

Springer Proceedings in Mathematics & Statistics

Marie Wiberg · Dylan Molenaar
Jorge González · Ulf Böckenholt
Jee-Seon Kim *Editors*

Quantitative Psychology

84th Annual Meeting
of the Psychometric Society, Santiago,
Chile, 2019

MOREMEDIA



Springer

**Springer Proceedings in Mathematics &
Statistics**

Volume 322

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at <http://www.springer.com/series/10533>

Marie Wiberg • Dylan Molenaar • Jorge González
Ulf Böckenholt • Jee-Seon Kim
Editors

Quantitative Psychology

84th Annual Meeting of the Psychometric
Society, Santiago, Chile, 2019

 Springer

Editors

Marie Wiberg
Department of Statistics, USBE
Umeå University
Umeå, Sweden

Dylan Molenaar
Department of Psychology
University of Amsterdam
Noord-Holland
Amsterdam, The Netherlands

Jorge González
Facultad de Matemáticas
Pontificia Universidad Católica de Chile
Santiago, Chile

Ulf Böckenholt
Kellogg School of Management
Northwestern University
Evanston, IL, USA

Jee-Seon Kim
Department of Educational Psychology
University of Wisconsin–Madison
Madison, WI, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-43468-7 ISBN 978-3-030-43469-4 (eBook)
<https://doi.org/10.1007/978-3-030-43469-4>

Mathematics Subject Classification: 62P15, 62P25, 91E45

© Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved 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, express 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: Gewerbstrasse 11, 6330 Cham, Switzerland

Preface

This volume represents presentations given at the 84th Annual Meeting of the Psychometric Society, organized by Centro de Extensión at the Pontificia Universidad Católica de Chile, in Santiago, Chile, on July 15–19, 2019. The meeting attracted 411 participants, and 383 papers were presented, of which 84 were part of a symposium. There were 4 preconference workshops, 11 keynote presentations, 8 invited presentations, 2 career-ward presentations, 4 state-of-the-art presentations, 66 poster presentations, 1 dissertation award winner, and 19 symposia.

Since the 77th meeting in Lincoln, Nebraska, Springer has published the proceedings volume from the annual meeting of the Psychometric Society to allow presenters to spread their ideas quickly to the wider research community while still undergoing a thorough review process. The previous seven volumes of the meetings in Lincoln, Arnhem, Madison, Beijing, Asheville, Zurich, and New York were enthusiastically received, and we expect these proceedings to be successful as well.

The authors of these proceedings were asked to use their presentations at the meeting as the bases of their chapters, possibly extended with new ideas or additional information. The result is a selection of 28 state-of-the-art chapters addressing a diverse set of psychometric topics, including but not limited to item response theory, factor analysis, hierarchical models, and computerized adaptive testing.

Umeå, Sweden
Amsterdam, Noord-Holland, The Netherlands
Santiago, Chile
Evanston, IL, USA
Madison, WI, USA

Marie Wiberg
Dylan Molenaar
Jorge González
Ulf Böckenholt
Jee-Seon Kim

Contents

Stories of Successful Careers in Psychometrics and What We Can Learn from Them	1
Carolyn J. Anderson, Susan Embretson, Jacqueline Meulman, Irina Moustaki, Alina A. von Davier, Marie Wiberg, and Duanli Yan	
Developing a Concept Map for Rasch Measurement Theory	19
George Engelhard Jr and Jue Wang	
Person Parameter Estimation for IRT Models of Forced-Choice Data: Merits and Perils of Pseudo-Likelihood Approaches	31
Safir Yousofi	
An Extended Item Response Tree Model for Wording Effects in Mixed-Format Scales	45
Yi-Jhen Wu and Kuan-Yu Jin	
The Four-Parameter Normal Ogive Model with Response Times	55
Yang Du and Justin L. Kern	
A Bayesian Graphical and Probabilistic Proposal for Bias Analysis	69
Claudia Ovalle and Danilo Alvares	
Comparing Hyperprior Distributions to Estimate Variance Components for Interrater Reliability Coefficients	79
Debby ten Hove, Terrence D. Jorgensen, and L. Andries van der Ark	
A Hierarchical Joint Model for Bounded Response Time and Response Accuracy	95
Sandra Flores, Jorge Luis Bazán, and Heleno Bolfarine	
Selecting a Presmoothing Model in Kernel Equating	111
Gabriel Wallin and Marie Wiberg	
Practical Implementation of Test Equating Using R	121
Marie Wiberg and Jorge González	

Predictive Validity Under Partial Observability	135
Eduardo Alarcón-Bustamante, Ernesto San Martín and Jorge González	
Multiple-Group Propensity Score Inverse Weight Trimming and Its Impact on Covariate Balance and Bias in Treatment Effect Estimation ...	147
Diego Luna-Bazaldua and Laura O'Dwyer	
Procrustes Penalty Function for Matching Matrices to Targets with Its Applications	161
Naoto Yamashita	
Factor Score Estimation from the Perspective of Item Response Theory ..	171
David Thissen and Anne Thissen-Roe	
On the Precision Matrix in Semi-High-Dimensional Settings	185
Kentaro Hayashi, Ke-Hai Yuan, and Ge Jiang	
Performance of the Modified Continuous α-Stratification Indices in Computerized Adaptive Testing	201
Ya-Hui Su and Yan-Ling Lai	
Constant CSEM Achieved Through Scale Transformation and Adaptive Testing	213
Dongmei Li	
Synergized Bootstrapping: The Whole is Faster than the Sum of Its Parts	227
Tim Loossens, Stijn Verdonck, and Francis Tuerlinckx	
Synchronized Time Profile Similarity in Applications to Nearest Neighbor Classification	247
Qimin Liu	
Topic Modeling of Constructed-Response Answers on Social Study Assessments	263
Jiawei Xiong, Hye-Jeong Choi, Seohyun Kim, Minh Kwak, and Allan S. Cohen	
Impact of Measurement Bias on Screening Measures	275
Oscar Gonzalez, William E. Pelham III, and A. R. Georgeson	
Reliability and Structure Validity of a Teacher Pedagogical Competencies Scale: A Case Study from Chile	285
Juan I. Venegas-Muggli	
Psychoperiscope	299
Joshua Chiroma Gandi	
Modeling Household Food Insecurity with a Polytomous Rasch Model ...	319
Victoria T. Tanaka, George Engelhard Jr, and Matthew P. Rabbitt	

Classical Perspectives of Controlling Acquiescence with Balanced Scales. 333
Ricardo Primi, Nelson Hauck-Filho, Felipe Valentini, and Daniel Santos

Testing Heterogeneity in Inter-Rater Reliability 347
František Bartoš, Patrícia Martinková, and Marek Brabec

An Application of Regularized Extended Redundancy Analysis via Generalized Estimating Equations to the Study of Co-occurring Substance Use Among US Adults 365
Sunmee Kim, Sungyoung Lee, Ramsey L. Cardwell, Yongkang Kim, Taesung Park, and Heungsun Hwang

Permutation Test of Regression Coefficients in Social Network Data Analysis 377
Wen Qu, Haiyan Liu, and Zhiyong Zhang

Index..... 389

Stories of Successful Careers in Psychometrics and What We Can Learn from Them



Carolyn J. Anderson, Susan Embretson, Jacqueline Meulman, Irini Moustaki,
Alina A. von Davier, Marie Wiberg , and Duanli Yan

Abstract This paper was inspired by the presentations and discussions from the panel “Successful Careers in Academia and Industry and What We Can Learn from Them” that took place at the IMPS meeting in 2019. In this paper, we discuss what makes a career successful in academia and industry and we provide examples from the past to the present. We include education and career paths as well as highlights of achievements as researchers and teachers. The paper provides a brief historical context for the representation of women in psychometrics and an insight into strategies for success for publishing, for grant applications and promotion. The authors outline the importance of interdisciplinary work, the inclusive citation

C. J. Anderson (✉)
University of Illinois, Champaign, IL, USA
e-mail: cja@illinois.edu

S. Embretson
Georgia Institute of Technology, Atlanta, GA, USA
e-mail: susan.embretson@psych.gatech.edu

J. Meulman
Leiden University, Leiden, The Netherlands
Stanford University, Stanford, CA, USA
e-mail: jmeulman@math.leidenuniv.nl; jmeulman@stanford.edu

I. Moustaki
London School of Economics & Political Science, London, UK
e-mail: i.moustaki@lse.ac.uk

A. A. von Davier
ACTNext, Iowa City, IA, USA
e-mail: Alina.vonDavier@act.org

M. Wiberg
Department of Statistics, USBE, Umeå University, Umeå, Sweden
e-mail: marie.wiberg@umu.se

D. Yan
Educational Testing Service, Princeton, NJ, USA
e-mail: dyan@ets.org

approaches, and visibility of research in academia and industry. The personal stories provide a platform for considering the needs for a supportive work environment for women and for work-life balance. The outcome of these discussions and reflections of the panel members are included in the paper.

Keywords Advice · Career paths · Psychometrics history · Gender gap

1 Introduction

In recent years, society has started to shift its narrative about scientists from the lonely genius (usually a white man) to more diverse images of the researchers, authors of papers, and to their supportive environment. The IMPS19 session, “Stories of Successful Careers in Psychometrics and What We Can Learn from Them,” is part of this expansion of acknowledgment of the contributions of contemporary fellow scientists to the field of psychometrics and their individual paths to successful careers. This proceedings volume provides a snapshot of the interests of members of the Psychometric Society in 2019 and as such it encompasses a historical and social perspective on ideas, creators, and life stories that are being mingled with the psychometric papers that these authors published in this volume or elsewhere.

In this paper, we loosely follow the structure of the symposium and allow the contributors to speak to her professional successes and to the personal context in which these successes took shape. The professional successes include breakthrough research ideas and projects, leadership acknowledgment, and social impact. The scientists will also share their lessons learned for the next generations of psychometricians. The team of established scientists is comprised of seven women from six countries, who now live and work across four countries. Some of these stories speak to the geopolitical influence, the immigrant’s experience, the struggle to publish in a foreign language, and the struggle to be authentic in a professional world with relatively narrow expectations.

There are many socio-historical, political, and cultural conditions that have led to marginalization of women in technical domains. STEM subjects in some societies are highly gendered often based on a belief that boys are better at math than girls due to biological differences. In the USA, women earn fewer PhDs in STEM domains and only 31.5% of women earned PhDs in mathematics and computer (Okahana and Zhou 2017). School and parental guidance have also contributed to the gender gap in STEM. Girls and boys often grow up with the idea that they will be bad and good at math, respectively (e.g., Math class is tough! Barbie is for girls) and that girls do not belong in a technical environment. All those reasons are in addition to systemic and structural biases such as opportunities for training, and later on, for recruiting.

An article published in *The Guardian* by Carol Black and Asiya Islam in 2014 is a response to over 50 senior Cambridge academics called on the university to change its staff appointment procedure because the existing system favored men. They stated that “Despite accounting for 45% of the academic workforce, women

hold only 20% of professorships in UK universities, and just 15.3% of such posts in Cambridge” (Black and Islam 2014). Though more women enter university than men and there is an almost equal representation of women and men at lower professional levels, only 27.5% of senior managers in higher education and 20.5% of professors in the UK are women. Worse, only 1.1% of senior managers in higher education and 1.4% of professors in the UK are black and minority ethnic women.

One would expect that in more gender-equal societies the gender gap in STEM scores and in higher managerial or academic positions is smaller. Different systems for tenure and promotion also lead to different outcomes. It looks like the problem is universal. This imbalance is spread in the world and the causes are often blatantly attributed to narrow views of women’s roles in society: *The Guardian* reported in June 2019, that after a medical school in Japan admitted rigging admission procedures to give men an unfair advantage, once the system became fair, women have outperformed their male counterparts in entrance examinations (McCurry 2019).

2 History

This section provides a brief historical tribute to women who have contributed to psychometrics and related disciplines. It is by no means complete or exhaustive. Psychometrics was founded by Thurstone’s vision for a mathematical underpinning for psychological research. The Psychometric Society was founded in 1935 by Louis Thurstone, Jack Dunlap, Paul Horst, Albert Kurtz, Marion Richardson, and John Stalnaker. Paul Horst and Albert Kurtz founded the journal in 1936 with the mission to create a journal that will be mathematically oriented to develop and disseminate work in psychological measurement. Much before that, Gauss in 1809 presented the theory of errors of observation following the normal distribution, Bessel’s presented “a personal equation” to correct observations for differences among observers, Galton in 1884 designed an apparatus to measure a variety of bodily dimensions, Cattell in 1889 established a laboratory of psychology with an interest in psychometric measures.

Where are the women in all those initiatives and contributions? At a time when women were destined to get married and bear children, Florence Nightingale (1820–1910), who was self-educated in statistics, pioneered in visual statistical graphs called Nightingale Rose Diagram or Polar Area Diagram. She was the first female member of the Royal Statistical Society and the founder of the nursing profession. Florence Nightingale David (1909–1993), named after Florence Nightingale, studied mathematics at Bedford College for Women after failing to go to University College of London. She published the *Tables of the Correlation Coefficient*, as well as *Combinatorial Chance* (with D.E. Barton) and *Games, Gods and Gambling: The Origins and the History of Probability*. She chaired the statistics department at the University of California, Berkeley, then founded the statistics department at the University of California, Riverside.

Ethel Elderton (1878–1954) is a true hidden figure, a female researcher who worked with Galton and Pearson in eugenics research. In 1905 she resigned her teaching post to become Galton’s assistant. Subsequently, she became a Galton Scholar and Fellow and Assistant Professor at University College London. In the same period, Gertrude Mary Cox (1900–1978) dreamed to be a missionary and saving souls in far-off lands. To be qualified as a missionary, she became a student of George Snedecor then published *Experimental Design* (Cochran and Cox 1957). She was the first female department chair in a men’s world and started the well-known North Carolina “Research Triangle.” In psychometrics, Thelma Thurstone (1897–1993) a psychometrician herself combined the theory of intelligence with its measurement to design instructional materials, like the tests she developed for the American Council on Education from 1924 to 1948. In 1955, Thelma Thurstone was asked to assume the directorship of the Psychometric Laboratory upon the death of her husband in order to continue his funded research projects. Barbara Stoddard Burks (1902–1943) worked in behavioral genetics and intelligence and was the first one who used a graph to represent a mediator. Her first paper published in 1926 was on the inadequacy of the partial and multiple correlation technique. Anne Anastasi (1908–2001) is known as the “test guru” psychometrician and the psychology’s female voice. She pioneered the development of psychometrics and chaired the department of psychology at the male-dominated school at Fordham University, and she won many awards including The American Psychological Foundation’s Gold Medal for Life Achievement. Her books on *Differential Psychology*, *Fields of Applied Psychology*, and [Psychological Testing](#) (with 7 editions) influenced generations of psychometricians. Fordham University established a special position named Anne Anastasi Chair Professor.

Another important contributor is Dorothy Adkins (1912–1975), an American psychologist who was interested in new (at the time) statistical techniques of factor analysis. She applied factor analytic techniques in order to examine and better understand curriculum, program evaluation, and affect in children. She was also co-editor of *Psychometrika* with Paul Horst (1958–1959, 1963–1966) and president of the Psychometric Society in 1949–1950. Forty-five years later Fumiko Samejima became the next female president of the society (1996–1997) who is known for her work on Item Response Theory (IRT) models for polytomous data. A few more women followed as presidents of the society, Susan Embretson (1998–1999), Jacqueline Meulman (2002–2003), Sophia Rabe-Hesketh (2014–2015), and Irini Moustaki (president-elect, 2020–2021). Susan Embretson was the first to integrate cognitive theory into IRT and test design whereas Jacqueline Meulman made significant contributions in the area of multivariate data analysis with optimal transformations of variables, and multidimensional scaling.

Many women without a PhD also made significant contributions. At Harold Gulliksen’s Gold Medal Award for Lifetime Achievement in Psychological Science (1991), he acknowledged his wife as his significant collaborator who did the programming and analyses for him. Similarly, Marilyn Wingersky worked mostly with Fred Lord and implemented algorithm, statistical models, and developed the LOGIST software for estimating latent traits and item parameters. Martha Stocking,

without a doctorate, also worked with Fred then furthered her contributions on computerized adaptive testing (CAT) research and development including automated test assembly (ATA) using weighted deviation and the conditional item exposure control algorithm with Charlie Lewis. Kikumi Tatsuoka (1930–2016) received her PhD later in life, after raising her children; she developed the Rule-Space model for diagnostic assessment. Dorothy Thayer has been an instrumental behind the scene figure. She worked with Mel Novick, Don Rubin, Paul Holland, Rebecca Zwick, Charlie Lewis, and Alina von Davier, and published numerous numbers of papers with them, always as the second author. Among other researchers we would like to note is Frances Swineford (1909–1997) who in 1937 together with Holzinger introduced the bifactor model (one general factor and multiple group factors) for mental abilities (Holzinger and Swineford 1937). Again, this oversight has been characteristic of the scientific world in the twentieth century. As discussed in Yong (2019) and Huerta-Sanchez and Rolfs (2019), our colleague professor Margaret Wu from Melbourne has been only thanked for an algorithm that she co-created to compute the “Watterson estimator.”

Finally, a very important initiative in 2004 is the Psychology Feminist Voices project directed by Alexandra Rutherford at York University in Toronto, Canada which aims to collect, preserve, and share the narratives of diverse feminist psychologists from all over the world (see <http://www.feministvoices.com/about>).

3 The Impact of the Structure of Society and Academia

3.1 Societal Structures

The structure of a society is important when pursuing an academic career. To have well-organized paid maternal and paternal leave tend to enhance gender equality. In the past, more men than women earned PhDs, but now in many countries many universities and colleges have more women than men earning PhD degrees (Okahana and Zhou 2017). The balance between work and family has gained attention with both men and women working. Mason and Wolfinger (2013) have examined the relationship between family formation and academic careers of men and women, including an examination of the family sacrifices women often have to make to get ahead in academia and consider how gender and family interact to affect promotion to full professor, salaries, and retirement. Although their research is from the USA it is seen in many countries that even if women and men work a similar number of hours, women tend to take more responsibility for their family. They concluded that men can get a career advantage when having children but for women it can be a career killer. Those women who advance through the faculty ranks tend to pay a high price by being less likely to be married with children. For a woman to facilitate her career it is thus important to be in a relationship which

believes in equality and to work in a country where the society helps women and men with this equality by, for example, paid maternal and paternal leave.

3.2 *Academic Structures*

It is not just the structure of the society which is important but also the academic structure. To have open calls and transparency in the career system is typically viewed as a way to frame gender equality. van den Brink et al. (2010) examined transparency in the Netherlands and concluded that transparency and accountability should be deployed to their full potential. In their study, transparency was limited to recruitment protocols, but transparency should also imply making the process and decisions more visible for the larger academic society, which is the case in Sweden and Finland.

Internal structures are also important. To be part of a supportive work environment, and to have role models, mentors, and colleagues all greatly enhance the chances of being able to pursue an academic career. Receiving constructive feedback is essential for career development for everyone. However, when Rubini and Menegatti (2014) examined the language in academia, they concluded that judgments of female applicants in academic personnel selection were formulated using negative terms at a more abstract level and positive terms at a more concrete level than those of male applicants. They also found that linguistic discrimination was perpetrated only by male committee members. The discrimination was mainly based on the use of negative adjectives and thus this could be a hindrance for women's academic careers. To counteract this tendency, institutions often try to have men and women represented on different committees; however, women should make sure not to get stuck doing committee work because they need a woman. It is important to say yes to exciting new projects and collaborations and often say no to the role of "female representative" unless you feel they asked you due to your competence. In summary, choose your service and work wisely.

4 Personal Reflections

In this section, each of the panel members has sketched a short biography together with some personal reflections.

4.1 *Personal Reflection by Carolyn J. Anderson*

Themes throughout CJA's career have included accepting opportunities that were offered to her and following her interests. Curiosity has been a driving force in her

career. In college, CJA was introduced to quantitative psychology by Bill Meredith and Barb Mellers at the University of California at Berkeley and took Bill's graduate seminars on factor analysis and latent class analysis. CJA was hooked!

CJA's first major challenge was choosing a dissertation topic upon which she built a career in academia. The University of Illinois at Urbana-Champaign (UIUC) was an ideal environment to pursue a PhD due to the breadth and depth of expertise of the faculty. Before CJA's ideas solidified, she did research on judgment and decision making with Michael Birnbaum and Elke Weber, and social network analysis with Stanley Wasserman. Stanley agreed to be her advisor and allowed her the freedom and support to pursue and explore whatever interested CJA. Starting with two papers by Leo Goodman that Stanley recommended, CJA read backward, forward, and side-ways in literatures on categorical data analysis, matrix decompositions, graphical models, optimal scaling, and computing algorithms. CJA's dissertation encompassed all of these areas and earned her the Psychometric Society and APA Division 5 Dissertation awards.

Dual career couples can face many challenges, especially finding positions in the same city and having a family. CJA was offered a tenure track position at UIUC and accepted it because she was expecting her first child and both parents would be employed. The policies at UIUC were nonexistent regarding childbirth and family policy. When CJA began, 80% of the tenured faculty in her primary college were men and attitudes of some senior faculty were not supportive of women. For example, after being denied a release from teaching due to childbirth, she was asked "doesn't it bother you that someone else is raising your child?". Fortunately, she also had very supportive colleagues. Stanley Wasserman and Rod McDonald stepped up and taught her courses until she was able to return to work.

After a rocky start, 15 months of little to no sleep, and becoming visually disabled, she needed to jump start her research program. She went back to the literature, including original sources. Typographic errors in a paper had carried through the literature and after correcting them it became obvious that row-column association models and their extensions were standard item response models. This led to an NSF grant and papers on graphical models and latent variables models starting with Anderson and Vermunt (2000).

4.2 Personal Reflection by Susan Embretson

SE's research direction has focused on understanding the cognitive processes, skills, and strategies that are involved in responding to test items. Her research has included developing item response theory models, perspectives on the validity concept, examining the impact of item design on test correlates and developing automatic item generators. SE has received career awards for this research, the 2019 Career Award for Lifetime Achievement from the Psychometric Society, the 2018 Saul Sells Award for Distinguished Multivariate Research from the Society for Multivariate Experimental Psychology, the 2013 Career Contribution Award from

the National Council on Measurement in Education, and the 2011 Distinguished Lifetime Achievement Award from the American Educational Research Association: Assessment and Cognition Division, as well as several scientific contribution awards. Although her personal journey to her research program was not direct, her interests in the topic began in high school after taking the Preliminary Scholastic Aptitude Test. Unfortunately, her score was not high enough to qualify for a Merit Scholarship. Why were test items involving Victorian novel vocabulary on a college admissions test? She complained to her high school counselor and a few weeks later, two individuals from the University of Minnesota came to administer an individual intelligence test. Shortly afterwards she received a full scholarship. Al Johnson, an engineer who built skyscrapers, decided he could fund ten students per year. He probably did not read Victorian novels either.

