_{p=0}^{P-1} \theta_{cjp} \mathcal{I}(p = y_j), \tag{21.25}$$

Definition 2 (Kruskal Rank) The Kruskal rank of matrix \mathbf{H} is the largest number j such that every set of j columns in \mathbf{H} is independent. If \mathbf{H} has full row rank, the Kruskal rank of \mathbf{H} is its row rank.

Theorem 3 (Allman et al. 2009) Consider a general latent class model with r classes and \mathcal{J} items, where $J \geq 3$. Suppose all entries of $\boldsymbol{\pi}$ are positive. If there exists a tripartition of the item set $\mathcal{J} = 1, 2, \dots, J$ that divides \mathcal{J} into three disjoint, nonempty subsets $\mathcal{J}_1, \mathcal{J}_2$, and \mathcal{J}_3 such that the Kruskal ranks of the three class-response matrices $\mathbf{H}_1, \mathbf{H}_2$, and \mathbf{H}_3 satisfy

$$I_1 + I_2 + I_3 \geq 2r + 2,$$

then the parameters of the model are uniquely determined, up to label switching.

To prove the model parameters are uniquely determined, we need to find three subsets of items in SLCMs that satisfy Theorem 3. Suppose we have three disjoint, nonempty subsets $\mathcal{J}_1, \mathcal{J}_2$, and \mathcal{J}_3 , the marginal probability of response \mathbf{Y} can be reframed as a three-way tensor \mathbf{T} of dimension $P^{|\mathcal{J}_1|} \times P^{|\mathcal{J}_2|} \times P^{|\mathcal{J}_3|}$. Specifically, the $(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3)$ -th element in \mathbf{T} is the marginal probability of the products of the three subsets items:

$$\begin{aligned} T_{(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3)} &= P(\mathbf{Y}_{\mathcal{J}_1} = \mathbf{y}_1, \mathbf{Y}_{\mathcal{J}_2} = \mathbf{y}_2, \mathbf{Y}_{\mathcal{J}_3} = \mathbf{y}_3 \mid \mathbf{B}, \boldsymbol{\pi}) \\ &= \sum_c \pi_c P(\mathbf{Y}_{\mathcal{J}_1} = \mathbf{y}_1, \mathbf{Y}_{\mathcal{J}_2} = \mathbf{y}_2, \mathbf{Y}_{\mathcal{J}_3} = \mathbf{y}_3 \mid \mathbf{B}, \alpha_c) \\ &= \sum_c \pi_c P(\mathbf{Y}_{\mathcal{J}_1} = \mathbf{y}_1 \mid \mathbf{B}_1, \alpha_c) P(\mathbf{Y}_{\mathcal{J}_2} = \mathbf{y}_2 \mid \mathbf{B}_2, \alpha_c) \\ &\quad \times P(\mathbf{Y}_{\mathcal{J}_3} = \mathbf{y}_3 \mid \mathbf{B}_3, \alpha_c). \end{aligned} \tag{21.26}$$

In other words, tensor \mathbf{T} can be decomposed as an outer product of three vectors:

$$T = \sum_c \pi_c \mathbf{H}_{\alpha_c}(\mathbf{D}_1, \mathbf{B}_1) \otimes \mathbf{H}_{\alpha_c}(\mathbf{D}_2, \mathbf{B}_2) \otimes \mathbf{H}_{\alpha_c}(\mathbf{D}_3, \mathbf{B}_3),$$

where $\mathbf{H}_{\alpha_c}(\mathbf{D}_i, \mathbf{B}_i)$ represents a row vector of M^K in the class-probability matrix $\mathbf{H}(\mathbf{D}_i, \mathbf{B}_i)$ of size $M^K \times P^{\mathcal{J}_I}$.

Kruskal (1976, 1977) and Allman et al. (2009) state that if the sum of the Kruskal ranks of $\mathbf{H}(\mathbf{D}_1, \mathbf{B}_1)$, $\mathbf{H}(\mathbf{D}_2, \mathbf{B}_2)$, and $\mathbf{H}(\mathbf{D}_3, \mathbf{B}_3)$ is greater or equal to $2M^K + 2$, the tensor decomposition is unique up to the row rescaling and label switching. Now we will give the proof by showing the existence of three item subsets \mathcal{J}_1 , \mathcal{J}_2 , and \mathcal{J}_3 . If the corresponding item-attribute matrices \mathbf{D}_1 , \mathbf{D}_2 , \mathbf{D}_3 satisfy the structure of \mathbf{D}_s , \mathbf{D}_s , and \mathbf{D}^* in Theorem 2.2, the Kruskal rank sum of their class-probability matrix \mathbf{H}_1 , \mathbf{H}_2 , and \mathbf{H}_3 satisfies the minimum requirement $2M^K + 2$, and the model parameters are uniquely determined.

Proof 4 shows if \mathbf{D}_1 and \mathbf{D}_2 take the form of \mathbf{D}_s in Theorem 2.2, the class-response matrices $\mathbf{H}(\mathbf{D}_1, \mathbf{B}_1)$ and $\mathbf{H}(\mathbf{D}_2, \mathbf{B}_2)$ have a full Kruskal row rank of M^K .

Proof 4 \mathbf{D}_s is the defined item structure matrix of dimension $K \times M^K$ in Theorem 2.2, wherein the k -th item loads on all of the levels in attribute k , i.e., $\delta_k = (\delta_{k,0}, \dots, \delta_{k,M-1}) = \mathbf{1}$, and the corresponding item parameters $\beta_{k,m} \neq 0$ for $m \in \{0, \dots, M-1\}$. The class-response matrix $\mathbf{H}(\mathbf{D}_s, \mathbf{B}_s)$ of dimension $M^K \times P^K$ can be written as the Kronecker product of K sub-matrices \mathbf{H}_k :

$$\mathbf{H}(\mathbf{D}_s, \mathbf{B}_s) := \bigotimes_{k=1}^K \mathbf{H}_k = \bigotimes_{k=1}^K \begin{bmatrix} \Psi(\tau_1 - \mu_{k,0}) & \Psi(\tau_2 - \mu_{k,0}) - \Psi(\tau_1 - \mu_{k,0}) & \dots & 1 - \Psi(\tau_{(P-1)} - \mu_{k,0}) \\ \Psi(\tau_1 - \mu_{k,1}) & \Psi(\tau_2 - \mu_{k,1}) - \Psi(\tau_1 - \mu_{k,1}) & \dots & 1 - \Psi(\tau_{(P-1)} - \mu_{k,1}) \\ \Psi(\tau_1 - \mu_{k,2}) & \Psi(\tau_2 - \mu_{k,2}) - \Psi(\tau_1 - \mu_{k,2}) & \dots & 1 - \Psi(\tau_{(P-1)} - \mu_{k,2}) \\ \Psi(\tau_1 - \mu_{k,M-1}) & \Psi(\tau_2 - \mu_{k,M-1}) - \Psi(\tau_1 - \mu_{k,M-1}) & \dots & 1 - \Psi(\tau_{(P-1)} - \mu_{k,M-1}) \end{bmatrix}, \quad (21.27)$$

where \mathbf{H}_k can be viewed as the attribute-category matrix of dimension $M \times P$ for the k -th item in \mathbf{D}_s . In \mathbf{H}_k , the rows indicate the attribute levels and columns indicate the response categories.

Given the item parameters are all nonzero $\beta_{k,m} \neq 0$, the latent class mean parameter $\mu_{k,m}$ is different from each other given $\mu_{k,0} = \beta_{k,0}$, $\mu_{k,1} = \beta_{k,0} + \beta_{k,1}$, $\mu_{k,2} = \beta_{k,0} + \beta_{k,1} + \beta_{k,2}$, and $\mu_{k,M-1} = \beta_{k,0} + \beta_{k,1} + \dots + \beta_{k,M-1}$. Then, the rows in matrix \mathbf{H}_k are not linearly dependent so that \mathbf{H}_k is of full row Kruskal rank, namely, $\text{rank}(\mathbf{H}_k) = M$. For each item k in \mathbf{D}_s , we have $\text{rank}(\mathbf{H}_k) = M$. According to the property of Kronecker products, $\mathbf{H}(\mathbf{D}_s, \mathbf{B}_s) = \bigotimes_{k=1}^K \mathbf{H}_k$ is also full Kruskal row rank with $\text{rank}(\mathbf{H}(\mathbf{D}_s, \mathbf{B}_s)) = M^K$.

The following Proof 5 shows if \mathbf{D}_3 takes the form of \mathbf{D}^* in Theorem 2.2, the class-response matrix $\mathbf{H}(\mathbf{D}_3, \mathbf{B}_3)$ has Kruskal row rank of 2.