SE began her studies with a goal to major in psychology. However, the required research experiences in the introductory course, which included running rats in mazes and learning nonsense syllables, did not pique her interest. She changed her major to Spanish, but after learning to speak the language, she found that she was not as enthusiastic about the literature and could not envision being a high school language teacher. By this time, she had a young daughter and a husband, which involved 2 h commuting as they could afford only one car. One very cold winter day, she decided to drop out of school. She hoped that the world of business would suit her better. It did not. After 6 months of the world of work, she decided to take two night school classes: Individual Differences and Psychological Statistics. Wow! SE found her interests. She returned full time to the University of Minnesota and, fortunately, the Al Johnson Foundation decided that they could fund her again and she finished in a little over 1 year.

SE applied for graduate school at the University of Minnesota. Required was a test used to select students for fellowships, the Miller Analogy Test. She remembers the test well. Why is knowing the answer to analogies such as “Moscow: Vodka:: Copenhagen:?” measuring aptitude? She did not know what the Danes drank. Again her score was not high. Despite that, she was selected, primarily because her Bachelor of Arts degree was awarded *summa cum laude*.

Her graduate career was exciting, as IRT was just entering the field and she was able to pursue her research interests in cognition and measurement. She delayed finishing by 1 year and then took a post doc position for 1 year so that her husband could finish his PhD. Afterwards she interviewed at the University of Georgia. However, the available teaching topic was not her major interest and the work-family balance did not work out. Thus with difficulty, she turned down their offer even though nothing else was pending. This was a good decision, as good luck came in a couple of weeks! The University of Kansas offered her a position to ease into teaching graduate statistics, and she could pursue whatever research topic interested her so long as it was successful. Also, women’s expertise in quantitative methods was not questioned, since Julie Shafer had been teaching statistics there. SE accepted the offer, to which she attributed much of her success. She spent 30 years

there and pursued her research interests with enthusiasm. Her current position at the Georgia Institute of Technology has been successful due to the solid base of research and teaching that she built at KU. In summary, SE characterizes her personal journey as involving some good luck, some good decisions, and lots of persistence.

4.3 Personal Reflection by Jacqueline Meulman

After JM was drawn into psychometrics while studying its history, preparing an undergraduate course in History of Psychology as TA at Leiden University, she abandoned everything else by becoming an RA at the Leiden Department of Data Theory in 1978. This department was founded at Leiden University in 1970 by the late John P. van de Geer, and its mission was the development of new and innovative methods for statistical multidimensional data analysis. Later on, Jan de Leeuw added to its mission the implementation in software for multivariate analysis of categorical data, and for multidimensional scaling and unfolding. JM had found the topic in statistics that she would cherish for the next 40+ years to come.

Like Jan de Leeuw and Willem Heiser, JM visited the famous AT&T Bell Telephone Laboratories in Murray Hill, NJ. The year 1982 that she spent in Doug Carroll's group in Mike Wish's department *Computer-Aided Information Systems* changed her life. Doug was a superb mentor who introduced her to all her heroes in psychometrics and beyond. It was Paul Tukey, the nephew of Bell Labs' Associate Executive Director John Tukey, who told her she was not a psychometrician, but a statistician. After returning to the Department of Data Theory in 1983, JM finished her dissertation in 1986 (advisors Jan de Leeuw and John P. van de Geer), and was awarded a 5-year fellowship from the Royal Netherlands Academy of Arts and Sciences, which allowed her to continue her career at the department that she loved.

In 1987, John P. van de Geer retired, and Jan de Leeuw took a position at UCLA, and Willem Heiser and JM were left some big shoes to fill. Their efforts resulted in Albert Gifi's *Nonlinear Multivariate Analysis* published by Wiley in 1990, and the incorporation of the associated software programs in the SPSS package CATEGORIES (also from 1990 onwards).

A next important period in JM's career started in 1992, by visiting the University of Illinois at Urbana-Champaign, where she was teaching and started collaborating with Larry Hubert. In 1994, JM was awarded the prestigious PIONEER Award by the Netherlands Organization for Scientific Research (NWO), which allowed her to start her own research group in Leiden, as well as spending time in Champaign-Urbana, where she had been appointed as Adjunct Professor in 1993. The collaboration with Larry Hubert and Phipps Arabie (in the so-called HAM team) resulted in a number of papers and two books.

In the meantime, Willem van Zwet, who was Professor of Mathematical Statistics in the Mathematical Institute in Leiden, took it upon him to support JM to become full professor. Her Chair was called Applied Data Theory, and she was leading a group of assistant professors, postdocs, and PhD students; the group was still called

Data Theory, but was relocated at the Department of Education. This association did not develop into a good synergy, and after a number of difficult years, the Data Theory Group left the Department of Education. However, good things also happened in this period: JM was elected as President of the Psychometric Society (in 2001), and as Member of the Royal Netherlands Academy of Arts and Sciences (in 2002).

In 2006, JM was offered a position in statistics at the Leiden Mathematical Institute for one day a week, and this appointment was extended to a full-time position with a Chair in Applied Statistics in 2009. In the meantime, the collaboration with SPSS had resulted in many new software programs, and royalties for Leiden University (first shared with Willem Heiser, and later under full control of JM) that increased to very impressive figures. The latter made it possible for JM to start anew within the Mathematical Institute (MI), with appointing assistant professors and a group of PhD students. At the MI, JM developed with Richard Gill, and later Aad van der Vaart, a new Master program called *Statistical Science (for the Life and Behavioral Sciences)*, in collaboration with other statisticians from the Leiden University Medical Center, the Methodology & Statistics Division at the Leiden Institute of Psychology, and Wageningen University and Research Center. From 2011 to 2016, JM was President of the *Netherlands Society of Statistics and Operations Research*, and she was appointed in the Department of Statistics at Stanford University, first in 2009 as Visiting, and later in 2017 as Adjunct Professor. The above story may sound as a dream, but the path has known many large obstacles, professional as well as medical. JM had to work very hard to pursue her ideals. But all is well that ends well: JM was honored with the Psychometric Society's Career Award for Life Time Achievement 2020.

4.4 Personal Reflection by Irini Moustaki

IM studied Statistics and Computer Science at the Athens University of Economics and Business and continued her studies at the London School of Economics from where she received a masters and PhD in Statistics. Initially, her PhD thesis was on sample surveys and variance estimators under the supervision of Colm O'Muircheartaigh but as soon as Colm was awarded a state grant as a co-investigator with David Bartholomew and Martin Knott on the Analysis of Large and Complex Data Sets, IM started working on latent variable models for mixed data closely also with Knott and Bartholomew. At LSE she has been very fortunate to have had a very supportive and encouraging environment in which to study and later to work. A year before she received her PhD, she got an appointment as a temporary lecturer at LSE and a year later a tenure track position in the same department. The Statistics Department at the time had no female professors and only one female lecturer. IM also spent a period of 5 years at the Athens University of Economics and Business as Assistant and Associate Professor before returning to LSE again in 2007 as associate

professor and in 2013 became a full professor. IM served both as head and deputy head in her department at LSE.

A turning point in her PhD studies was when she attended the IOPS meeting in Tilburg as a PhD student to discover to her surprise a whole community of researchers working on models with latent variables. At LSE and in the UK in general there wasn't much of a psychometric tradition or use of latent variable modeling in social sciences. The second opportunity came when her supervisor encouraged her to attend a workshop by Karl Joreskog on SEM in Heidelberg. This is also the place when she met with Alina and Matthias von Davier and also started a conversation with Karl Joreskog on IRT and SEM that later on led to two papers and a long-term friendship. The Psychometric meetings and community provided her with an academic family which allowed her to discuss her research developments, make collaborators, and make valuable friendships. IM is indebted to the continuous support she received in her early career by Martin Knott at the LSE, who trusted her capabilities and generously exchanged ideas of research and projects. Her collaborations with researchers from LSE but also other places in Europe and beyond led to publications in the areas of missing values, detection of outliers, and composite likelihood estimation. The highlights of her career were when she received an honorary doctorate from the University of Uppsala on the recommendation of her collaborators and friends Fan Wallentin and Karl Joreskog, served as the editor-in-chief of *Psychometrika*, and honored to be the president-elect of the Psychometric Society. The Psychometric Society has continuously provided a stimulated intellectual environment for her. Further to her teaching and research, IM finds the mentoring of junior academics and PhD students a very important part of her job.

4.5 Personal Reflection by Alina von Davier

AvD studied mathematics at the University of Bucharest and at the end of the studies was fortunate to experience the political change in a country that had been under an authoritarian regime for a long time. The political changes brought opportunities and hope and AvD went to work for a research institute (The Institute of Psychology of the Romanian Academy) instead of teaching math at a high school, as would have been the case under the previous system. Further on, she went to do her PhD in Germany. She started her work on falsifying causal hypotheses with Rolf Steyer, but she discovered interesting singularity points in the testing of hypotheses that captured her interest, and therefore her dissertation ended up back in mathematics, with a second advisor, Norbert Gaffke. In the 5 years she lived in Germany, she also learned German, married MvD, and had a son—efficiently, as she likes to describe it.

The von Daviers moved to the US and specifically to ETS, where their interests found a good match with the company's needs. ETS provided an incredible intellectually rich environment for the development and exploration of one's ideas. Her research journey went from research in test equating, to adaptive testing, and

to the measurement of collaborative problem solving and other complex constructs. She was fortunate to work closely with Paul Holland, Charlie Lewis, and Shelby Haberman. She also became increasingly involved with the operational testing and with the implementation of new methodologies and technologies. In 2015, she introduced the concept of Computational Psychometrics to define the blend of psychometric theory with the data-driven discovery. She moved to ACT in 2016 to establish and lead an innovation hub to help transform the company. With this move, a special opportunity was offered to her to redefine what the educational experience means in the twenty-first century and how psychometrics can be the foundation for the learning, measurement, and navigation efforts to support this experience for everyone everywhere.

4.6 Personal Reflection by Marie Wiberg

MW has in her career been driven by curiosity and she loves to try to solve new challenges and to collaborate with other curious persons. MW started her PhD in Statistics but worked at an educational measurement department where she came into contact with real test problems. From networking at conferences, she ended up as a visiting researcher with Professor Ramsay at McGill University and then moved on to do a postdoc with professor van der Linden at the University of Twente. These two research experiences had a major impact on her future career path. The work with nonparametric item response theory with Ramsay, which they both thought was an “easy” problem to solve, took more than 12 years to solve, but several papers and workshops have followed in recent years. An important lesson is that good ideas and how to solve them may take a while. The work in the Netherlands rerouted her to different test equating problems—a path she still follows and led to successful collaborations with researchers from around the world. Most of her collaborations spring from brief meetings at conferences where many new ideas have emerged. Since the start of her PhD program, MW has had an interest to work with real empirical test data (including national tests, admissions test, and the large-scale assessments TIMSS and PISA). MW recommends everyone who has a chance to work with real data to take the opportunity as many theoretical research problems may emerge. MWs work has been recognized nationally through large research grants and she has been a member of the Young Academy of Sweden which is an academy for talented young researchers within all research fields. Internationally, she has coauthored a test equating book (González and Wiberg 2017), worked as an associate editor for the Journal of Educational Measurement and is currently editor of the IMPS proceedings.

4.7 *Personal Reflection by Duanli Yan*

DY has been very fortunate to have many distinguished teachers and mentors who have had great influence through the decades on her career and life. DY became interested in statistics and optimization after earning her bachelor's degree in computer science and applications. When she completed her dual masters in statistics and in operations research in the statistics department at Penn State University, Professor C. R. Rao tried to persuade her to stay and do a PhD with him. However, she had been in school for almost all of her life by that time, and she wanted to work. Soon after she started working at ETS, she realized that she should have done a PhD.

While working at ETS, she learned the Rule-space model for cognitive diagnoses from Kikumi Tatusoka who came to ETS to join Charlie Lewis, her former dissertation advisor at the University of Illinois. DY learned many things from Bob Mislevy and they have been leading an annual NCME training session based on their book *Bayesian Networks in Educational Assessment* since 2002 (Almond et al. 2015). Charlie introduced DY to the world of CAT and they developed the tree-based CAT algorithm. She was always impressed by how Bob and Charlie solved problems. She was also impressed about 20 years ago, when Charlie hosted his former dissertation advisor John Tukey (from Princeton University in 1970) at ETS once a month to consult on their projects. DY brought modern computer outputs with analyses results and plots to show John. John didn't look at those outputs, instead he took a piece of paper and a pencil then started to draw a stem-leaf graph, and he asked everyone what they thought the results should be, which were the results DY produced after hours of computing! DY was astonished by how he explained things from his head, which is the way Charlie writes out the equations from his head at any point! She wanted to learn more. So, she later followed Charlie to Fordham University to finish her PhD in Psychometrics with her dissertation on computerized multistage testing (MST) which was co-advised by Charlie Lewis and Alina von Davier. They subsequently published a book (Yan et al. 2014) and DY received the 2016 AERA Significant Contribution to Educational Measurement and Research Methodology Award. DY was also honored to receive 2011 ETS Presidential Award, 2013 NCME Brenda Loyd Dissertations Award, and 2015 IACAT Early Career Award. Currently, she is responsible for ETS's automated scoring systems evaluations and analyses including a *Handbook of Automated Scoring: Theory into Practice* (Yan et al. 2020).

During her career, DY faced many challenges such as work and life balance including operational work versus research and development, schedule conflicts, family, and child raising. From her work on many operational programs and research and development projects, she gained experiences dealing with real-world practical issues and finding solutions. These helped her to create more innovative research questions and to develop and implement systems that increase accuracy and efficiency by using optimization and automation. Her daughter Victoria Song often slept on her desk or on the office floor. She grew up at ETS, volunteered

and interned at ETS, and is working on her dissertation advised by Fordham Anne Anastasi Chair Professor David Budescu. All DY's learnings and experiences are good lessons in her life. She appreciates her mentors who had great influence in her career and life.

5 Advice, Recommendations, and Lessons Learned

Although we all have worked at different academic departments and in different countries, we still have similar experiences and we have learned many things during our journeys. The some of the lessons all seven of us have learned are described below.

5.1 Things to Do

Dare to say yes to exciting projects and decline administrative committees if they just need a woman and not specifically your competence.

Try to find a supportive work environment with people you can be on the same level with. When you are young try to find a good mentor and once you are older try to be a good mentor: keep an open mind and learn. You never know when they become fruits in your life.

Work with people you enjoy spending time with and those you dare to say that you do not understand what they mean. It is more fun and it is more rewarding for both partners.

Don't be afraid to collaborate with new people. Some of the best collaborations come from just listening to a conference presentation and suggesting to collaborate with a joint topic, even if none of the people knew each other before.

Probably the most important lesson to share with junior psychometricians is to believe, respect, and acknowledge one's own ideas and at least in so much as to try them out. This would start by just writing down the idea, and then write the computer code, prove the theorem, and/or test the result empirically.

The "just do it!" approach is usually good. Even if an idea is not valuable but by trying them out one builds both expertise and confidence. It is tempting for a novice to talk about an idea but not pursue it, even in the face of statements in the literature that it is not possible, not feasible, or not true.

We all believe that for an academic career, what you study is your choice and therefore it is important to choose what interests you. Always go back to original sources and thoroughly read the literature. Do not rely solely on search engines because you might miss connections between different literatures.

Balancing family and career is possible, challenging, and rewarding. Finding a balance that works can be the harder part and this will change over time. A supportive husband or partner, friends, mentors, and colleagues, as well as a flexible

work schedule are invaluable. When looking for a position, consider the attitudes toward working women and family policies at the institution and laws within that country. For example, several of us choose to be parents and to be professors. The choice to be a parent can impact research productivity but recognize that this is temporary and does not imply that you are not serious about your career. It is not strange that there might be gaps in productivity, which coincide with major life events. Life does not always conform to an academic calendar. But planning can help your career a lot even when life events happen.

5.2 Things Not to Do

During our careers, there are also things we have learned that is better to avoid, and below is a short list of some “don’ts”:

Don’t say yes to committees just because they need to fill the female or minority spot, unless you really want to do the work.

Don’t say no to something you wish to do because you have never done it before or you are unsure about your capacity. If you are interested in the topic you will learn during the process.

Don’t let shyness or modesties stand in your way of your achievements. Many of us may be introverts and find it uncomfortable to present our work in public; however, recognize that those who are in the audience want to hear about your accomplishments.

Don’t just hang with the crowd you know at conferences. Try to meet and engage with new people in the area which interest you.

Don’t only attend sessions in your own area. Attending other sessions is an opportunity to learn and expand your knowledge.

Don’t be selfish in collaborations, especially with younger researchers. Your generosity most likely will be rewarded later.

Don’t discount or underestimate your knowledge. If you are in a meeting you are probably there as an expert.

Don’t always believe what you read in the literature. Knowledge and understanding evolve over time, and mistakes do sometimes slip by reviewers and editors.

Don’t despair if you get a reject/revise on a submitted paper. This means a bit more work and it will probably get published.

6 Future Directions

In the future, it is important to help our peers: to support young researchers and to help to build organizational structures to promote a healthy career. As senior researchers we should be aware of gender inequalities and make sure that scientific

conference program organizers make significant efforts to represent both genders and their scientific contributions in the keynote and invited talks. This goal serves to promote and acknowledge the work done by women researchers but also very importantly for creating role models for the younger generations. On one hand, we would like to acknowledge that we are fortunate to see how the science environment is changing and becoming more inclusive, while preserving and applying the high standards to all. On the other hand, it is crucial to continue to address the issues of gender and other inequalities that characterize most aspects of our jobs including recruitment, promotions, opportunities for collaborations, publishing our work, and other contributions to our respective working environments. Part of it is to understand that gender inequality has a negative impact on our profession and society. If we believe that our fields are exciting and important and that impact our and future generations then we should get all the talent we can get. There are many historical reasons as discussed in the introduction for the gender gap. We need to continue addressing how important it is for our generation of women to take an active role in promoting women's work and contributions, for mentoring women to help them progress and get promoted. Achieving these goals that are within our means can create a more balanced and healthy working environment and society for all.

There are also many systematic initiatives from recognized professional bodies. The London Mathematical Society (LMS) is committed to actively addressing the issues facing women in mathematics. It is concerned about the loss of women from mathematics, particularly at the higher levels of research and teaching, and at the disadvantages and missed opportunities that this represents for the advancement of mathematics. The LMS Council Statement on Women in Mathematics recognizes the need to give active consideration to ensuring that men and women are treated equally in their prospects, recognition, and progression.

The Association for Women in Mathematics' purpose (1971) is to encourage women and girls to study and pursue careers in the mathematical sciences, and to promote equal opportunity and equal treatment of women and girls in the mathematical sciences. There is also the "This is Statistics" campaign to pitch Big Data professions to middle and high school girls and minorities. This is very important since Data Science and Big Data analysis is an emerging field. Other initiatives include a yearly conference: Women in Statistics and Data Science (since 2016). The R-Ladies is a worldwide organization whose mission is to promote gender diversity in the R community.

Among the things we do and we should continue doing: address stereotyping in educational and training choices at school (and at home) at a young age, adopt teaching strategies to increase engagement of girls in mathematics, act as role models, achieve a better gender balance of teaching at all levels of education, and promote STEM professions among young women. In addition, we should organize and run regular workshops at conferences with themes that provide training on leadership skills (how to be influential and impactful): career events and workshops focusing on female students and junior academics on how to empower women and minorities. The importance of role models: strong representation of women

in keynote and invited talks as well as larger representation of women in editorial boards and editorships.

It is not enough to increase the quotas for female participation. We also need to create an environment in which women will have an equal voice and can prosper in their careers and personal lives, which is linked to the rate and time of promotions for women. To quote the character from *Ratatouille*, “anyone can be a chef, but not everybody can be a chef.”

References

- Almond, R., Mislevy, R., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian networks in educational assessment*. New York: Springer.
- Anderson, C. J., & Vermunt, J. K. (2000). Log-multiplicative association models as latent variable models for nominal and/or ordinal data. *Sociological Methodology*, *30*, 81–121.
- Black, C., & Islam, A. (2014). Women in academia: What does it take to reach the top? *The Guardian*. Retrieved from <https://www.theguardian.com/higher-education-network/blog/2014/feb/24/women-academia-promotion-cambridge>
- Cochran, W.G., & Cox, G.M. (1957). *Experimental Design*. 2nd Edition, John Wiley and Sons, New York.
- González, J., & Wiberg, M. (2017). *Applying test equating methods – Using R*. Cham: Springer.
- Holzinger, K. J., & Swineford, F. (1937). The bifactor method. *Psychometrika*, *2*, 41–54. <https://doi.org/10.1007/BF02287965>.
- Huerta-Sanchez, E., & Rolfs, R. (2019). *Hidden contributor to professor: An interview with Margaret Wu. From genes to genomes*. Retrieved from <http://genestogenomes.org/margaret-wu>
- Mason, M. A., & Wolfinger, N. H. (2013). *Do babies matter? Gender and family in the Ivory tower*. New Brunswick: Rutgers University Press.
- McCurry, J. (2019). Women outperform men after Japan medical school stops rigging exam scores. *The Guardian*. Retrieved from <https://www.theguardian.com/world/2019/jun/19/women-outperform-men-after-japan-medical-school-stops-rigging-exam-scores>
- Okahana, H., & Zhou, E. (2017). *Graduate enrollment and degrees: 2006 to 2016*. Washington, DC: Council of Graduate Schools.
- Rubini, M., & Menegatti, M. (2014). Hindering women’s careers in academia: Gender linguistic bias in personnel selection. *Journal of Language and Social Psychology*, *33*(6), 632–650.
- van den Brink, M., Benschop, & Janssen, W. (2010). Transparency as a tool for gender equality. *Organization Studies*, *31*, 1459–1483.
- Yan, D., von Davier, A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. Boca Raton: Chapman and Hall.
- Yan, D., Rupp, A. A., & Foltz, P. (2020). *Handbook of automated scoring: Theory into practice. Assessment*, *23*, 279–291. Taylor & Francis. <https://doi.org/10.1177/1073191115583714>.
- Yong, E. (2019). The women who contributed to science but were buried in the footnotes. *The Atlantic*. Retrieved from <https://www.theatlantic.com/science/archive/2019/02/womens-history-in-science-hidden-footnotes/582472>

Developing a Concept Map for Rasch Measurement Theory



George Engelhard Jr and Jue Wang

Abstract The purpose of this paper is to identify and describe the key concepts of Rasch measurement theory (Rasch G, Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen. (Expanded edition, Chicago: University of Chicago Press, 1980), 1960/1980). There have been several taxonomies describing item response theory (Kim S-H et al., A taxonomy of item response models in *Psychometrika*. In: Wiberg M, Culpepper S, Janssen R, Gonzáles J, Molenaar D (eds) *Quantitative psychology: 83rd annual meeting of the Psychometric Society*. Springer, New York City, pp 13–23, 2019; Thissen D, Steinberg L, *Psychometrika* 51:567–577, 1986; Wright BD, Masters GN, *Rating scale analysis: Rasch measurement*. MESA Press, Chicago, 1982), and this paper extends these ideas with a specific focus on Rasch measurement theory. Rasch’s measurement work reflects a key milestone in a paradigmatic shift from classical test theory to item response theory (van der Linden WJ, *Handbook of item response theory, volume 1: models*. CRC Press, Boca Raton, 2016). We include a categorization of measurement models that are commonly viewed as Rasch models (dichotomous, rating scale, partial credit, and many-faceted), as well as extensions of these models (mixed, multilevel, multidimensional, and explanatory models). Georg Rasch proposed a set of principles related to objectivity and invariance that reflect foundational concepts underlying science. Rasch measurement theory is the application of these foundational concepts to measurement. Concept maps provide useful didactic tools for understanding progress in measurement theory in the human sciences, and also for appreciating Rasch’s contributions to current theory and practice in psychometrics.