Proof 5 Condition (S2) in Theorem 2.2 ensures the main-effect components of each attribute to be nonzero in at least one item in \mathbf{D}^* so that \mathbf{D}^* can distinguish every pair of latent classes. Specifically, there must exist an item j_0 in \mathbf{D}^* that any two latent class c_1 and c_2 has different response probability matrix, i.e., $\Theta_{j_0, c_1} \neq \Theta_{j_0, c_2}$. Therefore, there must exist two rows in $\mathbf{H}(\mathbf{D}^*, \mathbf{B}^*)$ that are independent of each other, implying that the Kruskal rank of $\mathbf{H}(\mathbf{D}^*, \mathbf{B}^*)$ is at least 2.

With Proofs 4 and 5, we have $\text{rank}(\mathbf{H}_1) + \text{rank}(\mathbf{H}_2) + \text{rank}(\mathbf{H}_3) \geq 2M^K + 2$.

Generic Identifiability Proof

In the context of SLCMs, the item structure matrix $\mathbf{\Delta}$ is a sparse matrix, so the real parameter space should be of dimension less than $J \times M^K$. To be differentiated from Eq. 21.9, we denote the parameter space with a sparsity structure $\mathbf{\Delta}$ as

$$\Omega_{\mathbf{\Delta}}(\boldsymbol{\pi}, \mathbf{B}) = \{(\boldsymbol{\pi}, \mathbf{B}) : \boldsymbol{\pi} \in \Omega_1(\boldsymbol{\pi}), \mathbf{B} \in \Omega_{\mathbf{\Delta}}^*(\mathbf{B})\}, \quad (21.28)$$

where $\Omega_{\mathbf{\Delta}}^*(\mathbf{B})$ presents the parameter space for \mathbf{B} which have nonzero entry at position β when the corresponding $\delta = 1$. Then, we define the unidentifiable parameter set $C_{\mathbf{\Delta}}$ as

$$C_{\mathbf{\Delta}} = \{(\boldsymbol{\pi}, \mathbf{B}) : \mathbb{P}(\boldsymbol{\pi}, \mathbf{B}) = \mathbb{P}(\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{B}}), \quad (\boldsymbol{\pi}, \mathbf{B}) \not\sim (\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{B}}), \\ (\boldsymbol{\pi}, \mathbf{B}) \in \Omega_{\mathbf{\Delta}}(\boldsymbol{\pi}, \mathbf{B}), \quad (\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{B}}) \in \Omega_{\tilde{\mathbf{\Delta}}}(\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{B}})\}. \quad (21.29)$$

As Definition 2.3 stated, $\Omega_{\mathbf{\Delta}}(\boldsymbol{\pi}, \mathbf{B})$ is a generically identifiable parameter space if the unidentifiable set $C_{\mathbf{\Delta}}$ is of measure zero within $\Omega_{\mathbf{\Delta}}(\boldsymbol{\pi}, \mathbf{B})$.

Similar to the strict identifiability proof, we will use the tripartition strategy to find three item subsets \mathcal{J}_1 , \mathcal{J}_2 , and \mathcal{J}_3 that generate a tensor decomposition of \mathbf{D}_1 , \mathbf{D}_2 , and \mathbf{D}_3 . We proceed to show if \mathbf{D}_1 , \mathbf{D}_2 , and \mathbf{D}_3 satisfy the structure of \mathbf{D}_g , \mathbf{D}_g , and \mathbf{D}^* in Theorem 2.4, the Kruskal rank sum of the corresponding class-probability matrices \mathbf{H}_1 , \mathbf{H}_2 , and \mathbf{H}_3 satisfies the minimum requirement $2M^K + 2$.

Proof 6 For \mathbf{H}_1 and \mathbf{H}_2 , we use Theorem 7 to show that $\text{rank}(\mathbf{H}_1) = M^K$ and $\text{rank}(\mathbf{H}_2) = M^K$ hold almost everywhere in $\Omega_{\mathbf{D}_1}$ and $\Omega_{\mathbf{D}_2}$, respectively. Different from the Theorem 4 in Chen et al. (2020), we perform a transpose multiplication to the response-class matrix $\mathbf{H}(\mathbf{D}_g, \mathbf{B}_g)$ so that it can be transformed into a square matrix $\mathbf{H}(\mathbf{D}_g, \mathbf{B}_g)' \mathbf{H}(\mathbf{D}_g, \mathbf{B}_g)$ which has an accessible determinant function. Given Proof 10, we show $G_{\mathbf{D}}(\mathbf{B}) \rightarrow \mathbb{R}$ is a real analytical function of \mathbf{B} , and then we know $\lambda_{\Omega_{\mathbf{D}}}(A)$ has the Lebesgue measure zero. By Theorem 7, we can infer that $\mathbf{H}(\mathbf{D}_g, \mathbf{B}_g)$ is a full row rank matrix. Therefore, if $\mathbf{D}_1 \in \mathbb{D}_g$ and $\mathbf{D}_2 \in \mathbb{D}_g$, we have $\text{rank}(\mathbf{H}_1) + \text{rank}(\mathbf{H}_2) = 2M^K$ holds almost everywhere in $\Omega_{\mathbf{D}_1} \otimes \Omega_{\mathbf{D}_2}$.

Theorem 7 Given $\mathbf{D} \in \mathbb{D}_g$, the corresponding class-response matrix $\mathbf{H}(\mathbf{D}, \mathbf{B})$ is of full rank except some values of \mathbf{B} from a measure zero set with respect to $\Omega_{\mathbf{D}}$, i.e.,

$$\lambda_{\Omega_{\mathbf{D}}} \{\mathbf{B} \in \Omega_{\mathbf{D}} : \det[\mathbf{H}(\mathbf{D}, \mathbf{B})' \mathbf{H}(\mathbf{D}, \mathbf{B})] = 0\} = 0,$$

where $\lambda_{\Omega_{\mathbf{D}}}(A)$ denotes the Lebesgue measure of set A with respect to $\Omega_{\mathbf{D}}$.

Proposition 8 If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a real analytic function which is not identically zero, then the set $\{x : f(x) = 0\}$ has Lebesgue measure zero.

Remark 9 $G_{\mathbf{D}}(\mathbf{B}) = \det[\mathbf{H}(\Delta, \mathbf{B})' \mathbf{H}(\Delta, \mathbf{B})] : \Omega_{\mathbf{D}} \rightarrow \mathbb{R}$ is a real analytic function of \mathbf{B} .

Proof 10 $G_{\mathbf{D}}(\mathbf{B})$ is a composition function:

$$G_{\mathbf{D}}(\mathbf{B}) = \det[\mathbf{H}(\Delta, \mathbf{B})' \mathbf{H}(\Delta, \mathbf{B})] = h[(\theta_{\alpha_0}, \dots, \theta_{\alpha_{MK-1}})'(\theta_{\alpha_0}, \dots, \theta_{\alpha_{MK-1}})]$$

where $h(\theta) : [0, 1]^{K \times MK}$ denotes a polynomial function and θ_{α_c} represents the probability vector for the latent class α_c , which can be further written as the difference of two CDFs. A polynomial function is a real analytic function. Since the CDF is an integral of a real analytic function, the composition of real analytic functions (difference between two CDFs) is still a real analytic function. Furthermore, $h(\theta)$ is also a real analytic function of \mathbf{B} . $G_{\mathbf{D}}(\mathbf{B})$, as a determinant of $h(\theta)'h(\theta)$, is also a real analytic function.

Proof 11 For \mathbf{H}_3 , if \mathbf{D}_3 takes the form of \mathbf{D}^* in [S2], we can infer that there must exist an item j_0 in \mathbf{D}_3 that for any two latent classes c_1 and c_2 , we have $\mu_{j_0, c_1} \neq \mu_{j_0, c_2}$. Then we have at least two rows in matrix $\mathbf{H}(\mathbf{D}_3, \mathbf{B}_3)$ to be independent of each other, implying that the Kruskal rank of $\mathbf{H}(\mathbf{D}_3, \mathbf{B}_3)$ is at least 2. The exceptional case could exist when $\beta_{k, m} = 0$ holds for some k and m , which has Lebesgue measure zero with respect to Ω_{Δ^*} . Consequently, $\text{rank}(\mathbf{H}_3) \geq 2$ holds almost everywhere in Ω_{Δ^*} .

With Proofs 6 and 5, we have $\text{rank}(\mathbf{H}_1) + \text{rank}(\mathbf{H}_2) + \text{rank}(\mathbf{H}_3) \geq 2M^K + 2$ holds almost everywhere in $\Omega_{\Delta}(\pi, \mathbf{B})$.

References

- Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6), 3099–3132. <https://doi.org/10.1214/09-AOS689>
- Bolt, D. M., & Kim, J.-S. (2018). Parameter invariance and skill attribute continuity in the DINA model. *Journal of Educational Measurement*, 55(2), 264–280. <https://doi.org/10.1111/jedm.12175>
- Chen, Y., Culpepper, S., & Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika*, 85(1), 121–153. <https://doi.org/10.1007/s11336-019-09693-2>
- Chen, Y., & Culpepper, S. A. (2020). A multivariate probit model for learning trajectories: A fine-grained evaluation of an educational intervention. *Applied Psychological Measurement*, 44(7–8), 515–530. <https://doi.org/10.1177/0146621620920928>
- Chen, Y., Liu, Y., Culpepper, S. A., & Chen, Y. (2021). Inferring the number of attributes for the exploratory DINA model. *Psychometrika*, 86(1), 30–64. <https://doi.org/10.1007/s11336-021-09750-9>
- Cho, S. H. (2016). *An application of diagnostic modeling to a situational judgment test assessing emotional intelligence* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Culpepper, S. A. (2019). An exploratory diagnostic model for ordinal responses with binary attributes: Identifiability and estimation. *Psychometrika*, 84(4), 921–940. <https://doi.org/10.1007/s11336-019-09683-4>
- Culpepper, S. A., & Chen, Y. (2019). Development and application of an exploratory reduced reparameterized unified model. *Journal of Educational and Behavioral Statistics*, 44(1), 3–24. <https://doi.org/10.3102/1076998618791306>
- De La Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130. <https://doi.org/10.3102/1076998607309474>
- De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179–199. <https://doi.org/10.1007/s11336-011-9207-7>
- De La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333–353. <https://doi.org/10.1007/BF02295640>
- DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In R. L. B. P. D. Nichols & S. F. Chipman (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Routledge. <https://doi.org/10.4324/9780203052969>
- Furnham, A., Richards, S., Rangel, L., & Jones, D. N. (2014). Measuring malevolence: Quantitative issues surrounding the dark triad of personality. *Personality and Individual Differences*, 67, 114–121. <https://doi.org/10.1016/j.paid.2014.02.001>
- Furnham, A., Richards, S. C., & Paulhus, D. L. (2013). The dark triad of personality: A 10 year review. *Social and Personality Psychology Compass*, 7(3), 199–216. <https://doi.org/10.1111/spc3.12018>
- Garcia, D., & Rosenberg, P. (2016). *The dark cube: dark and light character profiles* (Vol. 4). PeerJ Inc. <https://doi.org/10.7717/peerj.1675>
- Glenn, A. L., & Sellbom, M. (2015). Theoretical and empirical concerns regarding the dark triad as a construct. *Journal of Personality Disorders*, 29(3), 360–377. https://doi.org/10.1521/pedi_2014_28_162
- Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions. *ETS Research Report Series*, 2008(2), i–25. <https://doi.org/10.1002/j.2333-8504.2008.tb02131>
- Hagenaars, J. A. (1993). *Loglinear models with latent variables* (No. 94). Sage Publications, Inc. <https://dx.doi.org/10.4135/9781412984850>

- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. <https://doi.org/10.1007/s11336-008-9089-5>
- Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment*, *21*(1), 28–41. <https://doi.org/10.1177/1073191113514105>
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. <https://doi.org/10.1177/01466210122032064>
- Karelitz, T. M. (2004). *Ordered category attribute coding framework for cognitive assessments* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Kruskal, J. B. (1976). More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, *41*(3), 281–293. <https://doi.org/10.1007/BF02293554>
- Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Its Applications*, *18*(2), 95–138. [https://doi.org/10.1016/0024-3795\(77\)90069-6](https://doi.org/10.1016/0024-3795(77)90069-6)
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. <https://psycnet.apa.org/doi/10.1007/BF02294535>
- Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, *42*(9), 1–21. <https://doi.org/10.18637/jss.v042.i09>
- McHoskey, J. W., Worzel, W., & Szyarto, C. (1998). Machiavellianism and psychopathy. *Journal of Personality and Social Psychology*, *74*(1), 192. <https://doi.org/10.1037/0022-3514.74.1.192>
- Narasimhan, B., Johnson, S. G., Hahn, T., Bouvier, A., & Kiêu, K. (2020). cubature: Adaptive multivariate integration over hypercubes [Computer software manual]. (R package version 2.0.4.1)
- Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. [https://doi.org/10.1016/S0092-6566\(02\)00505-6](https://doi.org/10.1016/S0092-6566(02)00505-6)
- Persson, B. N., Kajonius, P. J., & Garcia, D. (2019). Revisiting the structure of the short dark triad. *Assessment*, *26*(1), 3–16. <https://doi.org/10.1177/1073191117701192>
- Tatsuoka, K. K. (1987). Toward an integration of item-response theory and cognitive error diagnosis. In *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Lawrence Erlbaum Associates, Inc. <https://doi.org/10.4324/9780203056899>
- Templin, J. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis*. (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275. <http://dx.doi.org/10.1007/s00357-013-9129-4>
- Templin, J., Henson, R. A., et al. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. <https://doi.org/10.1037/1082-989x.11.3.287>
- Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, *32*(7), 559–574. <https://doi.org/10.1177/0146621607300286>
- Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, *14*(10), 2439–2468. <https://doi.org/10.1162/08997660260293292>
- von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, *2005*(2), i–35. <https://doi.org/10.1348/000711007x193957>
- von Davier, M. (2018). Diagnosing diagnostic models: From von neumann’s elephant to model equivalencies and network psychometrics. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 59–70. <https://doi.org/10.1080/15366367.2018.1436827>

- Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, 45(2), 675–707. <http://www.jstor.org/stable/44245820>
- Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, 113(523), 1284–1295. <https://doi.org/10.1080/01621459.2017.1340889>

Author Index

A

Abbasi, M.M., 78
Abdurrahim, S.H., 77
Achenbach, T.M., 346
Ackerman, T., 181–192
Ægisdóttir, S., 52
Aguinis, H., 81
Aiken, L.R., 56, 58
Akaïke, H., 408
Albert, J., 232
Allen, M.J., 198
Allman, E.S., 436, 437
Anastasi, A., 56, 59
Anderson, T.W., 154
Andrews, D.A., 48
Angelov, P., 85
Angoff, W.H., 186
Archer, C.O., 154
Arkes, H.R., 50, 62
Atkinson, J.W., 34, 35
Attneave, F., 22

B

Bacharach, V.R., 6
Bakker, M., 81
Bangerter, A., 50
Bar-Hillel, M., 289
Barnard, F.M., 7, 8
Bartolucci, F., 275, 277
Barton, M.A., 391
Bechger, T.B., 219–248
Béguin, A.A., 291
Beltiukov, A.P., 78

Bentler, P.M., 111, 135
Berger, S., 284–286
Bernaards, C.A., 144, 146
Bernoulli, D., 15
Binet, A., 5
Biot, J.-B., 13
Birch, D., 34, 35
Bird, H.R., 347
Birenbaum, M., 390
Birnbaum, A., 92, 203, 254, 333, 391
Blázquez-García, A., 81
Boake, C., 5
Bock, R.D., 15, 21, 22, 222
Böckenholt, U., 22
Bohrnstedt, G.W., 112, 113
Bolanowski, S.J., 24
Bollen, K.A., 92, 124
Bolsinova, M., 278, 307–323
Bolt, D.M., 414
Bonett, D.G., 135
Borgonovi, F., 141
Boring, E.G., 13, 14, 20
Borsboom, D., 3, 44, 89–106, 198
Boulton, M.J., 141
Bowman, M.L., 5
Bozdoğan, H., 376
Bradshaw, L., 390
Brennan, B.L., 287
Brennan, R.L., 181–183, 197, 203, 213, 258, 262
Brown, S.R., 8
Brown, W., 15, 183
Browne, M.W., 135
Buisman, R., 157

Bunch, M.B., 261, 286
Burt, C., 22

C

Camic, C., 20
Carroll, J.D., 155
Carver, C.S., 37
Cattell, J.M.K., 3–5, 20, 22, 23
Cattell, R.B., 8
Chalmers, R.P., 382, 405, 408
Chan, G.H.H., 355
Chang, J.-J., 155
Chang, Y.-P., 327–329, 332, 342
Chatterjee, S., 151, 154
Chen, J., 390
Chen, M., 346–364
Chen, W.-H., 314
Chen, Y., 413, 414, 418, 419, 422, 429–431, 438
Cheng, Y., 286, 287, 289, 300, 384
Chiu, C.-Y., 327, 328
Cho, E., 183
Cho, S.H., 430
Choi, K.M., 390
Christensen, K.B., 255
Cizek, G.J., 261, 286, 287
Clark, H.H., 51
Clauser, B., 163–179
Cohen, J., 359
Cohen, R.J., 56
Colonius, H., 24
Comrey, A.L., 147
Conijn, J.M., 346–364
Connors, C.K., 351–353
Cooper, C., 56, 57
Corrigan, B., 58
Cortina, J.M., 183
Costa, T.J.C.P. Jr., 92
Cowan, R.S., 23
Craig, S.B., 348, 349, 355
Cramer, A.O., 90, 104
CRAN, R., 70
Cronbach, L.J., 31, 48, 111, 120, 122, 182, 183
Csikszentmihalyi, M., 74
Cudeck, R., 135
Culpepper, S.A., 413–439
Curran, P.G., 355