G. Engelhard Jr (✉)

Quantitative Methodology Program, Educational Psychology, The University of Georgia, Athens, GA, USA

e-mail: gengeh@uga.edu

J. Wang

Research, Measurement, and Evaluation Program, The University of Miami, Coral Gables, FL, USA

e-mail: jue.wang@miami.edu

© Springer Nature Switzerland AG 2020

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 322, https://doi.org/10.1007/978-3-030-43469-4_2

Keywords Rasch measurement theory · Philosophy of measurement · Invariant measurement

1 Introduction

The concept of “objectivity” raises fundamental problems in all sciences. For a statement to be scientific, “objectivity” is required. (Rasch 1964, p. 1)

Rasch described several models for measurement that he developed to address problems encountered in his research work. His seminal book entitled *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch 1960/1980) introduced several models of measurement including models for misreadings, reading speed, and item analysis. These models became the basis for numerous advances in measurement theory.

Rasch measurement theory has been described as “a truly new approach to psychometric problems . . . [that yields] non-arbitrary measures” (Loevinger 1965, p. 151). As pointed out by van der Linden (2016), the first chapter of Rasch’s book is required reading for anyone seeking to understand the transition from classical test theory to item response theory (IRT). In his words, “One of the best introductions to this change of paradigm is Rasch (1960/1980, Chapter 1), which is mandatory reading for anyone with an interest in the subject” (van der Linden 2016, p. xvii). Wright (1980) commented that Rasch’s psychometric methods “go far beyond measurement in education or psychology. They embody the essential principles of measurement itself, the principles on which objectivity and reproducibility, indeed all scientific knowledge, are based” (p. xix). This study explores what Rasch did to receive these accolades.

In order to explore current perspectives on Rasch measurement theory, we conducted a Web of Science search using the topic phrase “Rasch measurement theory”. This bibliometric search was limited to the twenty-first century (2000–2019), and 754 references were identified. Figure 1 shows frequency of articles related to Rasch measurement theory. It is also interesting to note the distribution of these articles over various fields with psychology (N = 240), health care sciences (N = 125), and educational research (N = 109) identified as the top three areas.

The purpose of this study is to identify the key concepts that define Rasch measurement theory. Specifically, the following questions guide our research: (a) What is Rasch measurement theory? (b) What are the key concepts that define Rasch measurement theory?

2 What Is Rasch Measurement Theory?

One way to define Rasch measurement theory is by the specific models for measurement proposed by Rasch, and also the models that are considered extensions

Total Publications
754 Analyze

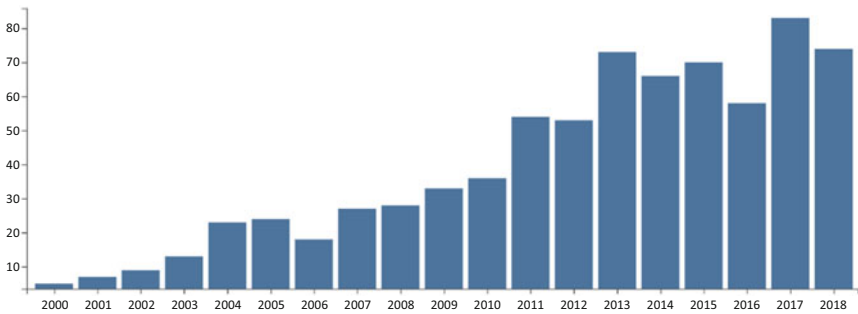


Fig. 1 Frequency of articles on Rasch measurement theory (Web of Science)

Name	Log-odd Forms of Rasch Models
Dichotomous Model	$\text{Ln} \left(\frac{P_{ni1}}{P_{ni0}} \right) = \theta_n - \delta_i$
Partial Credit Model	$\text{Ln} \left(\frac{P_{nik}}{P_{nik-1}} \right) = \theta_n - \delta_{ik}$
Rating Scale Model	$\text{Ln} \left(\frac{P_{nik}}{P_{nik-1}} \right) = \theta_n - (\delta_i + \tau_k) = \theta_n - \delta_i - \tau_k$
Many Facet Models	
Many Facet Partial Credit Model	$\text{Ln} \left(\frac{P_{nmik}}{P_{nmik-1}} \right) = \theta_n - \lambda_m - \delta_{ik}$
Many Facet Rating Scale Model	$\text{Ln} \left(\frac{P_{nmik}}{P_{nmik-1}} \right) = \theta_n - \lambda_m - (\delta_i + \tau_k) = \theta_n - \lambda_m - \delta_i - \tau_k$

Fig. 2 Commonly used Rasch Models. θ_n = person ability measure; δ_i = difficulty of item i ; δ_{ik} = difficulty of step k of item i (assuming unique scale structure of each item); τ_k = difficulty of step k (assuming common scale structure among all items); λ_m = scoring severity of rater m

of the unidimensional Rasch model. One of the earliest taxonomies of Rasch models is offered by Wright and Masters (1982). They described a family of Rasch models designed to analyze dichotomous and polytomous responses obtained from persons based on items that are developed to represent a unidimensional continuum. Specifically, Wright and Masters (1982) described five Rasch models: Dichotomous, Partial Credit, Rating Scale, Binomial Trials, and Poisson Count. Linacre (1989) extended this family of Rasch models to include raters. Figure 2 describes the most commonly used Rasch models.

Kim and his colleagues (2019) categorized IRT articles appearing in *Psychometrika* from 1960s to 2010s. They identified 157 articles related to Rasch measurement theory. About 41.64% of the total IRT articles in *Psychometrika* are related to Rasch measurement theory. Figure 3 shows the frequency of articles on Rasch measurement theory over time by model type.

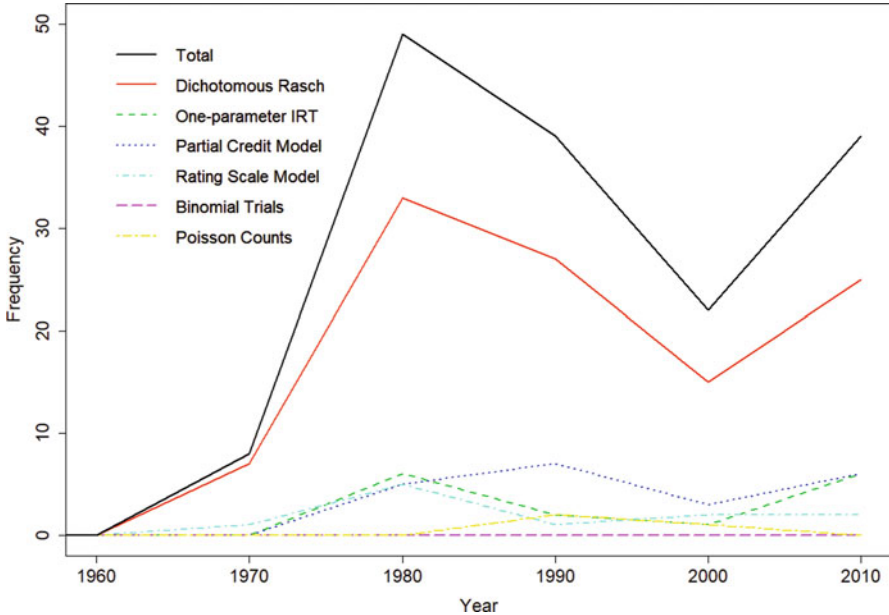


Fig. 3 Frequency of articles on Rasch models over time published in *Psychometrika* (Kim et al. 2019). Total (black solid line) shows the frequency of all Rasch models over time. Further bibliometric evidence shows that Rasch measurement theory continues to influence measurement research as provided by Aryadoust and Tan (2019)

There have been numerous extensions to Rasch models. Here is a partial list of the extensions: (a) mixed Rasch model (Rost 1990), (b) multilevel Rasch measurement model (Adams et al. 1997b), and (c) multidimensional random coefficients multinomial logit models (Adams et al. 1997a). Since research continues on extensions to Rasch measurement theory, this list should be considered incomplete.

In addition to defining Rasch measurement theory based on models for measurement that specifically include Rasch’s name, a complementary approach is to consider the key concepts that define Rasch measurement theory. These key concepts considered in the next section are based on Rasch’s views of science and the application of these concepts to measurement.

3 What Are the Key Concepts that Define Rasch Measurement Theory?

Looking then for concepts [of measurement] that could possibly be taken as primary it seems worthwhile to concentrate upon two essential characteristics of “scientific statements” 1. they are concerned with “comparisons”; 2. the statements are claimed to be “objective”; both terms of course calling for precise qualifications. (Rasch 1964, p.2)

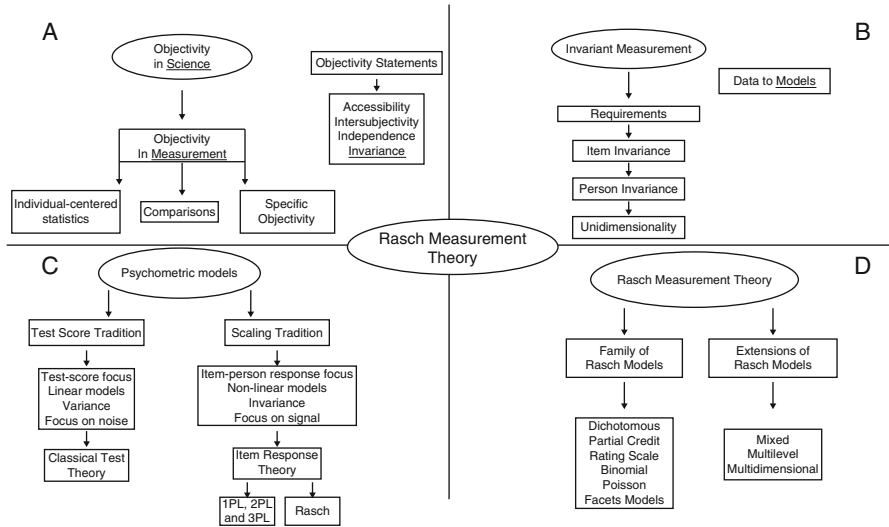


Fig. 4 Concept map for Rasch measurement theory

The overarching concepts that underly Rasch measurement theory are shown in Fig. 4. The four panels represent conceptual clusters that are important for understanding Rasch measurement theory including objectivity in science, invariant measurement, psychometric models, and (Rasch) models for measurement. Panels A through D in Fig. 4 are discussed in each section below

3.1 Objectivity in Science (Panel A)

Rasch was motivated by a desire to develop “individual-centered statistical techniques in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated” (Rasch 1960/1980, p. xx). In order to develop individual-centered statistical techniques, Rasch started with a perspective on scientific statements that included a concern with comparisons and objectivity.

Objectivity in science is based on objective statements that have several key features including accessibility, intersubjectivity, independence, and *invariance* (Nozick 1998, 2001). Objective statements are accessible from different angles implying that they can be repeated by different observers and at different times. Intersubjectivity implies that there is agreement among observers about a scientific fact. Next, objective statements are independent of the particular observers. Objectivity in science depends on objective statements that reflect accessibility, intersubjectivity, independence, and most importantly invariance that implies the first three characteristics.

Rasch's key insight was that principles of objectivity in science can also be applied to *objectivity in measurement*. In order to examine objectivity Rasch suggested that *comparisons* are a key concept in science. Rasch defined four requirements as follows:

The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.

Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (Rasch 1977, pp. 331–332)

Rasch also recognized the importance of identifying the specific conditions under which comparisons are invariant. Rasch's view of *specific objectivity* reflects what Nozick (2001) has pointed out in his work on invariance and objectivity:

What is objective about something, I have claimed, is what is invariant from different angles, across different perspectives, under different transformations. Yet often what is variant is what is especially interesting. We can take different perspective on a thing (the more angles the better), and it notice which of its features are objective and invariant, and also notice which of its features are subjective and variant (p. 102)

The next section expands on the requirements for specific objectivity based on the idea of invariant measurement.

3.2 *Invariant Measurement (Panel B)*

The scientist is usually looking for invariance whether he knows it or not. (Stevens 1951, p. 20)

As pointed out in the previous section, Rasch's views of objectivity can be interpreted as part of the quest for stable and invariant measurement. In seeking stable measures, Wright (1968) identified the following requirements based on Rasch measurement theory:

First, the calibration of measuring instruments must be independent of those objects that happen to be used for calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for the measuring (p. 87).

The first part of this quote refers to *person-invariant item calibration*. The basic measurement problem addressed by sample-invariant item calibration is how to minimize the influence of arbitrary samples of individuals on the estimation of item scale values or item difficulties. The overall goal of person-invariant measurement can be viewed as estimating the locations of items on a latent variable or construct of interest.

The second part refers to *item-invariant measurement of persons*. In the case of item-invariant measurement, the basic measurement problem is to minimize the

influences of the particular items that happen to be used to estimate a person's location on the latent variable or construct. Overall, both item and person locations should remain stable and consistent across various subsets of items and subgroups of persons.

It is interesting to note that the concept of invariance has played a key role in several theories of measurement (Engelhard 2013; Millsap 2011). The quest for invariance emerged in the work of early measurement theorists from Thorndike (1904) through Thurstone (1925, 1926) to Rasch (1960/1980). We view Rasch's concept of objectivity as essentially synonymous with the term invariance. Engelhard (2013) summarized five requirements of invariant measurement based on Rasch measurement theory as follows.

Person measurement:

1. The measurement of persons must be independent of the particular items that happen to be used for the measuring: *Item-invariant measurement of persons*.
2. A more able person must always have a better chance of success on an item than a less able person: *non-crossing person response functions*.

Item calibration:

3. The calibration of the items must be independent of the particular persons used for calibration: *Person-invariant calibration of test items*.
4. Any person must have a better chance of success on an easy item than on a more difficult item: *non-crossing item response functions*.

Variable map:

5. Items and person must be simultaneously located on a single underlying latent variable: *variable map*.

It is also important to recognize that Rasch specified the requirements of Rasch measurement theory a priori. This implies that model-data fit should stress the model over the data. In other words, we are confirming that a specific data set meets the requirements of invariant measurement. Statistical modeling typically stresses the reproduction of the data, and therefore privileges the data over the model.

3.3 Psychometric Models (Panel C)

Panel C (Fig. 4) embeds Rasch measurement theory in a broader historical story during the twentieth century. It is useful to view the history of psychometric models through two broad traditions: the test-score and scaling traditions.

The test-score tradition includes measurement theories that stress the total or sum scores as implied by the label. Linear models are used to model the person scores, and the goal is to estimate sources of error variance or uncertainty in scores—essentially, these models stress the reduction of noise reflected in the estimates of

error variance. Classical test theory is a clear example of measurement within the test-score tradition.

The scaling tradition on the other hand focuses on the individual responses of each person to each item. These item-person responses are modeled with non-linear probability models (e.g., logistic models). Measurement theories within the scaling tradition are used to define an invariant continuum with item and person locations representing a latent variable. These models emphasize the signal as defined by the invariant continuum. Most modern measurement theories including IRT (1PL, 2PL, 3PL), as well as Rasch measurement theory are within the scaling tradition. See Engelhard (2013) for a more detailed description of these two research traditions in measurement.

3.4 (Rasch) Models for Measurement (Panel D)

In a previous section, we listed measurement models that are included in the family of Rasch models (Dichotomous, Partial Credit, Rating Scale, Binomial Trials, Poisson Count, and Facets), and we also provided a brief list of extensions to the Rasch model (e.g., Mixed Rasch model, Multilevel Rasch measurement model, and Multidimensional random coefficients multinomial logit models). We have put Rasch's name in parentheses in this section because it is important to consider whether or not some of these models should be considered part of Rasch measurement theory. As pointed out by (Andersen 1995), "[Rasch] was very eager not to call the model the 'RM'. Instead, he suggested the name 'models for measurement' . . . his suggested name had the clear purpose of stressing the most important property of the model: that it solved a basic measurement problem in the social sciences, or as it became later, in Georg Rasch's opinion in all sciences." (p. 384).

Rasch believed that he had developed a general framework for models for measurement. A key part of Rasch's distinctive contribution to methodology is based on invariant comparisons within a frame of reference (Andrich 2018). In Rasch's words, "Let us imagine two collections of elements, O and A, denoted here objects and agents . . . the results of any comparison of two objects within O is independent of the choice of the agents A_i within A and also of the other elements in the collection of objects O" (Rasch 1977, p. 77). The converse is also true for the comparison of agents. Rasch's view of additivity based on invariant comparisons within a specific frame-of-reference can provide the basis for considering how extensions to Rasch measurement theory fit within this framework. According to Andrich (2018),

[Rasch articulated] the requirements of invariant comparison within an empirical, specified frame of reference, and the rendering of these in a probabilistic mathematical framework, being a very general and powerful theory of measurement relevant to both the natural and social sciences (p. 88).

4 Discussion

In 1977, when [Rasch] was primarily occupied with basic issues of philosophy of science, he would, I think, have been very disappointed that the developments up through the 80s and early 90s have been so much concerned with statistical techniques, while so few scientists have worked on basic philosophical issues. (Andersen 1995, p. 389)

In order to answer the guiding research questions, we used several approaches. First, we used bibliometric methods to document the continuing importance of Rasch measurement theory. Next, we described previous classifications and taxonomies related to Rasch models. Finally, we stressed key concepts that help explain Rasch's continuing influence on measurement theory and practice.

In returning to the guiding questions, many psychometricians when queried about "What is Rasch measurement theory?" tend to respond that it is a "logistic model with a slope of one and of course no lower asymptote". In this study, we argue that Rasch models represent a philosophy of measurement that includes the application of basic scientific concepts to models of measurement. The careful consideration of key concepts regarding invariant measurement identified by Rasch, and their realization in different measurement models is one of the challenges that remains for the next generation of measurement theorists. It is our hope that others will take on the challenge and extend the start that we made in this chapter.

The second guiding question is: What are the key concepts that define Rasch measurement theory? We have identified Rasch's quest for individual-level statistics, invariant comparisons, and specific objectivity as key aspects of his measurement theory. Invariant measurement offers both item-invariant person measurement and person-invariant item calibration. This aligns with Rasch's views of objective comparison and specific objectivity. Rasch measurement theory realizes the requirements of invariant measurement through the formation of a unidimensional scale based on a close examination of particular items. A key feature of Rasch measurement theory is seeking for the best scale that is defined by the good-quality items, instead of increasing the model complexity to fit the data.

In summary, we view Rasch measurement theory as representing scientific principles based on the concept of specific objectivity (invariant comparisons) applied to models for measurement. As Messick (1983) defined, "Theories of measurement broadly conceived may be viewed as loosely integrated conceptual frameworks within which are embedded rigorously formulated statistical models of estimation and inference about the properties of measurements and scores" (p. 498). The statistical models serve as tools to provide meaningful scores based on conceptual frameworks. Rasch measurement theory aims to create invariant scales and provide objective measures. Future research should consider the key characteristics that define whether or not a model for measurement meets the requirements of Rasch measurement theory.

References

- Adams, R. J., Wilson, M. R., & Wang, W. (1997a). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.
- Adams, R. J., Wilson, M., & Wu, M. (1997b). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*, 47–76.
- Andersen, E. B. (1995). What Georg Rasch would have thought about this book. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 383–390). New York: Springer.
- Andrich, D. A. (2018). A Rasch measurement theory. In F. Guillemin, A. Leplege, S. Briancon, E. Spitz, & J. Coste (Eds.), *Perceived health and adaptation in chronic disease* (pp. 66–91). New York: Routledge.
- Aryadoust, V., & Tan, H. A. H. (2019). *A systematic review of Rasch measurement in psychology, medicine, and education: The rise and progress of a specialty*. Manuscript submitted for publication.
- Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.
- Kim, S.-H., Bian, M., Feldberg, Z., Henry, T., Kwak, M., Lee, J., Olmetz, I. B., Shen, Y., Tan, Y., Tanaka, V. T., Wang, J., Xu, J., & Cohen, A. S. (2019). A taxonomy of item response models in Psychometrika. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzáles, & D. Molenaar (Eds.), *Quantitative psychology: 83rd annual meeting of the Psychometric Society* (pp. 13–23). New York City: Springer.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.
- Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*, 143–155.
- Messick, S. (1983). Assessment of children. In P. H. Mussen (Ed.), *Handbook of child psychology, volume 1: History, theory and methods* (pp. 477–526). New York: Wiley.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.
- Nozick, R. (1998). Invariance and objectivity. *Proceedings and addresses of the American Philosophical Association*, *72*(2), 21–48.
- Nozick, R. (2001). *Invariances: The structure of the objective world*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).
- Rasch, G. (1964). *On objectivity and models for measuring*. Lecture notes edited by Jon Stene. [Memo # 196z PDF \(1.0 MB\)](#).
- Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, *14*, 58–94.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.
- Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.
- Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York: Teachers College, Columbia University.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433–451.
- Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology*, *17*, 446–457.
- van der Linden, W. J. (Ed.). (2016). *Handbook of item response theory, volume 1: Models*. Boca Raton: CRC Press.

- Wright, B. D. (1968, February). Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems* (pp. 85–101).
- Wright, B. D. (1980). Foreword and afterword. In Rasch (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Expanded edition pp. ix–xix, 185–196). Chicago: University of Chicago Press, 1980.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Person Parameter Estimation for IRT Models of Forced-Choice Data: Merits and Perils of Pseudo-Likelihood Approaches



Safir Yousfi

Abstract The Thurstonian IRT model for forced-choice data (Brown A, Maydeu-Olivares A, *Educ Psychol Measur* 71:460–502, 2011) capitalizes on including structural local dependencies in the structural equation model. However, local dependencies of pairwise comparisons within forced-choice blocks are only considered for item parameter estimation by this approach but are explicitly ignored by the respective methods of person parameter estimation. The present paper introduces methods of person parameter estimation (MLE, MAP, and WLE) that rely on the exact likelihood of the response pattern that adequately considers local stochastic dependencies by multivariate integration. Moreover, it is argued that the common practice of ignoring local stochastic dependencies for person parameter estimation can be understood as a pseudo-likelihood approach (based on the independence likelihood) that will lead to similar estimates in most applications. However, standard errors and Bayesian estimation techniques are affected by falsely precise inference based on the independence likelihood. Fortunately, these distortions can be amended almost completely by a correction factor to the (independence) pseudo-likelihood for MLE and MAP estimation. Moreover, unbiased weighted (pseudo-)likelihood estimation becomes feasible without facing the prohibitive computational burden of weighted likelihood estimation with the proper likelihood based on multivariate integration.