D

Dailey, P.R., 120
Dalege, J., 103

Dana, J., 53, 61
Davenport, E.C., 183
Davidshofer, C.O., 56, 57, 62
Davison, M.L., 183
Dawes, R., 41
Dawes, R.M., 40, 42, 51–53, 58, 62
Dawid, A.P., 221, 227
De Boeck, P., 31–44, 314
de Boer, N.S., 98
De Candolle, A., 23
De Groot, A.D., 80
De la Torre, J., 364, 389–410, 413, 414, 430
De Leeuw, J., 21, 138
De Los Reyes, A., 346–348
Dehue, T., 20
Delaney, H.D., 200
Dempster, A.P., 146
Deng, W., 364
Deonovic, B., 244
DeSimone, J.A., 355
DiBello, L.V., 414
Diener, E.D., 135
Dietvorst, B.J., 53
Dinga, R., 82
Dirks, M.A., 346, 347
Doignon, J.P., 274
Dolan, C.V., 125, 126
Dolmans, T.C., 75
Domino, G., 56
Domino, M.L., 56
Dorans, N.J. 171
Dorz, S., 353
Douglas, J., 328, 329, 334, 393, 409, 413–439
Douglas, J.A., 430
Doyle, K.O., 5
Drake, F.L., 70
Drasgow, F., 349, 355
Drechsler, W., 8, 9
Drenth, P.J.D., 6, 47, 63
Dryden, I.L., 156
Dunn, T.J., 122
Dzhafarov, E.N., 24

E

Edelen, M.O., 371
Edgeworth, F.Y., 5
Efron, B., 154–155
Eggen, T.H., 373, 384
Eggen, T.J.H.M., 255, 284, 285
Ellis, J.L., 96, 102, 105, 212, 272, 327–342
Embretson, S.E., 72, 372, 390, 391
Emery, R.E., 40

Emons, W.H., 349, 355
 Emons, W.H.M., 195–213, 283–301
 Engzell, P., 284
 Epskamp, S., 101, 104
 Esary, J.D., 274
 Escalante, H.J., 77
 Euclid, 24
 Evers, A., 213, 264

F

Falk, C.F., 329, 342
 Falmagne, J.C., 274
 Faust, D., 41
 Fechner, G.T., 4, 6, 12–19
 Feldt, L.S., 164, 177, 182, 183, 196, 197,
 200–202
 Ferrando, P.J., 154
 Feskens, R., 265
 Feuerhahn, W., 9
 Feuerstahler, L.M., 329, 342
 Feurer, M., 83
 Fidell, L.S., 145
 Finnemann, A., 102
 Firth, D., 179
 Fischer, G.H., 291
 Fisher, R., 20
 Fisher, R.A., 144, 255
 Flanagan, J.C., 183
 Fleiss, J.L., 200
 Floridi, L., 70
 Fokkema, M., 38, 138
 Forcina, A., 275, 277
 Forster, O., 167
 Fox, J., 234, 235, 244
 Fox, J.-P., 315
 Friedman, S., 184
 Fuchs, L.S., 284
 Furby, L., 120, 122
 Furnham, A., 422, 423
 Furr, R.M., 6, 56, 57, 143

G

Galton, F., 4, 5, 20, 22, 23
 Garb, H.N., 40
 Garcia, D., 422
 Gelman, A., 235, 240, 242
 Geman, D., 219
 Geman, S., 219
 Gerstmann, E., 33
 Gescheider, G.A., 24
 Ghurye, S.G., 272

Girshick, M.A., 154
 Glas, C., 222
 Glas, C.A.W., 254, 284, 327
 Glenn, A.L., 423
 Gleser, G.C., 48
 Godin, B., 23
 Goldberg, L.R., 58
 Goldhammer, F., 307, 309, 320, 321
 Gonzalez, O., 385
 Goodman, L.A., 376
 Gore, L.R., 31–44
 Gorin, J.S., 390
 Gower, J.C., 151
 Grayson, D.A., 253, 272
 Green, B.F., 151
 Green, C.D., 19
 Green, P.J., 154
 Green, S.B., 183
 Gregory, R.J., 56, 58
 Grove, W.M., 49–52, 58, 62
 Grung, B., 145
 Gu, Z., 120–123, 125, 126, 130, 131, 137, 138,
 212
 Guay, J.P., 49, 53, 62
 Guilford, J.P., 15, 16, 18, 19
 Guion, R.M., 104
 Gulliksen, H., 186, 308
 Guo, F., 311
 Guttman, L., 99–102, 111, 123, 182, 196, 199,
 355

H

Haberman, S.J., 414
 Haertel, E.H., 391
 Hagen, G.F., 12
 Hagensaars, J.A., 414
 Haig, B.D., 91
 Hall, N.S., 20
 Hambleton, R.K., 72, 203, 290, 301
 Hanson, R.K.K., 62
 Harman, H.H., 143
 Harshman, R.A., 155
 Hart, S.G., 75
 Harvill, L.M., 196
 Haslbeck, J.M.B., 101
 Hastie, T., 83
 Hattie, J., 283
 Haviland, M.G., 209
 He, Q., 73, 74, 82
 He, S., 413–439
 Heise, D.R., 112, 113
 Heiser, W., 99
 Heiser, W.J., 3–24

Heitz, R.P., 311
 Hemker, B., 265
 Hemker, B.T., 251–268
 Hendricks, J., 175
 Henson, R.A., 392, 413, 414, 430, 431
 Herbart, J.F., 12
 Hermans, H.J.M., 42
 Hessen, D.J., 111–117
 Heymans, G., 6
 Highhouse, S., 50, 62, 63
 Hills, P., 141
 Ho, T.K., 373
 Hock, H.S., 16
 Hogan, T.P., 56, 58
 Hogarth, R.M., 52
 Hoijtink, H., 277
 Holland, P.W., 91–92, 96, 165, 196, 203, 212,
 272–274, 279
 Holt, R.R., 58
 Hopster-Den Otter, D., 195
 Hornstein, G.A., 24
 Hoskens, M., 165, 196, 203, 314
 Howard, G.S., 120
 Hoyt, C., 182, 200
 Hu, S., 77
 Hubert, L., 3
 Hulliger, B., 355
 Hunt, T., 131
 Hunter, I., 8
 Hunter, J.E., 49
 Hutter, F., 83
 Huynh, H., 253, 272, 333

J

Jabrayilov, R., 209, 210
 Jackson, P.H., 111, 199
 Jacobson, N.S., 208, 212
 James, G., 82, 83
 Jarjoura, D., 200
 Jastrow, J., 20
 Jennrich, R.I., 154
 Jensen, A.R., 92
 Jensen, J.L.W.V., 169
 Jeon, M., 44
 Jiang, Z., 232
 Joag-Dev, K., 272, 273
 Jodoin, M.G., 284
 Jogdeo, K., 274
 Johnson, J.A., 349, 355, 356
 Johnson, R.A., 355
 Jones, D.N., 423
 Jones, E., 173
 Jones, L.V., 6, 15, 21, 22

Josse, J., 144, 145, 154
 Junker, B., 232
 Junker, B.W., 6, 92, 96, 102, 105, 275, 328,
 329, 331, 334, 391, 410, 413

K

Kachur, A., 77
 Kahneman, D., 40, 42, 53, 54
 Kam, C.C.S., 355
 Kane, M.T., 84, 182
 Kant, I., 8, 12
 Kaplan, R.M., 56, 58
 Karelaia, N., 52
 Karelitz, T., 390, 392
 Karelitz, T.M., 414
 Karlin, S., 272
 Kass, R.E., 278
 Kaufman, J., 357
 Kaufmann, L.M., 120
 Kazdin, A.E., 346
 Kelley, T., 20
 Kelley, T.L., 20
 Kelly, G., 42
 Keszler, N.S., 77
 Kiers, H.A.L., 145, 149, 154, 155, 157
 Kim, D., 286
 Kim, J.-S., 414
 Kim, S., 183
 Kingsbury, G.G., 370
 Klein Entink, R.H., 234, 309
 Klein, G., 53, 54
 Kline, T.J., 56, 57
 Klugkist, I., 277
 Kohen, E.S., 58
 Kolen, M.J., 164, 177, 203, 258, 262, 287
 Kouwer, B.J., 5
 Kraemer, H.C., 347
 Krathwohl, D.R., 51
 Kreiner, S., 255
 Kroonenberg, P.M., 144, 150–153, 155–157
 Kruskal, J.B., 436–439
 Kruyen, P.M., 291
 Kuder, G.F., 182
 Kuijpers, R.E., 212, 283–301
 Kuncel, N.R., 48, 50, 52, 58

L

Lambert, Z.V., 154
 Lamiell, J.T., 42
 Laming, D., 317
 Lampinen, J., 423
 Lavine, M., 278