Keywords Forced-choice · Thurstonian IRT model · IRT · Person parameter estimation · Composite likelihood

S. Yousfi (✉)
German Federal Employment Agency, Nuremberg, Germany
e-mail: Safir.yousfi@arbeitsagentur.de

1 Introduction

Requiring respondents to assign ranks to questionnaire items that reflect their preference within a block of items (i.e., the forced-choice method) potentially reduces or eliminates item response biases (e.g., acquiescence, extreme responding, central tendency responding, halo/horn effect, social desirable response style) typically associated with direct responses (like Likert-type or Yes/No ratings). However, the ipsative nature of forced-choice data results in problematic psychometric properties of classical scoring methods (e.g., sum scores), i.e., construct validities and criterion-related validities, and reliabilities are distorted (Brown and Maydeu-Olivares 2013). Recently, Brown and Maydeu-Olivares (2011) proposed an IRT approach to modeling and analyzing forced-choice data that effectively overcomes these problems by binary coding and considering local dependencies of the binary response indicators in the process of estimating the *structural* model parameters, i.e., item parameters and latent trait correlations. However, the proposed methods of *person parameter estimation* explicitly neglect local dependencies of the binary response indicators that arise in blocks with more than two items.

Moreover, the approach of Brown and Maydeu-Olivares (2011, 2013) for quantifying the precision of the person parameter estimates relies on directional information derived from the curvature of the log-likelihood (of the observed binary comparison) in direction of the latent trait under consideration. This approach is generally not suited because it allows only very limited information about the precision of the estimates:

- The correlation of the estimation errors with respect to different dimensions cannot be quantified (cf. Brown and Maydeu-Olivares 2018).
- The inverse of directional information in the direction of the latent trait equals the squared standard error of vector-valued maximum likelihood estimates only under special circumstances (namely, if an eigenvector of the Fisher Information matrix points into the respective direction in latent space which happens only in case of zero correlations with the errors with respect to the other traits).

The precision of a vector-valued maximum likelihood estimate can be quantified by the (expected or the observed) Fisher information matrix whose inverse equals the asymptotic covariance of the estimator. Brown and Maydeu-Olivares (2018) used this approach to extend the Thurstonian IRT model to graded preference data, but the proposed method of person parameter estimation for ranking data still neglects local dependencies of binary response indicators.

Nevertheless, the empirical findings of Brown and Maydeu-Olivares (2011) demonstrate that their estimation technique shows good parameter recovery but with moderate overestimation of the precision. An explanation for these observations can be found in the literature of composite likelihood methods (see Varin et al. 2011 for an overview). Approximation of the likelihood by a product of marginal likelihoods (i.e., the independence likelihood according to Chandler and Bate 2007) generally leads to consistent estimators. However, the respective estimators lack the

property optimal efficiency of the genuine maximum likelihood estimator (Lindsay 1988). Moreover, the inverse of the sum of the Fisher information with respect to all marginal likelihood functions cannot be expected to be a consistent estimator of the covariance matrix of the maximum independence likelihood estimator (Chandler and Bate 2007).

In this paper, the genuine maximum likelihood estimator considering local dependencies (Yousfi 2018) and a calibrated maximum composite likelihood estimator of the latent trait will be introduced, and adequate estimates of the precision will be provided. Several estimation techniques for the latent trait emerge from the respective analyses.

2 Thurstonian MIRT Model of Forced-Choice Data

2.1 Notation

$(\)$ is used to extract elements from vectors or matrices. The entries in the brackets are positive integers and refer rows and columns, respectively.

$\langle \ \rangle$ is used to extract parts from vectors or matrices, respectively. The entries in the brackets are vectors of positive integers and refer to rows and columns, respectively.

The sign \bullet (i.e., a bold dot) indicates that all rows or columns are extracted.

2.2 Binary Coding of Forced-Choice Data

Let $\mathbf{y}_{\mathbf{b}}$ be a random variable whose values denote the response of a person to the forced-choice block \mathbf{b} which consists of $n_{\mathbf{b}}$ items. For instance, $\mathbf{y}_{\mathbf{b}} = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$ would indicate that the respondent shows the strongest preference for the third item of block \mathbf{b} and the lowest preference for second response options. The response pattern to a full forced-choice questionnaire of K blocks can be described by a sequence of K rankings $\mathbf{Y} : (\mathbf{y}_1, \dots, \mathbf{y}_{\mathbf{b}}, \dots, \mathbf{y}_K)$.

Let $\mathbf{Y}_{\mathbf{b}}$ be a random quadratic matrix of dimension $n_{\mathbf{b}} \times n_{\mathbf{b}}$, whereby the entry in p -th row and the q -th column refers to the binary response variable $y_{pq} : \mathbf{Y}_{\mathbf{b}(p,q)}$ with:

$$\mathbf{Y}_{\mathbf{b}(p,q)} = \begin{cases} 1 & \text{if } \mathbf{y}_{\mathbf{b}(p)} > \mathbf{y}_{\mathbf{b}(q)} \\ 0 & \text{if } \mathbf{y}_{\mathbf{b}(p)} \leq \mathbf{y}_{\mathbf{b}(q)} \end{cases} \quad (1)$$

Brown and Maydeu-Olivares (2011) referred only to the entries above the diagonal of \mathbf{Y}_b which results in a full description of the data as $y_{pq} = 1 - y_{qp}$.

2.3 Model Equations

Thurstone's law of comparative judgment states that the observed binary comparisons of the items (i.e., the entries of \mathbf{Y}_b) are determined by a vector of latent utilities $\mathbf{t}_b \in \mathbb{R}^{n_b}$ in the following way:

$$\mathbf{Y}_{b(p,q)} = \begin{cases} 1 & \text{if } \mathbf{t}_{b(p)} - \mathbf{t}_{b(q)} \geq 0 \\ 0 & \text{if } \mathbf{t}_{b(p)} - \mathbf{t}_{b(q)} < 0 \end{cases} \quad (2)$$

For each pattern of latent traits (i.e., true factor scores) $\boldsymbol{\theta} \in \mathbb{R}^m$, the entries in \mathbf{t}_b are assumed to be multivariate normally distributed:

$$\mathbf{t}_b \sim N(\boldsymbol{\mu}_b + \mathbf{A}_b \boldsymbol{\theta}, \boldsymbol{\Psi}_b) \quad (3)$$

The vector $\boldsymbol{\mu}_b \in \mathbb{R}^{n_b}$ refers to the intercepts of items and the real $(m \times n_b)$ matrix \mathbf{A}_b refers to the factor loadings of the items of block \mathbf{b} . $\boldsymbol{\Psi}_b$ is the covariance matrix of the latent utility residuals that reflect the stochastic nature of the observed response for a given trait pattern. $\boldsymbol{\Psi}_b$ is usually assumed to be diagonal. $f_{\mathbf{t}_b}$ is the probability density function of \mathbf{t}_b .

3 The Likelihood Function of a Forced-Choice Block

Let \mathbf{D}_b be the (random) matrix that transforms \mathbf{t}_b to the ordered vector of utilities, i.e.,

$$\mathbf{t}_{b(y_b)} = \mathbf{D}_b \mathbf{t}_b \quad (4)$$

and let \mathbf{T}_{n_b} be a quadratic matrix with n_b rows and columns, whereby:

$$\mathbf{T}_{n_b(p,q)} = \begin{cases} 1 & \text{if } p \in \{q, n_b\} \\ -1 & \text{if } p = q - 1 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Then the vector of utility-differences of items with subsequent ranks is given by:

$$\mathbf{d}_b := \mathbf{T}_{n_b} \mathbf{t}_{b(y_b)} = \mathbf{T}_{n_b} \mathbf{D}_b \mathbf{t}_b \sim N(\mathbf{T}_{n_b} \mathbf{D}_b \boldsymbol{\mu}_b + \mathbf{T}_{n_b} \mathbf{D}_b \boldsymbol{\Lambda}_b \boldsymbol{\theta}, \mathbf{T}_{n_b} \mathbf{D}_b \boldsymbol{\Psi}_b \mathbf{D}'_b \mathbf{T}'_{n_b}) \quad (6)$$

The likelihood of \mathbf{y}_b (the response to block \mathbf{b}) is given by (cf. Maydeu-Olivares 1999):

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_b) = \mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}_b) = \int_{\mathbf{S}_b} f_{\mathbf{t}_b}(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{D}_b \mathbf{S}_b} f_{\mathbf{t}_{b(y_b)}}(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{T}_{n_b} \mathbf{D}_b \mathbf{S}_b} f_{\mathbf{d}_b}(\mathbf{x}) d\mathbf{x} \quad (7)$$

\mathbf{S}_b refers to the region of \mathbb{R}^{n_b} where the following system of $n_b - 1$ inequalities holds true:

$$\mathbf{C}_b \mathbf{t}_{b(y_b)} \geq \mathbf{0}^{n_b-1} \quad (8)$$

whereby $\mathbf{0}^{n_b-1} \in \mathbb{R}^{n_b-1}$ is a vector of $n_b - 1$ entries of 0 and \mathbf{C}_b is a contrast matrix with $n_b - 1$ rows and n_b columns with:

$$\mathbf{C}_{b(p,q)} = \begin{cases} 1 & \text{if } q = p \\ -1 & \text{if } q = p + 1 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$\mathbf{D}_b \mathbf{S}_b$ is the image of \mathbf{S}_b with respect to the transformation described by \mathbf{D}_b (i.e., reordering the axes according the order of \mathbf{t}_b). $\mathbf{T}_{n_b} \mathbf{D}_b \mathbf{S}_b$ is the image of \mathbf{S}_b after (a) rotation of the first $n_b - 1$ axes of the coordinate system of the utility space onto axes that correspond to utility differences of items with subsequent ranks and (b) rotation of last axis onto an axis that corresponds to the sum of all n_b utilities.

Geometrically, $\mathbf{S}_b \subsetneq \mathbb{R}^{n_b}$ and $\mathbf{D}_b \mathbf{S}_b \subsetneq \mathbb{R}^{n_b}$ are unbounded full-dimensional polyhedral cones¹ (whereby all facets intersect in the only edge). In contrast, $\mathbf{T}_{n_b} \mathbf{D}_b \mathbf{S}_b \subsetneq \mathbb{R}^{n_b}$ consists of two neighbored orthants² of the coordinate system that results after (the oblique) rotation (and rescaling) in a way (a) that the first $n_b - 1$ coordinate axes correspond to utility differences of items with neighbored ranks and (b) the last coordinate axis corresponds to the sum of all item utilities. This facilitates integration by considering the set:

¹Describing the respective region as parallelepiped (Maydeu-Olivares 1999) or parallelotope (Yousfi 2018) is somewhat misleading as these figures are typically bounded and have more than one edge at the boundary of their facets. The term “conic region” (Maydeu-Olivares 1999) might also be misunderstood as usual three-dimensional cones are bounded and not polyhedral, i.e., they do not have flat facets connected to the apex.

²Orthants are generalizations of quadrants to coordinate systems with more than two dimensions.

$$\{\mathbf{d}_b \in \mathbb{R}^{n_b} | \mathbf{d}_{b(i)} > 0 \text{ for all } i \in \{1, \dots, n_b - 1\}\} \quad (10)$$

that results by pre-multiplying $\mathbf{T}_{n_b} \mathbf{D}_b$ to all utility patterns that result in the observed ranking (i.e., elements of \mathbf{S}_b). Consequently, the likelihood of the observed ranking can be expressed as an integral over an orthant region in \mathbb{R}^{n_b-1} :

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_b) &= \int_{\mathbf{T}_{n_b} \mathbf{D}_b \mathbf{S}_b} f_{\mathbf{d}_b}(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} \int_0^{\infty} \dots \int_0^{\infty} f_{\mathbf{d}_b}(x_1, \dots, x_{n_b}) dx_1 \dots dx_{n_b} \\ &= \int_0^{\infty} \dots \int_0^{\infty} f_{\mathbf{d}_{b(n_b-1)}}(x_1, \dots, x_{n_b-1}) dx_1 \dots dx_{n_b-1} \end{aligned} \quad (11)$$

whereby $f_{\mathbf{d}_{b(n_b-1)}}$ refers to the marginal distribution of the first $n_b - 1$ elements of \mathbf{d}_b . This integral can be solved by the methods developed by Genz (2004), for blocks with no more than 4 items, and Miwa et al. (2003), up to 20 items per block.

4 Person Parameter Estimation

4.1 Maximum Likelihood

Genuine Likelihood If local stochastic independence is given with respect to the responses to different blocks, then standard numeric optimization methods can be used to determine the maximum likelihood estimator of the latent trait vector:

$$MLE(\mathbf{Y}) = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = \operatorname{argmax}_{\boldsymbol{\theta}} \prod_{b=1}^K \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_b) \quad (12)$$

The observed Fisher information is defined as:

$$\mathbf{I}(\boldsymbol{\theta}; \mathbf{Y}) := -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{b=1}^K -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}; \mathbf{y}_b) =: \sum_{b=1}^K \mathbf{I}(\boldsymbol{\theta}; \mathbf{y}_b) \quad (13)$$

whereby $\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}; \mathbf{y}_b)$ (i.e., the Hessian matrix of the log-likelihood $\ell(\boldsymbol{\theta}; \mathbf{Y}) := \log_e \mathcal{L}(\boldsymbol{\theta}; \mathbf{Y})$) can be determined numerically by considering the log-likelihood in the neighborhood of $\boldsymbol{\theta}$. Under usual regularity conditions, the inverse of the (observed or expected) Fisher information is an asymptotically unbiased estimator of the covariance matrix of the maximum likelihood estimator:

$$\begin{aligned} \operatorname{cov}(MLE(\mathbf{Y}); \boldsymbol{\theta}) &\simeq E\left(\left(\mathbf{I}(\boldsymbol{\theta}; \mathbf{Y})\right)^{-1} | \boldsymbol{\theta}\right) \simeq E\left(\left(\mathbf{I}(MLE(\mathbf{Y}); \mathbf{Y})\right)^{-1} | \boldsymbol{\theta}\right) \\ &\simeq E\left(E\left(\left(\mathbf{I}(MLE(\mathbf{Y}); \mathbf{Y})\right)^{-1} | \boldsymbol{\theta}\right)\right) \end{aligned}$$

Independence Likelihood Person parameter estimates of Brown and Maydeu-Olivares (2011) are based on the independence likelihood of item comparisons within each block (i.e., the independence likelihood according to Chandler and Bate 2007):

$$\mathcal{L}_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_{\mathbf{b}}) := \prod_{p=1}^{n_{\mathbf{b}}-1} \prod_{q=p+1}^{n_{\mathbf{b}}} \mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}_{\mathbf{b}(p,q)}) = \prod_{p=1}^{n_{\mathbf{b}}-1} \prod_{q=p+1}^{n_{\mathbf{b}}-1} \mathbf{Y}_{\mathbf{b}(p,q)} P(\mathbf{Y}_{\mathbf{b}(p,q)} = 1 | \boldsymbol{\theta}) + (1 - \mathbf{Y}_{\mathbf{b}(p,q)}) (1 - P(\mathbf{Y}_{\mathbf{b}(p,q)} = 1 | \boldsymbol{\theta})) \quad (14)$$

whereby

$$P(\mathbf{Y}_{\mathbf{b}(p,q)} = 1 | \boldsymbol{\theta}) = \Phi \left(\frac{\boldsymbol{\mu}_{\mathbf{b}(p)} - \boldsymbol{\mu}_{\mathbf{b}(q)} + (\boldsymbol{\Lambda}_{\mathbf{b}(p,\bullet)} - \boldsymbol{\Lambda}_{(q,\bullet)}) \boldsymbol{\theta}}{\sqrt{\boldsymbol{\Psi}_{\mathbf{b}(p,p)} + \boldsymbol{\Psi}_{\mathbf{b}(q,q)}}} \right) \quad (15)$$

and Φ denotes the cumulative distribution function of the standard normal distribution.

The respective estimating equation:

$$\begin{aligned} MLE_{\text{ind}}(\mathbf{Y}) &:= \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}_{\text{ind}}(\boldsymbol{\theta}; \mathbf{Y}) = \operatorname{argmax}_{\boldsymbol{\theta}} \prod_{\mathbf{b}=1}^K \mathcal{L}_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_{\mathbf{b}}) \\ &= \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{b}=1}^K \ell_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_{\mathbf{b}}) \end{aligned} \quad (16)$$

results in a consistent estimator of the latent trait vector $\boldsymbol{\theta}$, but (unless all blocks consist of pairs) neither the observed nor the expected value of the inverse of the negative Hessian matrix of independence log-likelihood at the maximum independence likelihood estimator (MLE_{ind}) needs to be asymptotically equal to the covariance of the estimator (Varin et al. 2011), i.e.,

$$\operatorname{cov}(MLE_{\text{ind}}(\mathbf{Y}); \boldsymbol{\theta}) \neq -\frac{\partial^2}{\partial \boldsymbol{\theta}} \ell_{\text{ind}}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{\mathbf{b}=1}^K \sum_{p=1}^{n_{\mathbf{b}}-1} \sum_{q=p+1}^{n_{\mathbf{b}}} \mathbf{I}(\boldsymbol{\theta}; \mathbf{Y}_{\mathbf{b}(p,q)}) \quad (17)$$

Moreover, in contrast to the maximum likelihood estimator, the maximum independence likelihood estimator does not have the property of maximum asymptotic efficiency. However, the loss of efficiency needs not to be large (Varin et al. 2011; Chandler and Bate 2007) and due to the substantially lower computational burden, the maximum independence likelihood estimator $MLE_{\text{ind}}(\mathbf{Y})$ can be an attractive alternative, if no or only a cumbersome method of genuine maximum likelihood estimation $MLE(\mathbf{Y})$ is available.

Actually, simulations show that $MLE(\mathbf{Y})$ and $MLE_{\text{ind}}(\mathbf{Y})$ hardly differ from each other, whereby the intraclass correlation (ICC(2,1) according to Shrout and Fleiss 1979) generally exceeds 0.99, if all blocks have the same size n . Moreover, simulations suggest the following conjecture with respect to variability of the maximum independence likelihood estimator³:

$$\text{cov}(MLE_{\text{ind}}(\mathbf{Y}); \boldsymbol{\theta}) \simeq \frac{n+1}{3} \left(-\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \sum_{b=1}^K \ell_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_b) \right)^{-1} \simeq (\mathbf{I}(MLE(\mathbf{Y}); \mathbf{Y}))^{-1} \quad (18)$$

Composite Likelihood This suggest fine-tuning the estimation by calibration⁴ of the independence likelihood with adequate weights which results in the following composite likelihood:

$$\mathcal{L}_{\text{comp}}(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{b=1}^K (\mathcal{L}_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_b))^{n_b+1} \quad (19)$$

The respective maximum composite likelihood estimator is:

$$MLE_{\text{comp}}(\mathbf{Y}) := \text{argmax}_{\boldsymbol{\theta}} \mathcal{L}_{\text{comp}}(\boldsymbol{\theta}; \mathbf{Y}) = \text{argmax}_{\boldsymbol{\theta}} \sum_{b=1}^K \frac{\ell_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_b)}{n_b + 1} \quad (20)$$

Simulations suggest the following conjecture with respect to the precision of this estimator:

$$\begin{aligned} \text{cov}(MLE_{\text{comp}}(\mathbf{Y}); \boldsymbol{\theta}) &\approx -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log(\mathcal{L}_{\text{comp}}(MLE_{\text{comp}}(\mathbf{Y}); \mathbf{Y})) \\ &= \left(3 \cdot \sum_{b=1}^K \frac{\sum_{p=1}^{n_b-1} \sum_{q=p+1}^{n_b} \mathbf{I}(MLE_{\text{comp}}(\mathbf{Y}); \mathbf{Y}_{b(p,q)})}{n_b + 1} \right)^{-1} \end{aligned} \quad (21)$$

$MLE_{\text{comp}}(\mathbf{Y})$ more closely resembles $MLE(\mathbf{Y})$ than $MLE_{\text{ind}}(\mathbf{Y})$ if block sizes are heterogeneous. In case of homogenous block size, $MLE_{\text{comp}}(\mathbf{Y})$ equals $MLE_{\text{ind}}(\mathbf{Y})$.

³Alternatively, the precision of the maximum independence likelihood estimator can be quantified by referring to the Godambe information instead of the Fisher information (Varin et al. 2011). This practice is theoretically more convincing than referring to an empirically derived correction factor. However, it can lead to estimates of precision that exceed the precision of the maximum likelihood estimator, which does not correspond to the maximal efficiency of the maximum likelihood estimator.

⁴This procedure mimics the approach of Chandler and Bate (2007) but the proposed calibration factor is not based on the Godambe information but on the observed relation between the curvature of the independence likelihood and the covariance of maximum independence likelihood estimator.

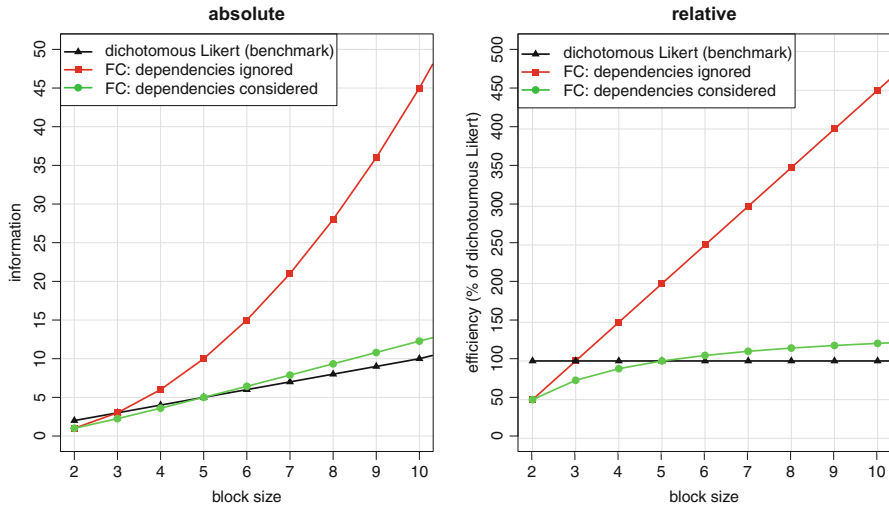


Fig. 1 Declared (red) and actual (green) efficiency of the forced-choice method with false local independence assumption in comparison to binary dichotomous Likert ratings. In the left panel, the ordinate (y-axis) refers to the Fisher information. In the right panel the Fisher information is divided by the Fisher information of the benchmark (dichotomous Likert ratings)

Efficiency⁵ If local independence is assumed, then the amount of information corresponds to the number of binary comparisons implied by the observed ranking, i.e., $n(n - 1)/2$ (Brown and Maydeu-Olivares 2011, cf. the red line in Fig. 1) which would imply that the forced-choice approach leads to efficiency gains if $n > 3$ and efficiency losses if $n = 2$. The proposed calibration factor of the independence likelihood dramatically reduces the expected efficiency of the forced-choice method for greater blocks (cf. the green line in Fig. 1). Nevertheless, the efficiency can still be expected to be a strictly monotonically increasing function of block size that reaches the efficiency of binary Likert ratings if blocks constitute of five items, but efficiency gains never reach 50% even if block size approaches infinity.

4.2 Bayesian Estimation

Genuine Likelihood Let $f(\theta)$ and $f(\theta|Y)$ be the prior and posterior density function of the latent trait θ , respectively, then the maximum a posterior estimator of θ is defined as:

$$MAP(Y) = \operatorname{argmax}_{\theta} f(\theta|Y) = \operatorname{argmax}_{\theta} f(\theta) \mathcal{L}(\theta; Y)$$

⁵The analyses in this paragraph refer to ceteris paribus conditions, i.e., they focus on the effect of block size and disregard the moderating effect of the item parameters.

$$= \operatorname{argmax}_{\boldsymbol{\theta}} \left(\log_e (f(\boldsymbol{\theta})) + \sum_{b=1}^K \ell(\boldsymbol{\theta}; \mathbf{y}_b) \right) \quad (22)$$

whereby the precision might be quantified by the posterior covariance of $\boldsymbol{\theta}$ given \mathbf{Y} that asymptotically equals the inverse of the negative Hessian matrix at $MAP(\mathbf{Y})$:

$$\operatorname{cov}(\boldsymbol{\theta}|\mathbf{Y}) \simeq \left(-\frac{\partial^2}{\partial \boldsymbol{\theta}} f(MAP(\mathbf{Y})|\mathbf{Y}) \right)^{-1} \quad (23)$$

If $\ell(\boldsymbol{\theta}; \mathbf{y}_b)$ is determined by the methods of Genz (2004) or Miwa et al. (2003) as outlined in the explanation of Eq. (11), it is straightforward to determine $MAP(\mathbf{Y})$ and $\left(-\frac{\partial^2}{\partial \boldsymbol{\theta}} f(MAP(\mathbf{Y})|\mathbf{Y}) \right)^{-1}$ by standard numerical optimization methods.

Independence Likelihood Neglecting local dependencies as suggested by Brown and Maydeu-Olivares (2011, 2013, 2018) leads to

$$\begin{aligned} MAP_{\text{ind}}(\mathbf{Y}) &= \operatorname{argmax}_{\boldsymbol{\theta}} \left(f(\boldsymbol{\theta}) \prod_{b=1}^K \mathcal{L}_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_b) \right) \\ &= \operatorname{argmax}_{\boldsymbol{\theta}} \left(\log_e (f(\boldsymbol{\theta})) + \sum_{b=1}^K \ell_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_b) \right) \end{aligned} \quad (24)$$

However, in contrast to maximum likelihood estimation (with respect to the independence likelihood), the falsely precise inference due to the misspecification of the likelihood affects not only the estimated precision but also the value (i.e., the point estimate) of the above-mentioned Bayesian trait estimator, as too much weight is given to the (independence) likelihood in relation to the prior distribution (cf. Pauli et al. 2011).

Composite Likelihood This problem might be amended by referring to the composite likelihood that results from the calibration of the independence likelihood of the respective blocks:

$$\begin{aligned} MAP_{\text{comp}}(\mathbf{Y}) &= \operatorname{argmax}_{\boldsymbol{\theta}} (f(\boldsymbol{\theta}) \cdot \mathcal{L}_{\text{comp}}(\boldsymbol{\theta}; \mathbf{y}_b)) \\ &= \operatorname{argmax}_{\boldsymbol{\theta}} \left(\log_e (f(\boldsymbol{\theta})) + 3 \sum_{b=1}^K \frac{\ell_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_b)}{n_b + 1} \right) \end{aligned} \quad (25)$$

Unlike the respective maximum likelihood estimators, $MAP_{\text{comp}}(\mathbf{Y})$ does not equal $MAP_{\text{ind}}(\mathbf{Y})$ even if block size is homogenous (unless $n = 2$). However, simulations show that $MAP_{\text{comp}}(\mathbf{Y}) \approx MAP(\mathbf{Y}) \not\approx MAP_{\text{ind}}(\mathbf{Y})$ and the precision of $MAP_{\text{comp}}(\mathbf{Y})$ can be quantified by:

$$\left(-\frac{\partial^2}{\partial \boldsymbol{\theta}} f(MAP_{\text{comp}}(\mathbf{Y} | \mathbf{Y}))\right)^{-1} \approx \left(-\frac{\partial^2}{\partial \boldsymbol{\theta}} f(MAP(\mathbf{Y} | \mathbf{Y}))\right)^{-1} \simeq \text{cov}(\boldsymbol{\theta} | \mathbf{Y}) \tag{26}$$

which asymptotically equals the posterior variance of the latent trait.

4.3 Weighted Likelihood Estimation

Genuine Likelihood Maximum Likelihood estimators are often biased. Warm (1989) introduced a strategy to prevent this bias by weighting the likelihood function. Wang (2015) showed how this approach can be applied to multidimensional IRT models and outlined that it is equivalent to Bayesian modal estimation with Jeffrey’s prior for the multivariate 2-pl model. Strictly speaking, this finding cannot be applied for normal-ogive Thurstonian IRT models of forced-choice data as they do not belong to the exponential family (Firth 1993). Nevertheless, due to the similarity between logistic and normal ogive models, Bayesian modal estimation of the latent trait with Jeffrey’s prior might result in an estimator with negligible bias in this setting, too:

$$\begin{aligned} WLE(\mathbf{Y}) &:= \operatorname{argmax}_{\boldsymbol{\theta}} \left(\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) \sqrt{|E(\mathbf{I}(\boldsymbol{\theta}; \mathbf{Y}))|} \right) \\ &= \operatorname{argmax}_{\boldsymbol{\theta}} \left(\frac{\log \left(\left| \sum_{b=1}^K E(\mathbf{I}(\boldsymbol{\theta}; \mathbf{y}_b)) \right| \right)}{2} + \ell(\boldsymbol{\theta}; \mathbf{y}_b) \right) \end{aligned} \tag{27}$$

Composite Likelihood However, the computational burden of $WLE(\mathbf{Y})$ is extensive and often prohibitive in practice, in particular for simulation studies. Capitalizing on the fact that the curvature of the composite likelihood closely resembles the genuine likelihood enables the following weighted maximum composite likelihood estimator of the latent trait which is computationally by far less demanding:

$$\begin{aligned} WLE_{\text{comp}}(\mathbf{Y}) &:= \operatorname{argmax}_{\boldsymbol{\theta}} \left(\mathcal{L}_{\text{comp}}(\boldsymbol{\theta}; \mathbf{Y}) \sqrt{\left| -E \left(\frac{\partial^2}{\partial \boldsymbol{\theta}} \log(\mathcal{L}_{\text{comp}}(\boldsymbol{\theta}; \mathbf{Y})) \right) \right|} \right) \\ &= \operatorname{argmax}_{\boldsymbol{\theta}} \left(3 \cdot \sum_{b=1}^K \frac{\ell_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_b)}{n_b + 1} + \frac{\log \left| \sum_{b=1}^K E \left(-\frac{\partial^2}{\partial \boldsymbol{\theta}} \frac{3 \cdot \ell_{\text{ind}}(\boldsymbol{\theta}; \mathbf{y}_b)}{n_b + 1} \right) \right|}{2} \right) \end{aligned} \tag{28}$$

The negative inverse of the Hessian matrix of the weighted composite likelihood at its maximum is a good candidate for quantifying the precision of $WLE_{\text{comp}}(\mathbf{Y})$.

5 Summary and Conclusions

The present paper introduces person parameter estimates for Thurstonian IRT models for forced-choice ranking data that rely on the genuine likelihood, whereby local dependencies of binary comparisons within blocks are adequately considered. Moreover, a theoretical foundation for the practice of neglecting local dependencies for person parameter estimation was offered by referring to composite likelihood methods. This approach explains the good parameter recovery as well as the problems with quantifying the precision of the estimates when local dependencies are ignored. Within the composite likelihood approach, robust estimates of precision are usually derived by referring to the Godambe information (instead of the Fisher information). However, this practice often leads to estimates of precision that exceed the estimated precision of the maximum likelihood estimator which can be attributed to the relative inefficiency of precision estimates based on the Godambe information (Kauermann and Carroll 2001). Due to these reasons, the calibration of the likelihood by the empirically derived correction factor seems to be superior to the adjustments suggested by Chandler and Bate (2007), unless the number of blocks is large.

In case of Bayesian estimation, the falsely precise inference due to referring to the independence likelihood is not limited to precision estimates but affects the trait estimates, too. As too much weight is given to the data in relation to the prior, the respective estimates are shifted towards the maximum likelihood estimator. Referring to the calibrated composite likelihood yields Bayesian trait estimates that closely resemble Bayesian estimates derived by means of the genuine likelihood. Nevertheless, in real-live applications, estimation should generally be based on the genuine likelihood, unless the need of computational resources poses a formidable obstacle, e.g., weighted likelihood estimation or extensive simulation studies.

References

- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*, 460–502.
- Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*, 36–52.
- Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling, 25*, 516–529.
- Chandler, R. E., & Bate, S. (2007). Inference for clustered data using the independence log-likelihood. *Biometrika, 94*, 167–183.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*, 27–38.

- Genz, A. (2004). Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, *14*, 251–260.
- Kauermann, G., & Carroll, R. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, *96*, 1387–1396.
- Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*, 220–239.
- Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, *64*, 325–340.
- Miwa, T., Hayter, A., & Kuriki, S. (2003). The evaluation of general non-centred orthant probabilities. *Journal of The Royal Statistical Society Series B—Statistical Methodology*, *65*, 223–234.
- Pauli, F., Racugno, W., & Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, *21*, 149–164.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
- Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*, 5–42.
- Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, *80*, 428–449.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.
- Yousfi, S. (2018). Considering local item dependencies: Person parameter estimation IRT models of forced-choice data. In M. Wieberg, S. Culpepper, R. Janssen, J. Gonzales, & D. Molenaar (Eds.), *Quantitative psychology*. Basel: Springer.

An Extended Item Response Tree Model for Wording Effects in Mixed-Format Scales



Yi-Jhen Wu and Kuan-Yu Jin

Abstract Likert scales are frequently used in social science research to measure an individual's attitude, opinion, or perception. Recently, item response tree (IRTree) models have been proposed to analyze Likert-scale data because they could provide insights into an individual's response process. A Likert-scale survey is often mixed with positively worded and negatively worded items, which might induce wording effects. Therefore, it is of interest to investigate how wording effects function in an IRTree model. In this study, we propose a new model—the bi-factor IRTree (BF-IRTree) model, in which combines an IRTree model and a bi-factor model in an IRT framework—to identify how wording effects influence response processes for negatively worded items. Twelve items of an extroversion construct from the Big Five personality inventory were used for demonstration. Results showed that the wording effects were varied on these negatively worded items.

Keywords Item response theory · IRTree · Multi-process · Wording effects · Bi-factor

1 Introduction

A multi-process item response theory tree (IRTree) provides insights into an individual's responding processes, thereby separating and quantifying response processes under an IRT framework (Böckenholt 2012, 2017; De Boeck and Partchev 2012; Thissen-Roe and Thissen 2013). Recently, IRTree models have drawn researchers' attention to investigate response styles (Khorramdel et al. 2019; Zettler et al. 2015).

Y.-J. Wu (✉)

Department of Medical Psychology and Medical Sociology, University of Göttingen, Göttingen, Germany

K.-Y. Jin

Assessment and Technology and Research Division, Hong Kong Examinations and Assessment Authority, Hong Kong, Hong Kong

e-mail: kyjin@hkeaa.edu.hk

However, little is known about how response processes function for negatively worded (NW) items relative to positively worded (PW) items. Given that NW items are frequently designed in a survey to reduce response bias, it is worth investigating what the extent of wording effects of NW items is relative to PW items in each response process in the IRTree framework. Therefore, in the current study, we propose a general model in which a bi-factor model is integrated into the IRTree approach, to investigate an underlying mechanism of NW items in each response process using empirical data.

1.1 IRTree

IRTree models have recently been drawing a lot of attention from IRT researchers. The reason of why IRTree models become gradually popular is that it combines IRT and cognitive psychology theories to explain the underlying response processes under a tree-like structure (Böckenholt 2012). Such a tree-like structure is helpful for researchers to investigate potential sequential processes behind observed responses. Currently, there are different structures of IRTree models. However, given that response processes are rather complex and could be influenced by other factors, there is no clear agreement about response processes in IRTree models. In this study, we adopted a three-step IRTree model from Böckenholt (2017) to demonstrate how wording effects influence each process in a five-point Likert scale. Furthermore, in order to be consistent with our empirical data, the three-step IRTree model seemed to be appropriate. As illustrated in Fig. 1, three response processes may exist, namely indifference (i.e., Process I), direction (i.e., Process II), and the intensity of attitude (i.e., Process III). Each branch is attached to two probabilities to lead to another process or to stop at the current process. In Process I, if individuals do not have clear attitudes or refuse to indicate their attitudes, they will endorse the midpoint category (i.e., 2) with probability P^I ; otherwise, they will move to Process II with probability $1 - P^I$. In Process II, individuals need to decide if they agree with an item (i.e., agreement categories, 3 or 4) with probability P^{II} or if they disagree with an item (i.e., disagreement categories, 0 or 1) with probability $1 - P^{II}$. After Process II, individuals move to Process III. In Process III, individuals decide whether they will endorse the extreme category (0 or 4) with probability P^{III} or the moderate category (1 or 3) with probability $1 - P^{III}$. The three response processes are assumed to be locally independent.

An original response can be decomposed into three sets of binary pseudo items (BPI) (Böckenholt 2012, 2017), and each BPI can be modeled by an IRT model:

$$\log \left(\frac{P_{ij}^d}{1 - P_{ij}^d} \right) = \alpha_j^d \theta_i^d - \delta_j^d, \quad (1)$$

Fig. 1 The three-step IRTree model for an item with a five-point Likert scale (scored from 0 to 4)

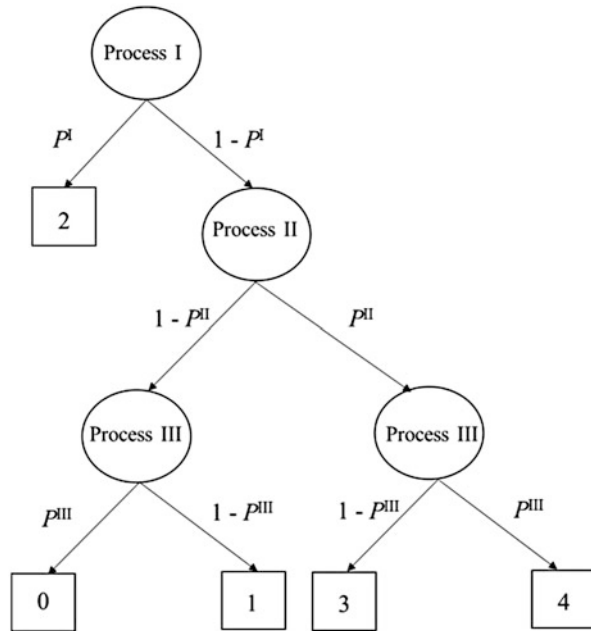


Table 1 Pseudo-items of the five-point response scale in the IRTree model

Original response	BPI coding			Probability
	I	II	III	
0	0	0	1	$\frac{1}{1+\exp(\alpha_j^I \theta_i^I - \delta_j^I)} \times \frac{1}{1+\exp(\alpha_j^{II} \theta_i^{II} - \delta_j^{II})} \times \frac{\exp(\alpha_j^{III} \theta_i^{III} - \delta_j^{III})}{1+\exp(\alpha_j^{III} \theta_i^{III} - \delta_j^{III})}$
1	0	0	0	$\frac{1}{1+\exp(\alpha_j^I \theta_i^I - \delta_j^I)} \times \frac{1}{1+\exp(\alpha_j^{II} \theta_i^{II} - \delta_j^{II})} \times \frac{1}{1+\exp(\alpha_j^{III} \theta_i^{III} - \delta_j^{III})}$
2	1	-	-	$\frac{\exp(\alpha_j^I \theta_i^I - \delta_j^I)}{1+\exp(\alpha_j^I \theta_i^I - \delta_j^I)}$
3	0	1	0	$\frac{1}{1+\exp(\alpha_j^I \theta_i^I - \delta_j^I)} \times \frac{\exp(\alpha_j^{II} \theta_i^{II} - \delta_j^{II})}{1+\exp(\alpha_j^{II} \theta_i^{II} - \delta_j^{II})} \times \frac{1}{1+\exp(\alpha_j^{III} \theta_i^{III} - \delta_j^{III})}$
4	0	1	1	$\frac{1}{1+\exp(\alpha_j^I \theta_i^I - \delta_j^I)} \times \frac{\exp(\alpha_j^{II} \theta_i^{II} - \delta_j^{II})}{1+\exp(\alpha_j^{II} \theta_i^{II} - \delta_j^{II})} \times \frac{\exp(\alpha_j^{III} \theta_i^{III} - \delta_j^{III})}{1+\exp(\alpha_j^{III} \theta_i^{III} - \delta_j^{III})}$

where d ($=$ I, II, or III) denotes the response process; α_j^I , α_j^{II} , and α_j^{III} are the slope parameters of item j ; θ_i^I , θ_i^{II} , and θ_i^{III} are the tendencies of indifference (i.e., Process I), direction (i.e., Process II), and intensity (i.e., Process III) for person i , respectively; δ_j^I , δ_j^{II} , and δ_j^{III} are the location parameters of item j . Table 1 shows the BPI coding and the response probability for each category in the IRTree model. Note these three latent traits measures in distinct processes could be correlated with each other.

1.2 Wording Effects

Surveys are frequently designed with positively worded (PW) and negatively worded (NW) items, because NW items might reduce response bias in cross-culture research (Baumgartner and Steenkamp 2001; Wong et al. 2003). When PW and NW items are designed in a survey, they generally intend to measure the same latent construct. As a result, when individuals tend to agree with PW items, they would disagree with NW items.

However, previous studies have shown that individuals exhibit different response behavior between PW and NW items (Riley-Tillman et al. 2009; Weems et al. 2003). For example, an individual has higher accuracy to measure their behavior in PW items compared with NW items (Riley-Tillman et al. 2009), or an individual is likely to endorse higher scores on PW items than NW items (Weems et al. 2003). Moreover, it has been shown that a cognitive process for an individual is longer for NW items than PW items (Clark 1976), because individuals need to convert negative wording to positive wording before endorsing a response. Altogether, it seems that individuals might have difficulty understanding NW items to judge their attitudes (Swain et al. 2008). Consequently, NW items might induce an unexpected effect to lead to an inaccurate judgment and a longer cognitive process.

Moreover, the issues of psychometric features of NW have been addressed (Baumgartner and Steenkamp 2001; Wong et al. 2003). First, NW items can reduce reliability and validity (Roszkowski and Soven 2010). Second, NW items may produce an additional trait that researchers do not intend to measure. In order to take into account an additional trait derived from NW items, a bi-factor model has been commonly applied for PW and NW items (Lindwall et al. 2012). In the bi-factor model, a general factor is for all items, and two additional factors are for PW items and NW items, respectively. Consequently, under the bi-factor model, there are three factors. However, Wang et al. (2015) argued that an interpretation of a general factor is problematic, because a general factor could not provide a clear definition of a latent trait that researchers intend to measure. Especially, when all items are PW items, an additional factor for capturing wording effects would not exist. Consequently, it is hard to interpret the meaning of a general factor. Thus, Wang et al. (2015) proposed another bi-factor model in the IRT framework (BF-IRT), in which an additional factor describing the wording effects of NW items relative to PW items was included. Therefore, this approach can provide a logical explanation of the second factor and match an original concept of survey development.

In the BF-IRT model, two latent variables are included: the intended-to-be-measured latent factor θ and the wording effect γ , where θ is measured by PW and NW items and γ is measured by NW items. This model keeps not only one intended-measured latent factor in a test development, but also captures a wording effect for NW items. The BF-IRT model is expressed as:

$$\log \left[\frac{P_{ijk}}{P_{ij(k-1)}} \right] = \alpha_j \theta_i + \beta_j \gamma_i - \delta_{jk}, \quad (2)$$

where P_{ijk} and $P_{ij(k-1)}$ are the probabilities of receiving scores k and scores $k-1$ on item j for person i ; θ_i and γ_i are independent variables referring to the targeted latent trait and the nuisance factor on particular items for the person i ; α_j and β_j are the slope parameters of item j on the two latent variables. Note that researchers could either treat PW or NW items as a reference, depending on the purpose of the test development. Given that the number of PW items is usually more than the number of NW items in practice, PW items were considered as a reference in this study. Thus, β_j was fixed at zero for all PW items. When θ and γ are constrained to follow a standard normal distribution, all α -parameters of PW items and α - and β -parameters of NW items can be freely estimated. NW items with a larger ratio of β_j/α_j index represent a large wording effect relative to PW items. As a result, this indicates that NW items do not exhibit similar features as PW items, even though the responses are reversely rescored.

1.3 Bi-factor IRTree

Given the advantages of the BF-IRT model, we attempted to integrate the BF-IRT model into the IRTree model to understand how wording effects influence underlying response processes. The bi-factor IRTree (abbreviated as BF-IRTree) model is expressed as:

$$\log \left(\frac{P_{ij}^d}{1 - P_{ij}^d} \right) = \alpha_j^d \theta_i^d + \beta_j^d \gamma_i^d - \delta_j^d, \quad (3)$$

The notations in Eq. 3 are similar to those defined for Eq. 2 above. Table 2 shows the response probability of each category for NW items in the BF-IRTree model. Moreover, the three targeted latent traits (i.e., θ) are intercorrelated, but they are independent of a wording effect (i.e., γ) in each process, respectively. These three nuisance factors (γ^I , γ^{II} , and γ^{III}) are independent of each other. Borrowing the concept of wording effects from the BF-IRT model (Wang et al. 2015), the ratio of β_j^d/α_j^d indicates the extent of wording effects for NW items in each response process.