- Leahy, S., 285
 Leary, D.E., 9, 12
 Lee, H.B., 147
 Lee, W.-C., 196, 197
 Lee, Y.-S., 390, 393
 Leibniz, G.W., 8, 9
 Lek, K., 283–301
 Lek, K.M., 196, 197
 Lewinsohn, P.M., 37
 Lewis, C., 253, 271, 286
 Lewis, J.B., 376, 382
 Lezak, M.D., 32
 Li, Q., 373
 Liang, H., 70
 Liem, C.C.S., 75
 Ligtvoet, R., 271–280
 Link, S.W., 16
 Linn, R.L., 120
 Linting, M., 151, 154
 Linzer, D.A., 373, 376, 382
 Little, R.J.A., 143
 Little, T.D., 154
 Liu, J., 232
 Liu, R., 390
 Livingston, R.B., 56
 Livingston, S.A., 286, 287
 Lloyd, S., 385
 Lockwood, J.R., 179
 Loevinger, J.A., 272
 Loeys, T., 309
 Lonigan, C.J., 351
 Lord, F., 17, 22
 Lord, F.M., 71, 72, 111–112, 116, 117, 120,
 165–167, 183, 196, 198, 199, 203,
 204, 258, 284, 287, 292, 391
 Lorenza-Seva, U., 154
 Lovett, B.J., 346, 363
 Lu, Y., 314
 Luce, R.D., 24, 33
 Luderer, M., 363
 Ludwig, K., 6–8
 Luecht, R., 181–192
 Luecht, R.M., 285
 Lukociene, O., 376
 Lumsden, J., 22, 212
 Lyddon, W.J., 42
- M**
- Ma, W., 405, 408
 Ma, Y., 181–192
 Maassen, G.H., 208
 Mackinnon, S.P., 135
 MacPhillamy, D.J. 37
 Maeda, Y., 183
 Magidson, J., 376
 Magis, D., 370
 Manne, R., 145
 Mardia, K.V., 156
 Marianti, S., 315
 Maris, E., 414
 Maris, G., 219, 230, 237, 244, 247, 248, 253,
 256
 Maris, G.K.J., 219–248
 Markus, M.T., 151, 154
 Marsman, M., 90, 101, 102, 105, 219–248
 Martel, M.M., 346–348, 363
 Martin, A.D., 424
 Masters, G.N., 222
 Mata, I., 141
 Maxwell, S.E., 200
 Maydeu-Olivares, A., 21
 McCaffrey, D.F., 179
 McConnel, K., 122
 McCrae, R.R., 92
 McCutcheon, A.L., 376
 McDonald, R.P., 112, 113, 124, 184
 McFall, R.M., 31, 33, 39
 McHoskey, J.W., 423
 McKeown, B., 8
 McKoon, G., 33
 McReynolds, P., 6–8
 Meade, A.W., 348, 349, 355
 Meehl, P.E., 49, 51–53, 57–59, 62, 63
 Mei, M., 10
 Meijer, R.R., 47–63, 72, 82, 348–350, 355
 Mellenbergh, G.J., 5, 92, 96, 103, 120, 196,
 199, 254
 Meng, X.B., 309, 317, 318
 Meredith, W., 121, 127
 Meredith, W., 92
 Meulman, J., 3
 Meulman, J.J., 144, 145
 Michell, J., 24
 Milan, L., 151, 154
 Milkman, K.L., 50
 Miller, G.A., 13–15
 Miller, L.A., 56, 57
 Millsap, R.E., 37, 92
 Minchen, N., 390
 Miotto, R., 70
 Mislevy, R., 219
 Mislevy, R.J., 285
 Möbius, A.F., 18, 21
 Mokken, R.J., 92, 253, 271, 273, 279, 327, 382
 Molden, D.C., 35
 Molenaar, D., 119–138, 317, 370
 Molenaar, I., 96

Molenaar, I.W., 72, 201, 255, 271, 272, 291
 Morgan, D.L., 286, 287
 Morris, S.B., 52
 Morton-Bourgon, K.E., 62
 Mülberger, A., 6
 Mulder, J., 278
 Müller, G.E., 18, 24
 Muraki, E., 222, 256
 Murphy, G.L., 51
 Murphy, K.R., 53, 56–58, 62
 Murray, D.J., 13, 14, 19, 24
 Murray, H.A., 35, 36
 Murray, I., 219, 223

N

Nagelkerke, E., 376
 Nagy, G., 321
 Narasimhan, B., 425
 Nelson, J.M., 346, 363
 Neumann, M., 47–63, 84
 Nicewander, A., 196
 Niessen, A.S.M., 349, 355
 Niessen, M., 47–63
 Nolan, K.P., 50, 63
 Novick, M.R., 71, 111–112, 116, 117, 120,
 183, 196, 198, 199, 287
 Nugester, R.J., 285

O

Ogasawara, H., 154
 Ohm, G.S., 13
 O'Neil, C., 70
 Oort, F., 138
 Oort, F.J., 121, 127
 Oosterwijk, P., 199, 202, 203, 212
 Ortner, T., 317
 Oshima-Takane, Y., 144, 145

P

Pál, J., 141
 Panch, T., 70
 Parent, G., 49, 53, 62
 Patz, R., 232
 Paulhus, D.L., 422, 423
 Payton, J., 207
 Pearl, J., 91, 92, 96, 98
 Peirce, C.S., 20
 Persson, B.N., 422, 430
 Pfadt, J.M., 122, 124, 125
 Pratt, C.C., 18
 Purpura, D.J., 351

Q

Qualls, A.L., 164, 177
 Qualls-Payne, A.L., 197

R

Rabbitt, P., 317
 Racine, J., 373
 Raftery, A.E., 278
 Raiker, J.S., 363
 Raju, N.S., 196
 Ramsay, J.O., 329, 334
 Ramul, K., 6, 9, 10, 12
 Ranger, J., 310, 317, 318
 Rasch, G., 90, 91, 95, 220, 222, 253, 391
 Ratcliff, R., 33
 Rawal, G., 78
 Raykov, T., 154
 Reckase, M., 244
 Reckase, M.D., 184
 Reeve, B.B., 371
 Reichenbach, H., 91, 92, 96
 Reise, S.P., 72, 92, 93, 209, 372, 391
 Revelle, W., 111, 186
 Reynolds, C.R., 56
 Rhemtulla, M., 125, 126
 Richardson, M.W., 182
 Rinott, Y., 272
 Röber, T.E., 75, 77
 Rodgers, B., 317
 Rodriguez, M.C., 183
 Roelofs, E.C., 265
 Rohatgi, A., 175
 Rose, N., 332
 Rosenbaum, P.R., 91–92, 96, 272–274, 331
 Rosenberg, P., 422
 Rosenberg, S., 42
 Rosenthal, J., 234
 Rosseel, Y., 131
 Roulin, N., 50
 Rozeboom, W.W., 24
 Rozell, E.J., 50
 Rubin, D.B., 143, 144, 146, 150, 157, 219
 Rulon, P., 183
 Russel, J.T., 48
 Russell, E.J., 255
 Ryan, A.M., 50

S

Saccuzzo, D.P., 56, 58
 Sackett, P.R., 50, 53
 Saffir, M., 22
 Samejima, F., 130, 203

- Sanford, E.C., 19
 Santos, K.C., 389–410
 Sarbin, T.R., 49
 Schafer, J.L., 150
 Scheier, M.F., 37
 Scherer, R., 309
 Schermelleh-Engel, K., 135
 Schervish, T., 278
 Schleicher, A., 80
 Schmidt, F.L., 49
 Schmiedek, F., 4
 Schöner, G., 16
 Schwab, A.P., 62
 Schwartz, C.E., 121
 Schwarz, G., 376, 408
 Sellbom, M., 423
 Shakespeare, W., 267
 Shane, F., 33
 Shang, Z., 419
 Shavelson, R.J., 183
 Shaw, A., 309
 Sheynin, O., 19, 20
 Shireman, E.M., 385
 Shrout, P.E., 200
 Sibley, M.H., 363, 364
 Sijtsma, K., 6, 34, 44, 47, 63, 71, 72, 82,
 89, 90, 92, 94–97, 103, 105, 111,
 122, 124, 125, 144, 146, 182, 183,
 196–202, 208, 211, 271, 275, 286,
 327–329, 331, 382, 391, 410, 413
 Simon, T., 5
 Simonton, D.K., 22
 Sireci, S.G., 314
 Skinner, B.F., 95
 Slinde, J.A., 120
 Smith, A.F., 348
 Smith, S.R., 346–348
 Smits, N., 369–386
 Soares, E., 85
 Sočan, G., 183
 Spearman, C., 6, 16, 20, 92, 112, 183
 Sprangers, M.A., 121
 Staveland, L.E., 75
 Stephenson, W., 8
 Stevens, S.S., 14
 Stigler, S., 3
 Stigler, S.M., 5, 6, 13, 14, 16–18, 20
 Straat, J.H., 283–301
 Straetmans, G.J.J.M., 284
 Streb, M.J., 33
 Sturm, T., 12, 24
 Suhr, J.A., 356
 Suppes, P., 273
 Susan, A., 47–63
 Swaminathan, H., 72, 203, 290
 Sweller, J., 298
 Swerdijk, M.E., 56
 Szabó, A.T., 23
- T**
 Tabachnick, B.G., 145
 Takane, Y., 21, 138, 144, 145
 Tanner, M., 220
 Tashakkori, A., 39
 Tatsuoaka, K.K., 274, 390, 414
 Taylor, B., 167
 Taylor, H.C., 48
 Teddie, C., 39
 Templin, J., 232, 414
 Templin, J.L., 392, 413, 431
 Ten Berge, J.M.F., 111, 151, 152, 183
 Tendeiro, J.N., 355
 Terpstra, D.E., 50
 Thiebes, S., 70
 Thissen, D., 6, 22, 203, 314, 372
 Thomas, D.B., 8
 Thomasius, C., 5–8, 22
 Thompson, B., 181, 199
 Thomson, G.H., 15, 20–21
 Thorndike, E.L., 19, 20
 Thorndike, R.L., 197
 Thurstone, L.L., 19, 21, 22
 Tibshirani, R.J., 154
 Tierney, L., 222, 226
 Tijmstra, J., 277, 278, 307–323, 341
 Tijssen, R., 3
 Timmerman, M.E., 154
 Timperley, H., 283
 Ting, K., 92
 Titchener, E.B., 15, 19, 20
 Tjoe, H., 390, 407
 Torgerson, W.S., 22
 Townsend, J.T., 31, 33, 39
 Traub, R.E., 32, 186
 Truax, P., 208, 212
 Tucker, L.R., 155
 Tuerlinckx, F., 34–36
 Tversky, A., 42
- U**
 Ulitzsch, E., 321
 Ünlü, A., 272, 333
 Urban, F.M., 16, 18, 21
 Urbina, S., 56

V

Vacha-Haase, 181
 Van Bork, R., 40, 105
 Van Borkulo, C.D., 90, 99
 Van Buuren, N., 373, 384
 Van Buuren, S., 150
 Van De Schoot, R., 196, 197
 Van den Brink, W.P., 254
 Van den Wollenberg, A.L., 212
 Van der Ark, L.A., 71, 125, 196–200, 202, 271, 278, 328, 346–364, 369–386, 410
 Van der Heijden, P.G.M., 6
 Van der Linden, W.J., 48, 73, 74, 203, 234, 284, 285, 301, 308–311, 327
 Van der Maas, H.L., 94, 237, 240, 244, 247, 248
 Van der Maas, H.L.J., 103
 Van der Palm, D.W., 373, 385
 Van Ewijk, H., 346–364
 Van Ginkel, J.R., 141–158, 355
 Van Mechelen, I., 42
 Van Rossum, G., 70
 Van Wingerde, B., 156
 Vehtari, A., 423
 Veldkamp, B., 80
 Veldkamp, B.P., 69–85, 285
 Verhelst, N., 222
 Verhelst, N.D., 254, 255, 260, 285
 Vermunt, J.K., 277, 278, 373, 376
 Verstralen, H.H.F.M., 260
 Verweij, A.C., 278
 Vidal, F., 8, 10, 12, 24
 Vispoel, W.P., 285
 Voigt, P., 81
 von Davier, A., 167
 von Davier, M., 163–179, 414
 Von dem Bussche, A., 81
 Von Mises, R., 19
 Von Plato, J., 20
 Von Rhein, D., 351, 357
 Vorst, H.C.M., 370
 Voulodimos, A., 70
 Vrieze, S.I., 50

W

Wagenmakers, E., 240
 Wainer, H., 284, 285, 314, 370, 371
 Waldorp, L.J., 101
 Walker, H.M., 6
 Walkup, D.W., 272, 275, 276
 Wallace, D.L., 272
 Waller, N.G., 92, 93
 Walls, B.D., 363
 Wanders, R.B., 349, 363

Wang, C., 309, 384
 Wang, M., 192
 Wang, M.C., 207
 Warm, T.A., 179, 335
 Warrens, M.J., 272
 Weber, E.H., 14, 16
 Webster, D.S., 23
 Weisberg, H.F., 184
 Weiss, D.J., 284, 370
 Whittaker, J., 151, 154
 Wichern, D.W., 355
 Wicherts, J.M., 81
 Wickham, H., 186
 Wiersma, E., 6
 Wiggins, N., 58
 Wilhelm, O., 4
 William, D., 284, 285
 Wilks, S.S., 333
 Williams, B.J., 120
 Williams, K.M., 422
 Wilson, M., 39
 Wingersky, M.S., 203
 Wissler, C., 22
 Wolff, C., 8–12, 23
 Wolter, K., 168
 Woodhouse, B., 111, 199
 Woodruff, D., 164, 177, 197, 200, 203
 Wright, W.F., 6
 Wu, P., 285
 Wundt, W., 3, 4, 24

X

Xie, Y., 20
 Xu, G., 309, 419
 Xu, X., 329, 342

Y

Yan, D., 284, 285, 373, 385
 Yang, Y., 183
 Yen, W.M., 198, 314
 Yin, P., 174–178
 Ysseldyke, J., 284
 Yu, M.C., 58

Z

Zanotti, M., 273
 Zehner, F., 307
 Zenisky, A.L., 285, 314
 Zhang, H., 5
 Zhu, X.X., 70
 Zinbarg, R.E., 111, 184
 Zudini, V., 14, 24
 Zwitser, R.J., 253, 256

Subject Index

A

Accountable AI, 70
Accuracy, 44, 53, 54, 57, 70, 77, 198, 204, 205, 209, 210, 237, 240, 248, 285–290, 293, 296, 297, 299, 308, 311, 315, 317, 319–322, 330, 348, 370, 423–425, 428–430
Adaptive testing, 285, 286, 298
Akaike information criterion (AIC), 376, 408
Amsterdam Chess Test (ACT), 164, 171–176, 237–241, 244
A-parameter/slope parameters, 205, 254, 256, 257, 259–261, 264–268, 291, 292, 371
Argument-based approach, 84
Artificial intelligence (AI), vi, 69–85
Attention, 75
Attention deficit hyperactivity disorder (ADHD), 346–364

B

Bayes factor, 272, 277–279
Bayesian information criterion (BIC), 376, 408
Bayesian statistics, 73, 74, 219
Bias, 5, 38, 41, 42, 63, 70, 77, 82, 120, 122, 123, 126, 132–134, 138, 146–149, 157, 163–179, 201, 205–208, 313, 320–322, 348–350, 360, 370, 392, 423, 429
Bias-corrected and accelerated (BCa) method, 154, 155
Bibliometrics, 23

Bifactor model, 184

Bootstrap confidence interval, 149, 154

C

Calibration, 266, 341, 370–376, 382
Careless responding, 348, 349, 355, 369
Carry-over effects, 122, 126, 129–134
Causality, 94, 96–98
Causal relation, 96–98
Centroid, 152, 153, 155, 186–191
Change, vii, 10, 14, 16, 17, 24, 31–33, 37, 38, 43, 75, 92, 94, 98, 101, 119–138, 151, 173, 175, 176, 179, 191, 207–212, 256, 261, 265–268, 287, 427
Classical test theory (CTT), vi, vii, 16, 32, 39, 48, 71, 72, 84, 103, 120–126, 131–133, 136–138, 164–169, 173, 177, 179, 182, 196, 198, 199, 203, 251–253, 263, 393
Classification certainty, 287–289, 294
Classification errors (CE), 147, 148
Classification tests, 298
Clinical prediction, 57
Coefficient alpha, 34, 37, 111, 112, 117, 182, 213
Coefficient (Cronbach's) alpha, vi, 34, 44, 103, 122, 123, 190, 191, 196, 199, 200, 353
Coefficient omega, 111–117
Cognitive diagnosis, 342
Cognitive diagnosis modelling, 328, 393