Table 2 Pseudo-items of the five-point response scale in the BF-IRTTree model

Response	Probability
0	$\frac{1}{1+\exp(\alpha_j^I \theta_i^I + \beta_j^I \gamma_j^I - \delta_j^I)} \times \frac{1}{1+\exp(\alpha_j^{II} \theta_i^{II} + \beta_j^{II} \gamma_j^{II} - \delta_j^{II})} \times \frac{\exp(\alpha_j^{III} \theta_i^{III} + \beta_j^{III} \gamma_j^{III} - \delta_j^{III})}{1+\exp(\alpha_j^{III} \theta_i^{III} + \beta_j^{III} \gamma_j^{III} - \delta_j^{III})}$
1	$\frac{1}{1+\exp(\alpha_j^I \theta_i^I + \beta_j^I \gamma_j^I - \delta_j^I)} \times \frac{1}{1+\exp(\alpha_j^{II} \theta_i^{II} + \beta_j^{II} \gamma_j^{II} - \delta_j^{II})} \times \frac{1}{1+\exp(\alpha_j^{III} \theta_i^{III} + \beta_j^{III} \gamma_j^{III} - \delta_j^{III})}$
2	$\frac{\exp(\alpha_j^I \theta_i^I + \beta_j^I \gamma_j^I - \delta_j^I)}{1+\exp(\alpha_j^I \theta_i^I + \beta_j^I \gamma_j^I - \delta_j^I)}$
3	$\frac{1}{1+\exp(\alpha_j^I \theta_i^I + \beta_j^I \gamma_j^I - \delta_j^I)} \times \frac{\exp(\alpha_j^{II} \theta_i^{II} + \beta_j^{II} \gamma_j^{II} - \delta_j^{II})}{1+\exp(\alpha_j^{II} \theta_i^{II} + \beta_j^{II} \gamma_j^{II} - \delta_j^{II})} \times \frac{1}{1+\exp(\alpha_j^{III} \theta_i^{III} + \beta_j^{III} \gamma_j^{III} - \delta_j^{III})}$
4	$\frac{1}{1+\exp(\alpha_j^I \theta_i^I + \beta_j^I \gamma_j^I - \delta_j^I)} \times \frac{\exp(\alpha_j^{II} \theta_i^{II} + \beta_j^{II} \gamma_j^{II} - \delta_j^{II})}{1+\exp(\alpha_j^{II} \theta_i^{II} + \beta_j^{II} \gamma_j^{II} - \delta_j^{II})} \times \frac{\exp(\alpha_j^{III} \theta_i^{III} + \beta_j^{III} \gamma_j^{III} - \delta_j^{III})}{1+\exp(\alpha_j^{III} \theta_i^{III} + \beta_j^{III} \gamma_j^{III} - \delta_j^{III})}$

2 An Empirical Study

2.1 Method

We used empirical data to explore wording effects on NW items in the BF-IRTTree model. We used the construct of extroversion from the Big Five personality inventory, including six PW items and six NW items from the English pilot study of non-cognitive skills of the Programme for the International Assessment of Adult Competencies (PIAAC; OECD 2018). The total sample size was 1442. The 12 items were scored from 0 (strongly disagree) to 4 (strongly agree). Based on Baumgartner, Weijters, and Pieters' classifications (2018), the six NW items belonged to the design of polar opposite items that phrases items with an opposite meaning as the intended-measured construct (i.e., extroversion). The NW items were reversely scored; therefore, higher scores meant relatively extroverted. The freeware WinBUGS (Lunn et al. 2000) was used to fit the IRTTree and the BF-IRTTree to the extroversion data in this study. For model comparison, the deviance information criterion (DIC; Spiegelhalter et al. 2002) and the correlations of residuals between pairs of items were used. The model with a lower DIC and approximately uncorrelated residuals was preferred.

2.2 Results

The DICs for the IRTTree and the BF-IRTTree models were 42,374 and 41,212, respectively, which suggested that the BF-IRTTree model yielded a better model fit. Furthermore, the correlations of residuals in the BF-IRTTree model were generally smaller than those in the IRTTree (Fig. 2). Altogether, we concluded that the BF-IRTTree model fits the data better than the IRTTree model.

In the BF-IRTTree, the estimated correlation between θ^I and θ^{II} was -0.36 ($SE = 0.04$), meaning that extroverts are less likely to endorse the middle response. The estimated correlation between θ^I and θ^{III} was -0.54 ($SE = 0.03$), suggesting

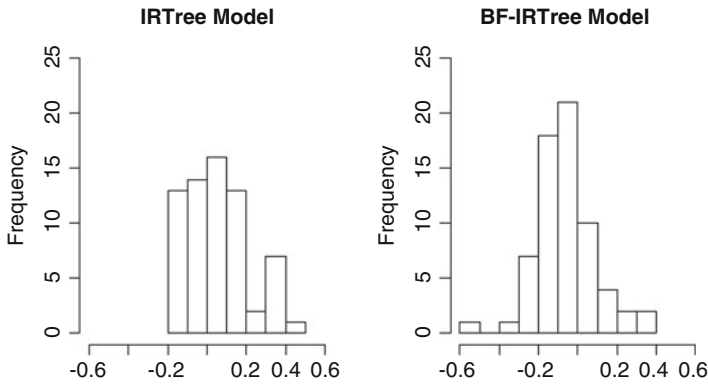


Fig. 2 Correlations among item residuals for the IRTree model and the BF-IRTree model

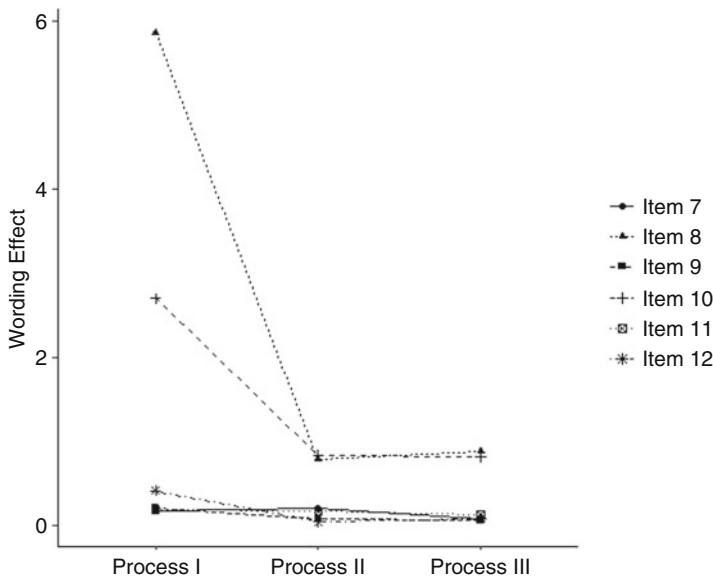


Fig. 3 Wording effects of the six NW items

that individuals who endorse the middle response were less likely to endorse extreme responses. The estimated correlation between θ^{II} and θ^{III} was -0.02 ($SE = 0.03$), indicating that the tendency to use extreme responses is independent of extroversion. Figure 3 shows the wording effects of the six NW items. It demonstrated that items 8 (“Tends to be quiet”) and 10 (“Is sometimes shy, introverted”) had the largest wording effects compared with other NW items. These results suggested that the two items should not be considered to have similar features as PW items after reversing scores.

3 Discussion

In this study, we developed the BF-IRTree model to understand how wording effects occur in underlying response processes. In the analysis of the PIAAC data, the BF-IRTree model yielded a better fit than the standard IRTree model, suggesting that it is necessary to consider the wording effects in the IRTree model when a survey has PW and NW items. We found that items 8 and 10 had pronounced wording effects in the three processes. This might be because people did not consider that the characteristics of item 8 and item 10 could directly correspond to the counterpart of extroversion. Thus, the findings suggest that the features of NW items might display different wording effects in each response process under the IRTree framework.

This study is not free of limitations. First, we only applied five-point Likert scales to demonstrate the performance of the BF-IRTree model. However, the BF-IRTree model is not limited to five-point Likert scales and the three-step IRTree model. In order for the BF-IRTree model to be applied in any condition, future studies should consider an even number of categories to obtain a better understanding of the wording effects in the BF-IRTree model. Furthermore, given that there are different types of IRTree models in the literature (Jeon and De Boeck 2016; Thissen-Roe and Thissen 2013), future studies should combine the BF-IRT approach with other types of IRTree models. Second, modeling item responses and response time jointly has become popular in recent years (Glas and van der Linden 2010; Molenaar et al. 2015). For example, it might be interesting to include response time as a predictor into the BF-IRTree model to investigate whether respondents spend more time endorsing an option, especially for NW items. Moreover, researchers could take the features of NW items into account to explain wording effects and examine whether there is an interaction between the features of NW items and response time.

References

- Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38, 143–156. <https://doi.org/10.1509/jmkr.38.2.143.18840>.
- Baumgartner, H., Weijters, B., & Pieters, R. (2018). Misresponse to survey questions: A conceptual framework and empirical test of the effects of reversals, negations, and polar opposite core concepts. *Journal of Marketing Research*, 55, 869–883. <https://doi.org/10.1509/jmr.15.0117>.
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, 17, 665–678. <https://doi.org/10.1037/a0028111>.
- Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, 22, 69–83. <https://doi.org/10.1037/met0000106>.
- Clark, H. H. (1976). *Semantics and comprehension*. The Hague: Mouton.
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48, 1–28. <https://doi.org/10.18637/jss.v048.c01>.
- Glas, C. A., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, 63, 603–626. <https://doi.org/10.1348/000711009X481360>.

- Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, 48, 1070–1085. <https://doi.org/10.3758/s13428-015-0631-y>.
- Khorramdel, L., von Davier, M., & Pokropek, A. (2019). Combining mixture distribution and multidimensional IRTree models for the measurement of extreme response styles. *British Journal of Mathematical and Statistical Psychology*, 538–559. <https://doi.org/10.1111/bmsp.12179>.
- Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., & Raudsepp, L. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, 94, 196–204. <https://doi.org/10.1080/00223891.2011.645936>.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337. <https://doi.org/10.1023/A:1008929526011>.
- Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, 68, 197–219. <https://doi.org/10.1111/bmsp.12042>.
- OECD. (2018). *Programme for the International Assessment of Adult Competencies (PIAAC), English pilot study on non-cognitive skills*. Data file version 1.0.0 [ZA6940]. Cologne: GESIS Data Archive. <https://doi.org/10.4232/1.13062>.
- Riley-Tillman, T. C., Chafouleas, S. M., Christ, T., Briesch, A. M., & LeBel, T. J. (2009). The impact of item wording and behavioral specificity on the accuracy of direct behavior ratings (DBRs). *School Psychology Quarterly*, 24, 1–12. <https://doi.org/10.1037/a0015248>.
- Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, 35, 117–134. <https://doi.org/10.1080/02602930802618344>.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639. <https://doi.org/10.1111/1467-9868.00353>.
- Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, 45, 116–131. <https://doi.org/10.1509/jmkr.45.1.116>.
- Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, 38, 522–547. <https://doi.org/10.3102/1076998613481500>.
- Wang, W.-C., Chen, H.-F., & Jin, K.-Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, 75, 157–178. <https://doi.org/10.1177/0013164414528209>.
- Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education*, 28, 587–607. <https://doi.org/10.1080/0260293032000130234>.
- Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, 30, 72–91. <https://doi.org/10.1086/374697>.
- Zettler, I., Lang, J. W., Hülshager, U. R., & Hilbig, B. E. (2015). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self-and observer reports. *Journal of Personality*, 84, 461–472. <https://doi.org/10.1111/jopy.12172>.

The Four-Parameter Normal Ogive Model with Response Times



Yang Du and Justin L. Kern

Abstract In recent years, interest in the four-parameter logistic (4PL) model (Barton and Lord, *ETS Res Rep Ser* 198(1):i-8, 1981), and its normal ogive equivalent, has been renewed (Culpepper, *Psychometrika*, 81(4):1142–1163, 2016; Feuerstahler and Waller (*Multivar Behav Res* 49(3):285–285, 2014)). The defining feature of this model is the inclusion of an upper asymptote parameter, in addition to those included in the more common three-parameter logistic (3PL) model. The use of the slipping parameter has come into contact with many assessment applications, such as high-stakes testing (Loken and Rulison, *Br J Math Stat Psychol* 63(3):509–525, 2010), low-stakes testing (Culpepper, *Psychometrika*, 81(4):1142–1163, 2016), and measuring psychopathology (Waller and Reise, *Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory*, 2010). Yet as mentioned in Culpepper (*Psychometrika*, 81(4):1142–1163, 2016), the recovery of the slipping parameter also requires larger sample sizes and longer iterations for the sampling algorithm to converge. Response time (RT), which has already been widely utilized to study student behaviors, such as rapid-guessing, was included in our model to help recover the slipping parameter and the overall measurement accuracy. Based on the hierarchical framework of response and RT (van der Linden, *Psychometrika* 72(3):287–308, 2007), we extended the four-parameter normal ogive model by incorporating RT into the model formulation. A Gibbs sampling approach to estimation was developed and investigated.

Keywords Four-parameter normal ogive model · Response times · Slipping · Hierarchical framework of response and response time

Y. Du (✉) · J. L. Kern
University of Illinois, Urbana-Champaign, Champaign, IL, USA
e-mail: yangd2@illinois.edu; kern4@illinois.edu

1 Introduction

Since Barton and Lord (1981) first investigated the possible upper asymptote for the three-parameter logistic model (3PL), there has been considerable growth in the studies of the four-parameter logistic model (4PL; see Eq. 1) and its normal ogive equivalent (4PNO) (Culpepper 2016; Loken and Rulison 2010; Rulison and Loken 2009; Liao et al. 2012).

$$P(Y_{ij} = 1|\theta_i, a_j, b_j, c_j, d_j) = c_j + \frac{d_j - c_j}{1 + \exp[-1.7a_j(\theta_i - b_j)]} \quad (1)$$

As opposed to the 3PL model, the 4PL and 4PNO models posit the existence of an upper asymptote, d_j ($0 \leq d_j < 1$), to the item response function addressing possible slipping behavior by examinees. Accordingly, the probability of answering an item correctly may never reach one.

While the 4PNO model has been widely applied to cognitive and non-cognitive tests (Loken and Rulison 2010; Culpepper 2016; Waller and Reise 2010), its application should be carefully handled given the relatively poor recovery of the slipping parameters (Culpepper 2016). However, with the aid of computerized test delivery software, apart from responses, response time (RT) data have also been taken into account to improve measurement accuracy and to resolve traditional psychometric problems, such as item selection as well as aberrant student behavior detection (Fan et al. 2012; Du et al. 2019; Choe et al. 2018; Wang and Xu 2015; van der Linden et al. 2007). In a similar vein, we hypothesize that RT may also contain useful information about person slipping behavior. For instance, if a student spent much less RT than their expected RT on an item, they may get a wrong response (i.e., slipped) on that item. In this study, hence, we investigate whether we can improve the parameter recovery of the four-parameter model by adding response time to the model. We propose both an extension of the hierarchical framework of response and RT (van der Linden 2007) using the 4PNO model and a Gibbs sampler for the model estimation.

The rest of this manuscript is organized as follows: we first briefly present the literature review of 4PNO and RT. Next, our model and the full conditional distributions for our posterior are provided. Then, a simulation study is done to investigate estimation accuracy. Finally, we will end our manuscript with the simulation results and conclusions.

2 Literature Review

With the recently renewed interest in the four-parameter models, studies have successfully applied the model to high-stakes tests (Loken and Rulison 2010), low-stakes tests (Culpepper 2016), and psychopathology measurement (Waller and

Reise 2010). Moreover, multiple studies have used the 4PL to address the slipping behavior in early stages of computerized adaptive tests (Liao et al. 2012; Rulison and Loken 2009).

Building upon the estimation techniques developed for the logistic and normal ogive models (Albert 1992; Béguin and Glas 2001; Patz and Junker 1999a,b), estimation methods for the four-parameter model include Metropolis-Hastings (MH) algorithms (Loken and Rulison 2010), Gibbs sampling (Culpepper 2016), and marginal maximum likelihood (MML) (Feuerstahler and Waller 2014). While the MH and MML methods entail tuning the proposal distribution or incorporating informative priors to satisfy the model identification condition, respectively, the Gibbs sampler is often more efficient. However, according to Culpepper (2016), the Gibbs sampler also requires more iterations for the model to converge and larger sample size to accurately recover all parameters.

Among all the existing RT models, the lognormal model proposed by van der Linden (2006) has become popular due to its feasibility and simplicity. The hierarchical framework proposed by van der Linden (2007) further provided plug-and-play framework to jointly model response and RT. For instance, based on this framework, Wang et al. (2013) proposed a semiparametric model that incorporated 3PL with the Cox PH model (Cox 1972) to model response and RT. To date, however, there is no investigation of the viability of 4PNO in the hierarchical model for responses and response times.

3 Model Specification

Relying on van der Linden’s (2007) hierarchical framework, our model formulation is presented in Fig. 1, where the 4PNO model maps out the probability of responses (i.e., Y_{ij}), while the lognormal RT model captures the probability of RT (i.e., T_{ij}).

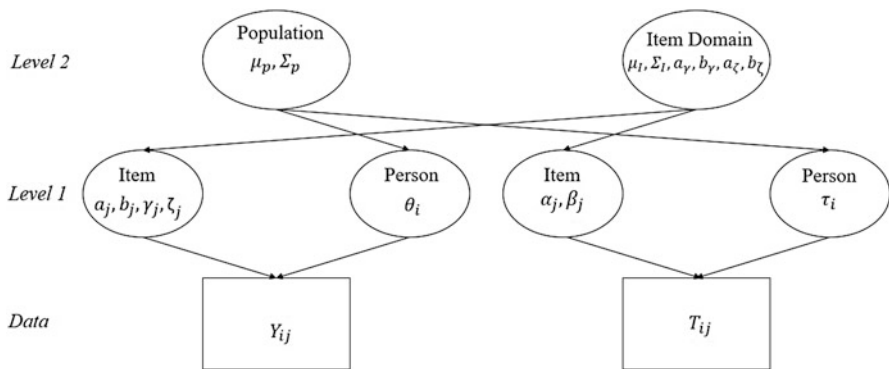


Fig. 1 Hierarchical framework of response and response time

Specifically, the 4PNO (Barton and Lord 1981) is given as

$$P(Y_{ij} = 1 | \eta_{ij}, \gamma_j, \zeta_j) = \gamma_j + (1 - \zeta_j - \gamma_j)\Phi(\eta_{ij}) \quad (2)$$

where $\eta_{ij} = a_j\theta_i - b_j$; $\Phi(\cdot)$ is the cumulative standard normal distribution and a_j , b_j , γ_j , and $1 - \zeta_j$ denote the j th item's discrimination, threshold, lower asymptote, and upper asymptote, respectively. Additionally, this probability is only valid when $0 \leq \gamma_j < 1$, $0 \leq \zeta_j < 1$, and $0 \leq \gamma_j + \zeta_j < 1$. The RT model (van der Linden 2006) is given by Eq. 3,

$$f(T_{ij} \leq t_{ij} | \alpha_j, \beta_j, \tau_i) = \frac{\alpha_j}{\sqrt{2\pi}} \exp - \frac{1}{2} \left(\alpha_j [t_{ij} - (\beta_j - \tau_i)] \right)^2 \quad (3)$$

where α_j and β_j denote the j th item's time discrimination and time intensity parameters, respectively, and τ_i is the latent speed parameter for the i th person.

4 Posterior Distribution

To approximate the posterior distribution of the current model, we introduced two augmented variables, Z and W , denoting the augmented continuous and dichotomous variables, respectively. W arises because of categorizing continuous Z . This is shown as follows:

$$p(\mathbf{Z}, \mathbf{W}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\tau}, \boldsymbol{\psi} | \mathbf{Y}, \mathbf{T}) \propto p(\mathbf{Y} | \mathbf{W}, \boldsymbol{\gamma}, \boldsymbol{\zeta}) p(\mathbf{Z}, \mathbf{W} | \boldsymbol{\theta}, \boldsymbol{\xi}) p(\mathbf{T} | \boldsymbol{\psi}, \boldsymbol{\tau}) p(\boldsymbol{\gamma}, \boldsymbol{\zeta}) p(\boldsymbol{\theta}, \boldsymbol{\tau}) p(\boldsymbol{\psi}, \boldsymbol{\xi}). \quad (4)$$

Here, $\boldsymbol{\xi} = (\mathbf{a}, \mathbf{b})$ denotes the discrimination and threshold parameters, and $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ denotes the time discrimination and time intensity parameters. The likelihoods in the model are

$$p(\mathbf{Y} | \mathbf{W}, \boldsymbol{\gamma}, \boldsymbol{\zeta}) = \prod_{i=1}^I \prod_{j=1}^J \left[(1 - \zeta_j)^{W_{ij}} \gamma_j^{1-W_{ij}} \right]^{Y_{ij}} \left[(1 - \gamma_j)^{1-W_{ij}} \zeta_j^{W_{ij}} \right]^{1-Y_{ij}} \quad (5)$$

$$p(\mathbf{Z}, \mathbf{W} | \boldsymbol{\theta}, \boldsymbol{\xi}) = \prod_{i=1}^I \prod_{j=1}^J \left\{ \phi(Z_{ij}; \eta_{ij}, 1) \times [I(Z_{ij} \leq 0)I(W_{ij} = 0) + I(Z_{ij} > 0)I(W_{ij} = 1)] \right\} \quad (6)$$

$$p(\mathbf{T}|\boldsymbol{\psi}, \boldsymbol{\tau}) = \prod_{i=1}^I \prod_{j=1}^J \phi\left(\log t_{ij}; \beta_j - \tau_i, \frac{1}{\alpha_j^2}\right) \quad (7)$$

where ϕ denotes the normal density. In terms of the priors, due to the constraints that $0 \leq \gamma_j < 1$, $0 \leq \zeta_j < 1$, and $0 \leq \gamma_j + \zeta_j < 1$, the guessing and slipping parameters are naturally dependent, and thus we chose a joint beta distribution as their prior (shown in Eq. 8), corresponding to $p(\boldsymbol{\gamma}, \boldsymbol{\zeta})$ in Eq. 4. For the prior distributions of person parameters and the other four item parameters, we chose bivariate and multivariate normal distributions, respectively. Their means, covariances, and the corresponding hyperpriors are specified in Eqs. 9 and 10, corresponding to $p(\boldsymbol{\theta}, \boldsymbol{\tau})$ and $p(\boldsymbol{\psi}, \boldsymbol{\xi})$ in Eq. 4, respectively.

$$(\gamma_j, \zeta_j) \propto \gamma_j^{a_\gamma-1} (1 - \gamma_j)^{b_\gamma-1} \zeta_j^{a_\zeta-1} (1 - \zeta_j)^{b_\zeta-1} \quad (8)$$

$$\boldsymbol{\Sigma}_p \sim \text{Inv-Wishart}(\boldsymbol{\Sigma}_{p_0}^{-1}, v_{p_0}), \boldsymbol{\mu}_p | \boldsymbol{\Sigma}_p \sim \text{MVN}(\boldsymbol{\mu}_{p_0}, \boldsymbol{\Sigma}_p / \kappa_{p_0}) \quad (9)$$

$$\boldsymbol{\Sigma}_I \sim \text{Inv-Wishart}(\boldsymbol{\Sigma}_{I_0}^{-1}, v_{I_0}), \boldsymbol{\mu}_I | \boldsymbol{\Sigma}_I \sim \text{MVN}(\boldsymbol{\mu}_{I_0}, \boldsymbol{\Sigma}_I / \kappa_{I_0}) \quad (10)$$

With these in place, the full posterior distribution of our model can be approximated. The full conditional distributions used in the proposed Gibbs sampler are presented in the Appendix.