- Cognitive diagnosis models (CDMs), vi, 328, 332, 389–410
- Combination rules, 51, 146, 150–151, 155–158
- Communality, vii, 111–117
- Comparative fit index (CFI), 135–138
- Compensatory IRT model, 184
- Component
loading, 144–158
- Computational feasibility, 376
- Computer-adaptive testing, 327–342
- Computerized adaptive testing, vii, 72, 300, 369–386
- Concordance, 164, 171–179
- Conditional association, 105, 272
- Conditional independence (CI), 34, 73, 92, 96, 273, 313–320, 327
- Conditional standard error of measurement (CSEM), vii, 164, 177, 195–213
- Confidence ellipse, 154
- Connors' rating scales, 360
- Consistency, 33–37, 96, 183, 197, 201–203, 286–290, 292–295, 298
- Constant method, 17, 20
- Continuity, 9, 10, 12
- Convex hull
peeling, 154
- Core array, 155
- Correlated errors, 120, 122, 124, 126–127
- Correlated residuals, 127, 129
- Coverage percentage, 153, 155
- Cronbach's alpha, vi, 34, 44, 103, 122, 123, 190, 191, 353
- Cross validation, 83, 423
- Curse of dimensionality, 385
- Cut-off scores, 258, 261–264, 267, 268, 287, 356, 359
- D**
- Data collection, 42, 51–52, 54, 56, 75, 78, 143, 144, 363, 423
- Data combination, 51–52, 54, 56
- Data mining, 83, 84
- Data visualization, 84
- Decision-making, vi, 41, 47–63, 196, 211, 286, 289, 291, 294, 298, 301, 346, 348–351, 356, 360–363
- Decision rules, 52, 54
- Decision trees, 373, 374, 385
- Deep neural nets, 75, 76, 84, 85
- Density estimation, 375, 376, 385
- Deterministic input, noisy “and” gate model (DINA model), 391, 410, 413
- Diagnostic algorithm, 350, 351, 357–358, 360–363
- Difficulty matching
latent, 332, 333, 336, 339, 340
observed, 333, 336, 339–341
- Difficulty parameter, 17, 22, 100, 224, 255, 290, 316, 331, 333, 337, 338, 391, 407
- Dimensionality, vii, 101, 123, 131, 138, 141, 183, 186, 191, 192, 268, 385, 410
- Dimensional structure, 141, 142
- Discrepancy, vii, 323, 332–334, 346–364
- Discrimination parameter, 130, 185, 186, 192, 224, 290, 333, 337, 338, 393, 404–407
- Dissemination of research findings, 50
- Dissimilarity, 24
- Distance, 16–19, 24, 32, 76, 151, 329, 354, 355, 405
- Divisive latent class model, 373, 385
- E**
- Educational survey, 219, 220, 222, 223, 241
- Educational testing, 5, 47–63, 91
- Eigenvalue, 142, 145, 146, 184
- EM-covariances, 146, 148
- Equal appearing intervals, 18, 19, 21
- Equating, vii, 163–179, 241, 258–265, 267, 268, 287
- Error score, 120
- Evidence-based applied science, 50
- Expectation-maximization (EM) algorithm, 146
- Expected a posteriori (EAP), 241, 372
- Expected moments, 335–336
- Explainability, 69, 70, 80, 83
- Explainable AI, 70
- Explicability, 70
- Explicable AI, 70
- Extreme responding, 356, 363
- F**
- Face recognition, 4, 75–78, 81
- Factor analysis, vii, 4, 22, 42, 48, 99–102, 119–138, 429
- Fairness, 80, 83
- Fechner's law, 13, 14, 19, 21, 24
- FlexCAT, vii, 370, 372–380, 382–385
- Four-parameter logistic (4PL) model, 391, 403, 405–409
- Full information maximum likelihood, 146, 149

G

Generalized DINA (G-DINA) model, 392, 395, 397, 406, 408
 Generalized Procrustes analysis, 151–152, 155, 156, 158
 Gibbs sampler, 220, 232, 234, 237, 239, 241–243, 278, 279, 429
 Gibbs sampling, 219, 415, 421–422, 431, 432
 Goodness of fit, 138, 376, 385
 GPA . sps, 156
 Graded response model (GRM), 130, 203, 205
 Greatest lower bound, 111, 199
 Guttman's Lambda-1, 202
 Guttman's Lambda-2, 111, 202
 Guttman's Lambda-3, 182, 200, 202

H

Hamming distance, 329
 Heron's rule, 153
 Hierarchical model, vii, 307–323
 Holistic prediction, 49, 51–55, 59, 63
 Hyperparameter, 83, 278, 421, 424, 432

I

Ideal response patterns, 329, 332
 Identifiability, 95, 376, 385, 415, 417–421, 429, 430, 436–439
 Individual change assessment, 198, 207–210
 Individual decision making, 211, 286, 287, 289, 291, 294, 298, 346, 348–351, 363
 Individual differences, 4, 6, 11, 18, 23, 44, 93, 430
 Informant discrepancy, 347–349, 352, 353, 356, 359, 363
 Informant report, 346, 347, 351
 Intensity, 9–11, 18, 23, 35, 235, 308, 310, 311, 316
 International Test Commission, 48, 60
 Interpretability, 83, 84
 Interpretable AI, 70, 83
 Intuitive prediction, 49
 Invariant item ordering, 331–333
 Isotonic regression, 334, 341
 ITC test guidelines, 61
 Item difficulty, 17, 22, 91, 101, 185, 187–190, 262, 284, 287, 290, 308, 310, 316, 331–333
 Item response theory (IRT), v, vi, vii, 17, 22, 24, 34, 36, 43, 44, 72, 73, 84, 90–96, 99–102, 165, 179, 184, 192, 197, 201, 203, 209, 212, 220–224, 229,

230, 232, 237, 241, 242, 251–253, 255, 258, 267, 268, 272, 273, 301, 308, 309, 311, 312, 314, 315, 319, 320, 322, 327–329, 349, 354, 355, 364, 369–386, 389–410

Item-rest regression, 329, 330, 333–336, 339–342
 Item selection, 330, 336, 338, 341, 371, 372, 374, 385

J

Jensen's inequality, 169
 Just noticeable difference (jnd), 14, 16, 18, 20, 24

K

Kernel smoothing, 329, 342
 Knowledge dissemination, 47–63

L

Landmark, v, 75–77, 81, 156
 Latent class model (LCM), 373, 375–382, 384, 385, 391, 436
 Latent distributions, 191, 292
 Latent factor, 112, 117, 124, 125, 127, 129, 130
 Latent regression, 244
 Latent trait estimation
 expected a posteriori, 372
 maximum likelihood, 372
 Latent variable(s), 34, 38, 39, 90–93, 96, 99–102, 105, 106, 165, 203, 272, 275, 290, 308–310, 314–317, 322, 327, 332, 375, 421, 424, 429, 430
 Latent variable model, 34, 92, 96, 100, 103, 104, 272, 273, 393
 Law of comparative judgment, 21
 Layer, 75
 Limen, 16, 18, 20
 Linear equating, 164
 Linear function, 163, 168, 175
 Listwise deletion, 144, 145, 148, 149
 Loading, 38, 111–113, 115, 124, 128–130, 136–138, 143–145, 147, 149, 151–155, 184, 191, 192, 425
 Loading plot, 152, 153
 Local independence, 72, 92, 203, 253, 267, 290, 314, 373, 382, 383, 416
 Local non-negative dependence, 273
 Local optimum, 376, 385
 Log-odds ratios, 272, 275–276

- Long-string index, 354, 355
 Lower-bound reliability indices, 198–200
 I_z^p and I_z^p person-fit index, 354, 355, 363
- M**
- Magnitude, 10, 14, 15, 17, 19, 21, 24, 164, 169, 175, 179, 187, 190, 191
 Mahalanobis distance (MD), 354, 355, 359, 360
 Malingering, 348, 349, 356
 Markov Chain Monte Carlo (MCMC), 415, 421–424, 429, 430
 Maximum likelihood (ML), 116, 117, 129, 131, 146, 149, 229, 277, 332, 335, 336, 372
 Measurement errors, 44, 71, 95, 103, 112, 113, 121, 122, 124, 125, 165, 182, 196, 198, 207, 252, 263, 289, 294, 301, 310
 Measurement invariance, 33, 37–39, 43, 92, 121, 127–130, 135–138
 Measurement precision, 120, 196–200, 211, 212, 298, 300
 Mechanical prediction, 49, 52, 55
 Median, 17, 19, 21, 259, 337, 383
 Memory, 4, 11, 23, 35, 93, 97, 234, 360, 362
 Mental testing, 4–6, 20, 22
 Method of right and wrong cases, 17, 19, 20
 Metropolis algorithm, 222
 MH model, vii, 271–275, 279, 330
 Missing at random (MAR), 143, 144, 146, 148
 Missing completely at random (MCAR), 143–146, 148
 Missing data, vii, 141–158, 351
 Missing data passive, 144–146, 148
 Missingness mechanism, 143–144, 148
 Missing not at random (MNAR), 143, 144, 148
 Model assumptions, 272
 Model fit, 82, 120, 126, 129–131, 135–137, 197, 255, 256, 259–260, 263, 313, 349, 408, 415, 425
 Mokken scale analysis, 271, 272
 Monotone homogeneity (MH), vii, 255, 327–331, 334–336
 Monotone latent variable model, 91, 92, 273
 Monotonicity, 72, 92, 165, 273, 274, 327, 382, 383, 398, 403, 415, 419–421, 429
 Monotonicity of the substest characteristic curves, 274
 Monotonic polynomial model, 342
 Multidimensional discrimination MDISC, 184–191
- Multidimensionality, 120, 122–124, 137, 138, 183, 191, 192, 266
 Multidimensional models, 104, 241, 242
 Multinomial model, 277, 278
 Multinomial parameters, 272
 Multiple imputation
 combination rules, 150–151, 155–158
 Multivariate total positivity, 272
- N**
- National Council on Measurement in Education (NCME), 48
 Neighborhood score, 101–103
 Network model, 100–102, 104
 NeuroIMAGE study, 351, 357, 361–364
 Noise, 40, 81, 103, 191
 Non-linear equating, 163–179
 Non-negative partial correlations, 272
 Nonparametric classification (NPC), 328, 329
 Nonparametric item response theory, vi, 201
 Normal ogive, 17, 18, 21, 22, 224, 232
- O**
- One-parameter logistic model (OPLM), 222, 253–263, 265, 266, 290, 391, 403, 408, 409
 Operations triad model, 347, 348, 350
 Ordinal reliability, 331, 338, 341
 Ordinal response, 415, 418, 421
 Ordinal variables, 126
 Organizing principles, 89–106
 Overreporting, 346, 348, 363
 Overreporting index, 354, 356, 359, 360, 362
- P**
- Paired comparisons, 19, 21
 Pairwise deletion, 144–146, 148, 149, 157, 158
 Parafac model, 155
 Peeling, 154
 Percentile method, 154, 155
 Personality, 5–8, 22, 32, 35, 74, 75, 77, 91, 92, 94, 96, 97, 130, 141, 422, 423, 429, 430
 Person-fit index, 349, 354, 355, 363
 Personnel selection, 49, 196
 Phi-gamma function, 18, 21
 Plausible values, 8, 219, 220, 223
 Polychoric variance, 126
 Posterior distribution, 74, 219, 220, 223–229, 232, 234, 242, 334, 424, 430

- Prediction, v, vii, 18, 42, 44, 48–55, 57–59, 62, 63, 79, 82–85, 263, 300, 369–386, 425, 430
- Preprocessing, 81, 149
- Pretest-posttest design, 119, 122, 127
- Principal component analysis (PCA), vii, 141–158
- Probit analysis, 18
- Professional standards, 60–61
- Proportion, vii, 7–9, 15, 18, 23, 82, 112, 115–117, 157, 183, 184, 201, 202, 210, 225, 231, 232, 237, 277, 278, 289, 292–294, 296, 353, 383, 408, 419, 428
- Proportional reasoning assessment, 390, 407, 408
- Psychological testing, 48, 54, 55, 57, 59, 60, 62
- Psychometria, 9–12
- Psychometric function, 18–22, 24
- Psychometric Society, v, 20, 48, 211
- Psychophysics, 6, 12, 13, 15, 17, 18, 20, 21, 24
- Q**
- Q-methodology, 8
- Quantity objection, 24
- Quasi-IRF, 332–334
- R**
- Rank correlation, 330, 331, 336
- Rasch model, v, 72, 91, 92, 96, 102, 222, 224, 226, 242, 247, 248, 255–265, 267, 291, 330, 332
- Rater effects, 348–350, 356, 362–364
- Rating, 6–8, 22, 31, 39, 241, 351–353, 357, 364
- Reaction-time, 4
- Reference composite, 192
- Regularized PCA, 144–146, 148
- Rejection sampler, 221–222, 224–226, 230–232
- Reliability, 5, 34, 48, 71, 90, 111, 120, 164, 181, 196, 286, 331, 347
- Reliable change index (RCI), 208–212
- Repetitive responding, 349, 354, 355, 360, 362, 363
- Response bias, 345, 348–350, 360
- Response inconsistency, 360
- Response shift, 38, 120, 121, 124, 126–127, 129, 138
- Response time modeling/models, 74, 80, 221, 307–323
- Response times, v, vii, 33, 72–74, 80, 221, 236, 237, 240, 307–323
- Rest score, 334, 338
- Retrofitting, 390, 392, 410
- Robustness of weights, 54, 56
- Root Mean Square Error of Approximation (RMSEA), 135–138
- Rotation
- matrix, 151, 152
 - oblique, 143, 152
 - varimax, 143, 147, 153, 156
- Rotational freedom, 151
- S**
- Scalability coefficients, 272, 382
- Scale value, scientometrics, 15–17, 19, 21, 328, 330
- Scalogram, 329, 332
- Scholastic Aptitude Test (SAT), 164, 171, 172, 391
- School attitude questionnaire, 370
- Science-practitioner gap, 50
- Score transformations, 163–179
- Self-narratives, 72–74
- Sentiment analysis, 78
- Shapes, 156, 329, 337
- Signal, 16, 75, 81, 186, 187, 191
- Simple structure, 191, 312–314, 430
- Single variable exchange (SVE), 219, 222–223, 229
- Singular value, 142, 143, 145, 152, 155, 157
- Singular value decomposition, 142, 143, 145, 152
- Smoothing, 145, 172, 175, 201, 329, 341, 342
- Sparse latent class models (SLCM), vii, 413–439
- Split-half reliability, 183
- Standard error of measurement (SEM), 71, 164, 175, 196–200, 202, 204, 208, 211, 212
- Standard error of the latent trait, 315
- Standards for psychological and educational testing, 60–61
- Stanines, 374, 383
- Statistical prediction, 49–55, 57, 58, 63
- Statistical rules, 53, 58, 62
- Stochastic ordering, 253, 255, 256, 410
- Stochastic ordering on the latent variable by the sum, 410
- Stopping rules, 372, 374, 380, 382–385
- Strongly positive orthant dependence (SPOD), 272–276
- Structural relation, 96

Sufficient statistics, 221, 227, 229, 240, 251, 255, 256, 333
 Sum score, 38, 39, 120–123, 131, 136, 166, 196, 197, 200–202, 204–206, 253, 256, 258, 267, 286, 292, 293, 296, 300, 301, 333, 356, 380, 410

T

Target testing, 166
 Taylor series, 167–169, 173, 175
 Termination criterion, 372
 Test set, 82, 83
 Test standards, 50, 59–63
 Test theory, v, vi, vii, 16, 22, 32, 48, 71, 84, 120–126, 131–133, 136–138, 165–166, 177, 179, 182, 196, 251, 252, 393
 Test use, 42, 48–51, 54, 55, 59–63
 Textbooks on psychological testing, 48, 54, 55, 57, 62
 Theory of social representation, 50–51, 59
 Three-mode analysis, 155–157
 Three-parameter logistic (3PL) model, 224, 391, 403, 405–409
 3WayPack, 156
 Thurstone scale, 21, 22
 Total scores, vii, 48, 49, 111–117, 182, 186, 187, 191, 206, 213, 251–268, 292, 335, 370, 374, 376–380, 382–386, 430
 Training set, 82, 382
 Transparency, 54, 56, 252, 267, 285
 Transparency of decision making, 54
 True reliability, 132, 190, 191
 True score, 39, 71, 102, 124, 125, 164–170, 173, 177, 179, 182, 185, 186, 190, 198, 199, 203, 204, 206, 252, 258, 262, 264, 265, 267, 286, 287, 289, 294
 True score equating, 164, 258–265, 267, 287
 Trustworthy artificial intelligence, vi, 69–85
 Tucker-Lewis Index (TLI), 135–137

Tucker2 model, 155
 Tucker3 model, 155
 Two-parameter logistic (2PL) model, 203, 224, 244, 254–257, 264, 265, 327–332, 334–338, 341, 371, 382, 391–393, 403, 405–409
 Two-way mixed ANOVA, 197

U

Underreporting, 348, 354, 356, 359, 363
 Underreporting index, 354, 356, 359
 Unidimensional, vii, 20, 90–92, 102, 105, 120, 124, 125, 127, 183–185, 187, 191, 192, 203, 242, 263, 273, 355, 364, 370, 389–410
 Unidimensionality, 72, 90, 101–102, 120, 123, 126, 131, 253, 266, 267, 290, 327, 382, 383
 Unidimensional models, 90, 92
 Uniform relative difficulty, 331
 Unsystematic error, 348–350, 363
 Unweighted total score, vii, 252–255, 257, 264, 267, 268

V

Validity, vii, 5, 40, 42, 48, 49, 60, 61, 71, 80, 82–84, 90, 93–95, 104, 112–116, 197, 266, 267, 279, 285, 311, 346–364, 410, 430
 Validity index, 354–359, 363, 364
 Validity-index approach, 349–351, 356, 360–363
 Variables being associated, vii, 274–275
 Variations, 4, 5, 9, 10, 16, 37, 42, 71, 76, 81, 93, 125, 151, 152, 175, 237, 254, 256, 265–267, 284, 316, 347, 350

W

Weber's Law, 13–15, 18, 21
 Working memory, 4, 93