5 Simulation Design and Model Evaluation

To evaluate our model, we generated 5000 students and 20 items from multivariate normal distribution whose means and covariances are shown below. Additionally, a_j and α_j are constrained to be positive, and γ and ζ are sampled from beta distributions with means of $0.2 = 2/(2 + 8)$. The simulation was replicated 20 times.

$$\boldsymbol{\mu}_p = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}, \boldsymbol{\Sigma}_p = \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.0 \end{bmatrix} \quad (11)$$

$$\boldsymbol{\mu}_I = \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \\ 0.0 \end{bmatrix}, \boldsymbol{\Sigma}_I = \begin{bmatrix} 1.00 & 0.03 & -0.04 & -0.11 \\ 0.03 & 1.00 & 0.23 & 0.30 \\ -0.04 & 0.23 & 1.00 & 0.18 \\ -0.11 & 0.30 & 0.18 & 1.00 \end{bmatrix} \quad (12)$$

$$\gamma \sim \text{Beta}(2, 8), \zeta | \gamma \sim \text{Beta}(2, 8) \quad (13)$$

In order to evaluate model convergence, we ran a total of five chains with varying initial values, each of which included 100,000 iterations.

Uninformative priors were used in the simulation. $\Sigma_I = I_4$, $\Sigma_p = I_2$, $\mu_I = (1, 0, 1, 0)^T$, $\mu_p = (0, 0)^T$, where I_4 and I_2 are four- and two-dimensional identity matrices. Uniform priors were used for the lower and upper asymptote parameters. In terms of the hyperprior parameters, $\nu_{p0} = 2$, $\kappa_{p0} = 1$, $\nu_{I0} = 4$, and $\kappa_{I0} = 1$. In sampling the time discrimination parameters, α , the proposal distribution variance was set to be 0.001 to achieve a proper acceptance rate, and the Gibbs-within-Gibbs iteration for sampling γ , ζ was set to be 10.

To evaluate both models, we adopted three sets of criteria. First, for model convergence criterion, a value of \hat{R} (Gelman and Rubin 1992; Brooks and Gelman 1998; Gelman et al. 2013) less than 1.1 was adopted. Second, the overall model fit was evaluated via posterior predictive p values (ppp) (Sinharay et al. 2006), shown in (Eq. 14), where the discrepancy statistic is given by Eq. 15.

$$P(D(y^{rep}, \theta) \geq D(y, \theta)|y) = \int_{D(y^{rep}, \theta) \geq D(y, \theta)|y} p(y^{rep}|\theta)p(\theta|y) dy^{rep} d\theta \quad (14)$$

$$D(y, \theta) = odds\ ratio(OR) = \frac{n_{11}n_{00}}{n_{10}n_{01}} \quad (15)$$

Lastly, the parameter estimation accuracy was evaluated by bias, root mean square error (RMSE), posterior standard deviation (SD), and the correlations between true parameters and their posterior means.

6 Results

In the simulation, the 4PNO with RT model successfully converged after 50,000 iterations. All analyses were thus based on the remaining 50,000 iterations. Next, we evaluated the overall model fit of our model, and only 3.7% of the ppp values are extreme, suggesting appropriate model fit.

Finally, we examined the overall parameter estimation of the 4PNO with RT model in the simulation, and the results are shown in Table 1. Note that in this table, the results are averaged across all replications.

Based on the bias and RMSE, it seems that except for the item discrimination parameters, most of the parameters were recovered well with bias and RMSE close to zero.

However, if we take a close look at the correlations between the true parameters and their posterior means, the surprisingly low correlations of the guessing and slipping parameters (γ , ζ) caught our attention. In other words, given that the guessing and slipping parameters are bounded between 0 and 1, the magnitude of their values could be small, and their bias and RMSE may not yield as much information as the other criterion, such as the correlation between population parameters and posterior

Table 1 Overall model parameters estimation accuracy

Parameters	Bias	RMSE	SD	Correlation
θ	0.01	0.50	0.49	0.87
τ	0.00	0.19	0.13	0.99
a	0.22	0.41	0.38	0.82
b	-0.03	0.20	0.22	0.95
γ	0.03	0.07	0.05	0.54
ζ	0.06	0.08	0.06	0.45
α	0.00	0.02	0.02	1.00
β	0.00	0.01	0.02	1.00

means. To examine the parameter recovery more specifically, we summarized the population values, bias, RMSE, and standard deviations of all the item parameters in Table 2. We also plotted the true parameters against their posterior means, shown in Fig. 2.

From Table 2 and Fig. 2, item threshold, time discrimination, and time intensity parameters are recovered well, most of which fall onto the identity line. Consistent with what we mentioned earlier, item discrimination parameters were not recovered well, particularly those less discriminating items. It is manifest that RMSE and bias are higher if the magnitude of item discrimination parameters are less than 1. This can also be seen from Fig. 2. Those lower discriminating parameters were overestimated in our model. Such result was not found in Culpepper (2016) since the item discrimination parameters in his simulation studies are mostly greater than 1 (in his simulations, $a \sim N(2, 0.5)$). While this result is partially in line with what Culpepper (2016) claimed that harder items (i.e., larger value of b/a) may lead to less accurate estimates of ζ , it's also thought-provoking to see that incorporating slipping parameters would potentially make items more discriminating. One reason could be that the result of overestimating the slipping parameters lowers the upper asymptote, thus making fewer students' probability to answer the item correctly reach one and consequently items become more discriminating.

Similar overestimation issues are present in guessing and slipping parameters whose true values of guessing and slipping parameters are smaller than 0.25. Such overestimation might be caused by the uniform prior we set and smaller sample size. As pointed out in Culpepper (2016), sample sizes of at least 2500 and 10,000 are needed for educational items and psychopathology items, respectively, to accurately recover all parameters. When the sample size is small and item difficulty parameters are extremely high or low, few students will have both latent binary variables and actual responses (W and Y) equal to one or zero. In other words, due to the limited valid responses from our data, the posteriors of guessing and slipping parameters will be greatly dominated by the priors. Given the uniform prior we have, for those items whose difficulty parameters are extremely high or low, they will always be overestimated. By increasing the sample sizes, we could potentially increase the number of extremely high or low ability students.

Table 2 Item parameter estimates across 20 replications

Item	True parameters					Bias					RMSE					SD								
	a	b	γ	ζ	α	a	b	γ	ζ	α	a	b	γ	ζ	α	a	b	γ	ζ	α				
1	1.71	-1.30	0.37	0.10	2.43	0.34	0.02	-0.03	-0.05	0.00	0.01	0.00	0.36	0.20	0.09	0.01	0.03	0.01	0.42	0.19	0.10	0.01	0.03	0.02
2	1.29	-0.17	0.14	0.07	1.94	0.85	0.10	0.02	0.01	0.00	0.00	0.00	0.23	0.11	0.04	0.02	0.03	0.01	0.24	0.10	0.04	0.03	0.02	0.02
3	1.68	0.43	0.07	0.08	1.39	0.77	0.10	0.02	0.00	0.00	0.00	0.00	0.22	0.05	0.01	0.02	0.01	0.01	0.26	0.09	0.02	0.03	0.01	0.02
4	0.84	0.28	0.16	0.12	0.62	-0.23	0.16	-0.07	0.00	0.04	0.00	0.01	0.23	0.13	0.03	0.05	0.01	0.02	0.28	0.15	0.05	0.07	0.01	0.03
5	1.05	-0.07	0.19	0.08	1.57	0.45	0.18	-0.02	0.00	0.01	0.00	0.00	0.34	0.13	0.06	0.04	0.01	0.01	0.26	0.12	0.05	0.03	0.02	0.02
6	0.73	-0.36	0.18	0.19	0.38	0.03	0.31	0.07	0.06	0.01	0.00	-0.01	0.44	0.18	0.08	0.04	0.00	0.03	0.35	0.20	0.08	0.05	0.00	0.04
7	0.17	-0.13	0.19	0.07	1.36	0.38	0.77	0.10	0.26	0.20	0.00	0.01	0.82	0.35	0.26	0.20	0.01	0.01	0.58	0.55	0.11	0.09	0.01	0.02
8	0.35	0.35	0.10	0.17	0.66	-0.68	0.61	-0.10	0.09	0.21	0.00	0.01	0.69	0.32	0.10	0.21	0.00	0.02	0.46	0.33	0.06	0.12	0.01	0.03
9	1.38	-0.25	0.13	0.12	2.02	0.06	0.02	-0.02	-0.02	0.00	0.00	0.00	0.23	0.10	0.05	0.03	0.03	0.01	0.27	0.11	0.04	0.03	0.02	0.02
10	2.25	0.12	0.06	0.12	2.05	-0.33	-0.04	0.02	0.00	-0.01	0.00	0.00	0.28	0.06	0.02	0.02	0.02	0.01	0.32	0.08	0.02	0.02	0.02	0.02
11	2.86	0.17	0.09	0.12	1.86	0.91	-0.25	0.00	-0.01	-0.01	0.00	0.00	0.35	0.09	0.01	0.02	0.02	0.01	0.37	0.09	0.01	0.02	0.02	0.02
12	2.09	-0.15	0.30	0.23	1.07	-0.95	-0.19	-0.02	-0.03	-0.02	0.00	0.00	0.40	0.14	0.04	0.03	0.01	0.01	0.46	0.15	0.04	0.03	0.01	0.02
13	0.73	-1.28	0.04	0.14	1.89	0.07	0.57	0.10	0.29	0.01	0.00	0.00	0.64	0.22	0.30	0.02	0.02	0.01	0.44	0.26	0.13	0.02	0.02	0.02
14	0.95	-1.00	0.16	0.27	2.07	-0.27	0.31	-0.01	0.05	0.01	0.00	0.00	0.35	0.16	0.08	0.01	0.02	0.01	0.41	0.25	0.10	0.02	0.02	0.02
15	1.30	-0.01	0.14	0.24	2.05	-0.19	0.18	0.02	0.00	-0.01	0.00	0.00	0.37	0.10	0.03	0.03	0.02	0.01	0.37	0.13	0.04	0.04	0.02	0.02
16	0.54	1.23	0.17	0.11	1.28	-0.07	0.48	-0.38	0.00	0.31	0.00	0.00	0.56	0.53	0.03	0.32	0.01	0.01	0.49	0.38	0.04	0.15	0.01	0.02
17	0.70	0.47	0.36	0.15	2.48	0.56	0.29	-0.11	-0.02	0.05	0.00	0.00	0.39	0.30	0.05	0.06	0.03	0.01	0.42	0.29	0.07	0.08	0.03	0.02
18	0.61	1.63	0.11	0.18	1.10	-0.44	0.57	-0.09	0.01	0.32	0.00	0.00	0.60	0.37	0.01	0.33	0.01	0.01	0.47	0.41	0.01	0.18	0.01	0.02
19	1.24	0.91	0.23	0.37	2.30	1.10	-0.06	-0.03	-0.01	-0.05	-0.01	0.00	0.32	0.24	0.03	0.08	0.02	0.01	0.45	0.27	0.03	0.12	0.03	0.02
20	1.11	0.82	0.23	0.06	2.15	-0.18	0.35	0.01	0.01	0.07	0.01	0.00	0.43	0.14	0.02	0.09	0.02	0.01	0.34	0.17	0.02	0.06	0.02	0.02

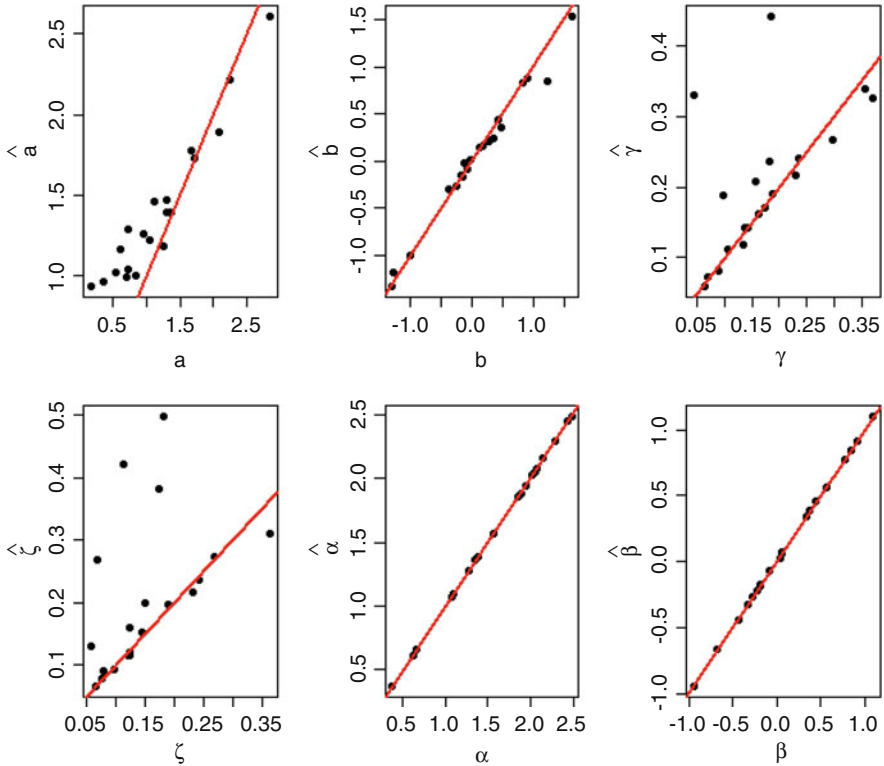


Fig. 2 Item parameters estimation accuracy

To answer our research question that whether including RT could help with item parameter recovery, we compared the item parameter recovery of both 4PNO with and without RT models. The results are shown in Table 3. The item response parameter estimates yielded by both models were very similar. Simply based on the limited simulation conditions in our study, it seems that including RT in the hierarchical framework does not help with item parameter recovery in 4PNO. However, this by no means suggests that RT cannot aid in item parameter recovery at all. Additionally, as the guessing and slipping parameters in our model were not captured in the item parameter covariance structure in the second level, the uniform prior would still play a role when the data do not provide sufficient information. Nevertheless, as an important source of collateral information, RT may help boost the parameter estimate accuracy if they are incorporated in the response accuracy model part. Furthermore, future studies may investigate on utilizing RT to directly model the guessing and slipping parameters in the item response model. Lastly, a broader investigation with a wider array of item parameters may further elucidate the nature of the current estimation approach.

Table 3 Item parameter recovery comparison

Model	Bias				RMSE				SD			
	a	b	γ	ζ	a	b	γ	ζ	a	b	γ	ζ
4PNO with RT	0.22	-0.03	0.03	0.06	0.41	0.20	0.07	0.08	0.38	0.22	0.05	0.06
4PNO	0.10	-0.06	0.01	0.06	0.10	0.06	0.01	0.06	0.35	0.21	0.06	0.06

7 Conclusions

Based on recent developments of the 4PNO (Culpepper 2016) and the hierarchical framework of response and response time (van der Linden 2007), our 4PNO with RT model showed appropriate model fit to our simulated data. The successful convergence of our 4PNO with RT model verifies that the 4PNO model can be incorporated to the hierarchical framework. But based on the limited simulation conditions, our study also suggests that RT included in this hierarchical framework does not assist in recovering the guessing and slipping parameters. Consistent with Culpepper (2016), large sample sizes are still needed in order to accurately recover the guessing and slipping parameters. Future studies may investigate the possibility of including RT in the item response model or employing RT to directly model guessing and slipping parameters in more simulation conditions.

Appendix: Gibbs Sampler

- Step 1: sample W_{ij}

$$W_{ij}|Y_{ij}, \theta_i, \eta_{ij}, \gamma_j, \zeta_j \sim \begin{cases} \text{Bernoulli}\left(\frac{\zeta_j \Phi(\eta_{ij})}{1-\gamma_j-(1-\gamma_j-\zeta_j)\Phi(\eta_{ij})}\right), & Y_{ij} = 0 \\ \text{Bernoulli}\left(\frac{(1-\zeta_j)\Phi(\eta_{ij})}{\gamma_j+(1-\zeta_j-\gamma_j)\Phi(\eta_{ij})}\right), & Y_{ij} = 1 \end{cases}$$

- Step 2: sample Z_{ij}

$$Z_{ij}|\eta_{ij}, \theta_i, W_{ij} \sim \begin{cases} N(\eta_{ij}, 1)I(Z_{ij} \leq 0), & W_{ij} = 0 \\ N(\eta_{ij}, 1)I(Z_{ij} > 0), & W_{ij} = 1 \end{cases}$$

- Step 3: sample θ_i

$$\theta_i|\mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p \sim N\left(\frac{\sigma_{\theta|\tau_i}^{-2}\mu_{\theta|\tau_i} + \sum_{j=1}^J a_j(z_{ij}+b_j)}{\sigma_{\theta|\tau_i}^{-2} + \sum_{j=1}^J a_j^2}, \left(\sigma_{\theta|\tau_i}^{-2} + \sum_{j=1}^J a_j^2\right)^{-1}\right)$$

where

$$\mu_{\theta|\tau_i} = \mu_{\theta} + \frac{\sigma_{\theta\tau}}{\sigma_{\tau}^2}(\tau_i - \mu_{\tau}), \sigma_{\theta|\tau_i}^2 = \sigma_{\theta}^2 - \frac{\sigma_{\theta\tau}^2}{\sigma_{\tau}^2}$$

- Step 4: sample $\xi_j = (a_j, b_j)$

$$\begin{aligned} \xi_j &= (a_j, b_j) | \mathbf{Z}_j, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I \\ &\sim N \left(\frac{\boldsymbol{\mu}_{a,b|\alpha,\beta} \boldsymbol{\Sigma}_{a,b|\alpha,\beta}^{-1} + \mathbf{X}' \mathbf{z}_j}{\boldsymbol{\Sigma}_{a,b|\alpha_j,\beta_j}^{-1} + \mathbf{X}' \mathbf{X}}, \left(\boldsymbol{\Sigma}_{a,b|\alpha_j,\beta_j}^{-1} + \mathbf{X}' \mathbf{X} \right)^{-1} \right) \end{aligned}$$

where $\mathbf{X} = [\boldsymbol{\theta} \ -1]$, $\boldsymbol{\mu}_{a,b|\alpha,\beta}$ and $\boldsymbol{\Sigma}_{a,b|\alpha,\beta}$ follow directly from $\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I$

- Step 5: sample γ_j, ζ_j , based on the Gibbs within Gibbs in Culpepper (2016),

$$f_{\zeta|\gamma} = \frac{f_{\gamma\zeta}}{f_{\gamma}} = \frac{f_{\zeta}}{F_{\zeta(1-\gamma)}} I(0 \leq \zeta \leq 1 - \gamma)$$

- Step 6: sample τ_i

$$\tau_i | \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p \sim N \left(\frac{\sigma_{\tau|\theta_i}^{-2} \mu_{\tau|\theta_i} + \sum_{j=1}^J \alpha_j^2 (\beta_j - t_{ij})}{\sigma_{\tau|\theta_i}^{-2} + \sum_{j=1}^J \alpha_j^2}, \left(\sigma_{\tau|\theta_i}^{-2} + \sum_{j=1}^J \alpha_j^2 \right)^{-1} \right)$$

- Step 7: sample β_j

$$\begin{aligned} \beta_j | \mathbf{t}_i, \boldsymbol{\tau}, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I &\sim N \left(\frac{\sigma_{\beta|a_j,b_j,\alpha_j}^{-2} \mu_{\beta|a_j,b_j,\alpha_j} + \alpha_j^2 \sum_{i=1}^I (t_{ij} + \tau_i)}{\sigma_{\beta|a_j,b_j,\alpha_j}^{-2} + I \alpha_j^2}, \right. \\ &\left. \left(\sigma_{\beta|a_j,b_j,\alpha_j}^{-2} + I \alpha_j^2 \right)^{-1} \right) \end{aligned}$$

- Step 8: sample α_j by Metropolis-Hastings algorithm, the acceptance probability is

$$\min \left(1, \frac{f(\alpha_{jt}^* | t_{ij}, \tau_i, \beta_j) g(\alpha_{j(t-1)}) | \alpha_{jt}^*}{f(\alpha_{j(t-1)}^* | t_{ij}, \tau_i, \beta_j) g(\alpha_{jt}^* | \alpha_{j(t-1)})} \right)$$

– Step 9: sample hyperprior μ_P, Σ_P

$$\begin{aligned} \Sigma_P | \delta &\sim \text{Inverse - Wishart}(\Sigma_{P*}^{-1}, v_{P*}), \quad \mu_P | \Sigma_P, \delta \sim N_2(\mu_{P*}, \Sigma_P / \kappa_{P*}) \\ \Sigma_{P*} &= \Sigma_{P_0} + S_\delta + \frac{I \kappa_{P_0}}{I + \kappa_{P_0}} (\bar{\delta} - \mu_{P_0})(\bar{\delta} - \mu_{P_0})', \quad v_{P*} = v_{P_0} + I, \quad \kappa_{P*} = \kappa_{P_0} + I \\ \mu_{P*} &= \frac{\kappa_{P_0}}{\kappa_{P_0} + I} \mu_{P_0} + \frac{I}{\kappa_{P_0} + I} \bar{\delta}, \quad S_\delta = \sum_{i=1}^I (\delta - \bar{\delta})(\delta - \bar{\delta})' \end{aligned}$$

– Step 10: sample hyperprior μ_I, Σ_I

$$\begin{aligned} \Sigma_I | \mathbf{v} &\sim \text{Inverse - Wishart}(\Sigma_{I*}^{-1}, v_{I*}), \quad \mu_I | \Sigma_I, \mathbf{v} \sim N_4(\mu_{I*}, \Sigma_I / \kappa_{I*}) \\ \Sigma_{I*} &= \Sigma_{I_0} + S_v + \frac{J \kappa_{I_0}}{J + \kappa_{I_0}} (\bar{\mathbf{v}} - \mu_{I_0})(\bar{\mathbf{v}} - \mu_{I_0})' \\ v_{I*} &= v_{I_0} + J, \quad \kappa_{I*} = \kappa_{I_0} + J \\ \mu_{I*} &= \frac{\kappa_{I_0}}{\kappa_{I_0} + J} \mu_{I_0} + \frac{J}{\kappa_{I_0} + J} \bar{\mathbf{v}}, \quad S_v = \sum_{j=1}^J (\mathbf{v} - \bar{\mathbf{v}})(\mathbf{v} - \bar{\mathbf{v}})' \end{aligned}$$

References

- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17(3), 251–269.
- Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66(4), 541–561.
- Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 198(1), i-8.
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455.
- Choe, E. M., Kern, J. L., & Chang, H.-H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 43, 135–158.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202.
- Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, 81(4), 1142–1163.
- Du, Y., Li, A., & Chang, H.-H. (2019). Utilizing response time in on-the-fly multistage adaptive testing. In M. Wiberg, S. A. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology. IMPS 2017. Springer Proceedings in Mathematics & Statistics*. Cham: Springer.
- Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, 37(5), 655–670.
- Feuerstahler, L. M., & Waller, N. G. (2014). Estimation of the 4-parameter model with marginal maximum likelihood. *Multivariate Behavioral Research*, 49(3), 285–285.

- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton: Chapman and Hall/CRC.
- Liao, W. W., Ho, R. G., Yen, Y. C., & Cheng, H. C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: An International Journal*, 40(10), 1679–1694.
- Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63(3), 509–525.
- Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24(4), 342–366.
- Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of educational and behavioral Statistics*, 24(4), 146–178.
- Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, 33(2), 83–101.
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34(3), 213–232.
- Sinharay, S., Johnson, M., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4), 298–321.
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287–308.
- van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, 44(2), 117–130.
- Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. Embretson (Ed.), *Measuring psychological constructs: Advances in model based approaches*. Washington, DC: American Psychological Association.
- 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.
- Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, 38(4), 381–417.

A Bayesian Graphical and Probabilistic Proposal for Bias Analysis



Claudia Ovalle and Danilo Alvares

Abstract One of the main concerns in educational policies is to analyze whether a national test is fair for all students, especially for the economically and socially disadvantaged groups. In the current literature, there are some methodological proposals that analyze this problem through comparative approaches of performance by groups. However, these methodologies do not provide an intuitive graphical and probabilistic interpretation, which would be useful to aid the educational decision-making process. Therefore, the objective of this work is to bridge these gaps through a methodological proposal based on the one-, two-, and three-parameter logistic models, where we evaluate the performance of each group using the difficulty parameter estimated from a Bayesian perspective. The difference between parameters of each group and their respective 95 credible interval are displayed in graphical form. In addition, we also calculate the mean of the posterior probability of all the differences of each parameter for the groups compared. This probabilistic measurement provides a more accurate perception of intergroup performance by analyzing all items together. Our methodology is illustrated with the Chilean University Selection Test (PSU) data of 2018, where the analyzed groups are students from (regular) high schools versus technical high schools. A sensitivity analysis between the two logistic models is presented. All analyzes were performed using the R language with the JAGS program.

Keywords Differential item functioning · Mantel Haenszel · Item analysis · Graphical representation

C. Ovalle (✉)

Centro de Justicia Educacional, PUC, Santiago de Chile, Chile

e-mail: claudia.ovalle@uc.cl

D. Alvares

Departamento de Estadística, PUC, Santiago de Chile, Chile

© Springer Nature Switzerland AG 2020

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 322, https://doi.org/10.1007/978-3-030-43469-4_6

1 Introduction

In the field of education, it is necessary to measure and evaluate the learning of students. However, being able to obtain items that allow measuring different population groups is hindered by the potential bias and by aspects such as differences in the parameters of difficulty, pseudo-chance, and discrimination that may favor a majority group compared to a minority. In this sense, it is necessary to explore the performance of the test items with high-stake consequences for students. One way of doing this is the measurement and graphic representation of the difference between groups of the parameters of difficulty, guessing, and discrimination proposed in the Item Response Theory or IRT. In this theory, student's ability is a latent trait, and the characteristics of the items can be measured and represented graphically (e.g., by means of the item information curve). The present proposal is novel since it is not limited to an interpretation based on the 3PL model, but it incorporates measures of sensitivity integrating the interpretation of the parameters in the 1PL-G, 1PL, 2PL, and 3PL models, which are calculated with a Bayesian approach in the language R with the JAGS program. Likewise, it proposes a new graphic representation that incorporates the difference in the parameters between the groups, to facilitate the detection of biases in the test in a visual way. This graphic representation allows the researcher to observe the same parameter for several items, which is not possible whenever information curves are drawn for each item. The graphic representation provides a more informative analysis of the performance between the groups since all the items are analyzed together and not separately as it is done with the item information curve.

2 IRT Models for Parameter Estimation

The use of the guessing parameter is due to Birnbaum (1968) and corresponds to the asymptote with a value greater than 0 in information curves. The guessing parameter represents the probability of response of an individual with a very low ability. On the other hand, the difficulty refers to the probability that a student responds correctly to a given item with a certain level of ability (San Martín and De Boeck 2015). Finally, discrimination refers to the ability of the item to differentiate which students know the content as opposed to those who do not (Tuerlinckx and De Boeck 2005). This parameter is represented by the slope of the information curve of the item. There are different models to determine the parameters of the items. The first is the logistic model of a parameter with guessing (1PL-G) which is a case of the 3PL model in which the discrimination parameters are set at 1 (San Martín et al. 2006). The second is the Rasch model (Rasch 1960) that focuses on the difficulty parameter. The third is the 2PL model, in which discrimination and difficulty parameters are included. The fourth is the 3-PLG Model (difficulty, discrimination, and guessing) of Birnbaum (1968). All models are presented in Table 1.

Table 1 IRT models

Model	$G(\theta_i, \omega_i)$	Item parameter	Parameter space
1 PL	$F(\theta_i, -\beta_i)$	$\omega_j = \beta_j$	$(\theta_{1i}, \omega_{1j}) \in R^i X R^j$
2 PL	$F(\alpha_i \theta_i, -\beta_i)$	$\omega_j = \alpha_j, \beta_j$	$(\theta_{1i}, \omega_{1j}) \in R^i X R^j X R^j$
1PL-G	$F(\gamma_i + (1 - \gamma_i)F(\alpha_i, -\beta_j))$	$\omega_j = \beta_j, \gamma_j$	$(\theta_{1:i}, \omega_{1:j}) \in R^i X R^j X (0, 1)^j$
3 PL	$F(\gamma_i + (1 - \gamma_i)F(\alpha_i \theta_i, -\beta_j))$	$\omega_j = \alpha_i, \beta_j, \gamma_j$	$(\theta_{1i}, \omega_{1:j}) \in R^i X R^j X R^j X (0, 1)^j$

1PL-G models, in which the discrimination parameter is constant and it is equivalent to 1, are widely used in the literature (Fariña, Gonzales, San Martín 2019; San Martín et al. 2013). The 1PL-G is also preferred since issues are raised when interpreting the parameters in the 3PL model (Maris and Bechger 2009) and the convenience of using binary models (two parameters) under different specifications has been reported (San Martin 2016). In the present study, we opted for the difficulty parameter, since the main objective is to compare two groups (focal and reference groups) so that we can find the differences in the items that affect the minority group (technical high school students vs. the academic track students). For this we focus on the 1PL, 1PL-G, 2PL, and 3PL models. From these four models, a proposal was developed to establish a selection criteria for the items that must integrate a standardized test applicable to different groups of students based on the differences between groups in the difficulty parameter. Specifically, this was done to compare students of the technical high school vs. students of the academic track and thus to reduce the bias in favor of one or the other group. This research provides a novel approach to item bias by means of a graphical representation of the difference in the difficulty parameter.

3 Item Bias

Bias or differential item functioning (DIF) arises when the probability of a correct response between people with the same value of the latent trait (ability) differs between groups, for example, whenever the difficulty of an item depends on the membership to a subgroup based on race, ethnicity, or gender (Berger and Tutz 2016). The present study is focused on uniform bias, that is, when individuals from different subgroups with the same skill level have different probabilities of solving an item and these differences do not depend on their ability. Zwick (2012) reviews the criteria for the detection of biased items and identifies the flagging rules used by ETS (Educational Testing Service). The author concludes that rule “C” is insufficient to establish critical bias of the items even when the samples are large. The rule indicates that an item that has bias must have a χ^2 Mantel-Haenszel Delta, MH D statistic, with an absolute value greater than 1.5 and it must be significant at the 5 percent level. A similar rule indicates that the critical value of the MH

Delta is 1,645 or 95th percentile. If the classification of the ETS bias is used with the MH Delta statistic, which focuses on the difficulty parameter, then an item of category “A” or without bias has a nonsignificant delta value of 1. A category “B” item has a delta between 1 and 1.5 and is significant, and one of category “C” has a delta of at least 1.5 and is significant too. In terms of the parameters of the items, a MH Delta of 1.0 is equivalent to $[-2.35 * \beta^2]$ and therefore indicates the cutoff point for an item to have type B or type C bias in terms of the difficulty parameter. The present proposal is based on a Bayesian approach, and it is centered on the values of the difference of the difficulty parameter for different groups. For example, we compared the difficulty parameter between groups of technical students versus academic track students who took a standardized test. While the no-Bayesian approach seeks to find the values of the (estimated) parameters that maximize the probability of the data that has been observed, the Bayesian approach used in the present study makes use of the prior distributions of the parameters of interest, and the inferences are based on samples of the posterior distributions, which can be used to summarize all the information about the parameters of interest (Gonzalez 2010a,b, p. 2). That is, the probability distribution of the parameter of interest is used.

4 Method

Since the purpose of the present study is to analyze if a standardized test is fair for all students, in particular for those who come from a vocational/technical high school using the measurement of the parameters in the selected IRT models (1PL, 1PL-G, 2PL, 3PL), we will proceed to use a graphical Bayesian interpretation of the differences in the parameters in the tests of mathematics from a national standardized test (80 items from the PSU). The question that guides the present study is: “Does a differential functioning or DIF persists in the items (in the mathematics subtest) that is not due to the ability of students (latent trait) but can be conditional to aspects such as the type of curriculum (academic versus technical/vocational)?”.

4.1 Descriptive Analysis

In the present proposal, the DIF analysis will be developed with the χ^2 Mantel-Haenszel. This is a descriptive analysis that will be developed with the DIFAS software for dichotomous items.

4.2 Bias Analysis Comparing Difficulty Parameter with 1PL, 1PL-G, 2PL, and 3PL Models

The following models were used:

One parameter logistic (1PL) model

$$P(Y_{ij} = 1) = \frac{\exp \{ \theta_i - \beta_j^* \}}{1 + \exp \{ \theta_i - \beta_j^* \}}$$

One parameter logistic with guessing (1PL-G) model

$$P(Y_{ij} = 1) = \gamma_j + (1 - \gamma_j) \frac{\exp \{ \theta_i - \beta_j^* \}}{1 + \exp \{ \theta_i - \beta_j^* \}}$$

Two parameter logistic (2PL) model

$$P(Y_{ij} = 1) = \frac{\exp \{ \alpha_j (\theta_i - \beta_j^*) \}}{1 + \exp \{ \alpha_j (\theta_i - \beta_j^*) \}}$$

Three parameter logistic (3PL) model

$$P(Y_{ij} = 1) = \gamma_j + (1 - \gamma_j) \frac{\exp \{ \alpha_j (\theta_i - \beta_j^*) \}}{1 + \exp \{ \alpha_j (\theta_i - \beta_j^*) \}}$$

All models predict the probability of correct response $Y_{ij} = 1$. The parameter $\beta_{ij}^* = \beta_j + g_i \Delta_j$ represents the difficulty β_j and an interaction term β_j^* which represents the negative or positive increment of the difficulty parameter for the TP (vocational/technical) group compared to the SH (academic) group of students.

The Bayesian approach was used to calculate the item parameters in all IRT models. In the estimation, the following priors were used: The guessing parameter γ_i was represented as a beta distribution (0.5, 0.5), discrimination α_i was represented as a uniform distribution (0,100), and difficulty β_i was represented as a normal distribution (0,100). The person latent ability θ_i was a parameter estimated in all models, and it is distributed as normal(0, $\sigma_{g_i}^2$), where $\sigma_{g_i}^2$ is the standard deviation of the ability parameter for each of the groups g_i of vocational or academic students.

4.3 Graphical Representation

For each parameter, calculated by means of the IRT models (1PL-G, 1PL 2PL, 3PL), a graphical representation was done. The horizontal axis represents all of the test items, and the vertical axis represents the difference in the value of the difficulty parameter for each one of these items. This representation helps to establish if the difference affects the minority group (vocational/technical students) in comparison to a majority group (academic students). The difference between groups in the difficulty parameter is represented for each item by a point in the graph. The graphical representation includes the credibility intervals that indicate the probability that the difference between groups in the parameter (conditional on the data) is greater or less than zero.

4.4 Data

In the year 2017, for the university admission of the year 2018, approximately 295.531 students registered for the PSU standardized test, and 262.139 (89%) took the subtests of language and mathematics (DEMRE, 2017). Among these students, almost 90.000 belonged to the technical/vocational track.

4.5 Sample

We sampled 136.918 students from the academic track and 56.798 students from the technical/vocational track.

5 Results

5.1 Descriptive Statistics

The analysis was done separately for the four equivalent forms of the test of the mathematics PSU test (each form with 80 items, which can be combined and repeated in different ways). With the estimations of DIF, it was established which items have potential bias. In order to detect DIF, the χ^2 MH (chi square of Mantel-Haenszel (MH)) was calculated. We used the R language to establish the model parameters, and we used the software DIFAS 5.0 for the χ^2 MH analysis. In order to establish if an item has DIF, two criteria were used: χ^2 MH has to be significant and the CDR (combined decision rule) should be true. The last rule implies that MH LOR – log odds ratio – is equivalent to a value ranging between 2.0 -2.0 (indicating

that the item has DIF) and LOR Z (negative values indicating DIF is in favor of the minority group and positive values favoring the majority group).

5.2 Graphical Representation

The first graphic representation obtained with the 1PL-G model is presented in Fig. 1. The horizontal axis corresponds to the PSU test items ($n = 180$ items), while the vertical axis corresponds to the difference between groups (technical vs. academic track) for the value of the difficulty parameter.

The vertical axis value is the difference between the medians of the posterior distributions obtained with multiple samples. These samples were obtained using a Bayesian approximation for the calculation of the difficulty parameter. The difference between groups in the difficulty parameter has a range between 3 and -3 (on the vertical axis).

In this sense, the differences that approach 0 indicate that the item is appropriate and it is not presenting an important difference between the two groups of students (technical vs. academic). On the contrary, those items in which the differences between groups are closer to 3 or closer to -3 are the items which are “problematic” since they do not measure the groups in the same way. The items in the upper band (above 0.5) indicate that the difficulty is greater for the technical student group compared to the academic group. According to the representation, the vast majority of the items in the different areas are over 0.5 and should be reconsidered before including them in the PSU test.

The graphical Bayesian representations of the difference between groups in the difficulty parameters based on the 1PL-G, 1PL, 2PL, and 3 PL models are displayed in Figs. 1, 2, 3, and 4. In summary, the representations show that bias, defined as a probability of a difference between the groups above 0, is present in a large percentage of the items:

$$P(\beta_{TP} - \beta_{SH} > 0 \mid \text{data}) : 97\%(\mathbf{1PL}) \quad 100\%(\mathbf{1PL - G}) \quad 89\%(\mathbf{2PL}) \quad 93\%(\mathbf{3PL})$$

6 Conclusion

A visual representation of item parameter differences can help determine item bias, and it can help in decision-making regarding item selection to benefit minority groups. In the present study, the differences in the difficulty parameter between academic vs. technical track students were represented (Figs. 1, 2, 3, and 4). Although ETS (Educational Testing Service) flagging rules may underestimate item bias affecting minorities, our Bayesian visual representation determined a more

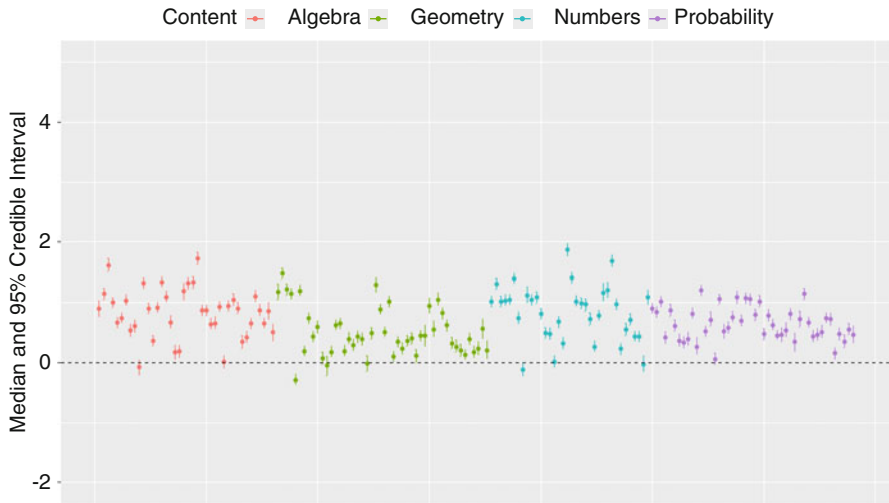


Fig. 1 Difference (TP-SH) between difficulty parameters. 1PL model

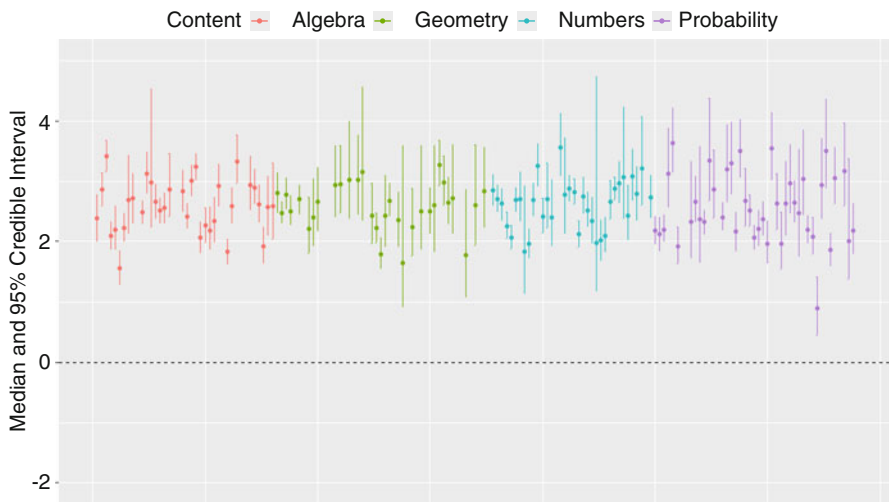


Fig. 2 Difference (TP-SH) between difficulty parameters. 1PL-G model

precise approach: it showed bias in 97% of items according to the 1PL and 1PL-G models, 89% in the 2PL model, and 93% in the 3PL model. When we used the ETS flagging rules, they showed that all items had a type “A” or minimal bias (Table 2), underestimating the differences between student groups. Also, our visual Bayesian analysis is more effective to establish bias against minorities (such as vocational/technical students) compared to traditional measures such as χ^2 Mantel-Haenszel.

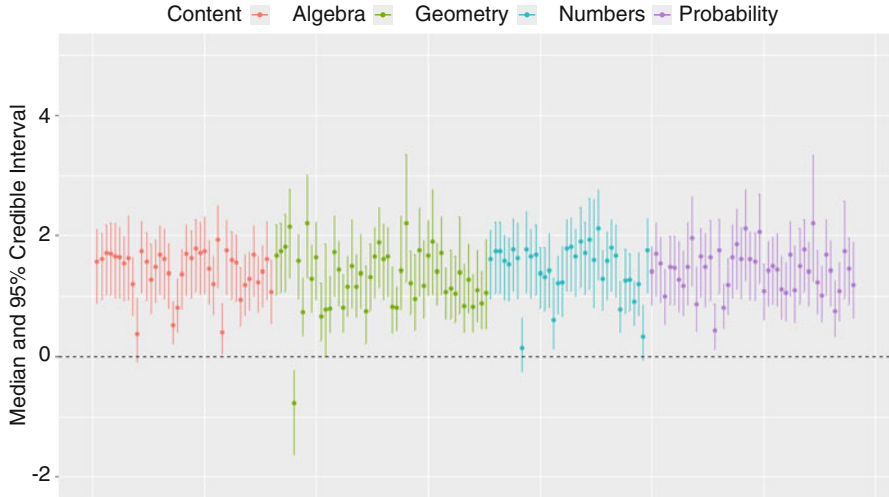


Fig. 3 Difference (TP-SH) between difficulty parameters. 2PL model

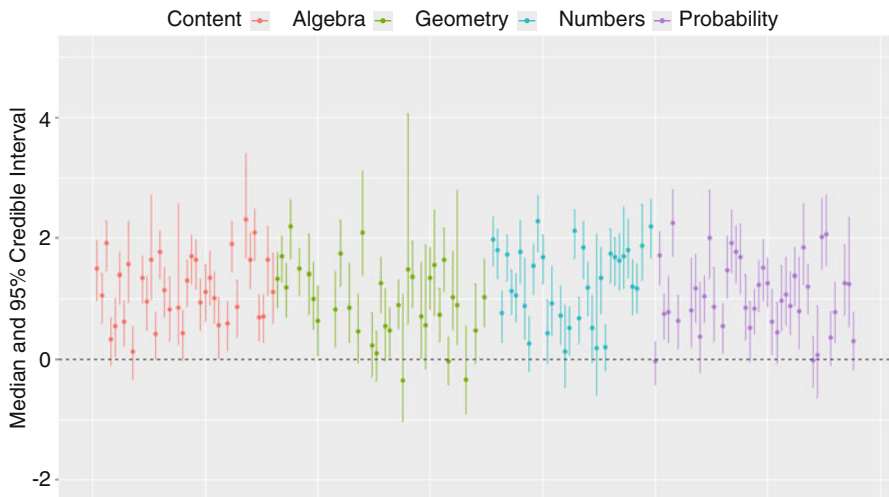


Fig. 4 Difference (TP-SH) between difficulty parameters. 3PL model

In the present study, the Mantel-Haenszel statistic detected item bias between 28% and 45% in each of the test forms (see Table 2) missing the large bias found in the present study. Also, the Bayesian analysis permitted the calculation of posterior distributions of the difficulty parameter making bias analysis more accurate. Finally, the difficulty parameter is suitable to compare groups in a bias analysis.

Acknowledgments This study was funded by project Conicyt PIA CIE 160007.

Table 2 MH statistic for all forms of the PSU test (2018)

Form	Items	MH LORZ majority	MH LORZ minority	CDR	ETS
111	80	19(23,7%)	21(26,2%)	36(45%)	A
112	80	14(17,5%)	13(16,2%)	23(28,7%)	A
113	80	17(21,2%)	18(22,5%)	33(41,2%)	A
114	80	12(15%)	12(15%)	27(33,7%)	A

References

- Berger, M., & Tutz, G. (2016). Detection of uniform and non-uniform differential item functioning by item focused trees. *Journal of Educational and Behavioral Statistics*, 41(6), 559–592.
- Birnbaum, A. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.
- Fariña, P., González, J., & San Martín, E. (2019). The use of an Identifiability-based strategy for the interpretation of parameters in the 1PL-G and Rasch models. *Psychometrika*, 84, 511–528.
- Gonzalez, J. (2010a). Bayesian estimation in IRT. *International Journal of Psychological Research*, 31(1), 164–176.
- Gonzalez, J. (2010b). Bayesian Methods in Psychological Research: The case of IRT. *International Journal of Psychological Research*, 3(1), 164–176.
- Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement*, 7(2), 75–88.
- Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.
- San Martín, E. (2016). Identification of item response theory models, in van der Linden, W. (Ed.). (2016). *Handbook of item response theory*. New York: Chapman and Hall/CRC, <https://doi.org/10.1201/b19166>.
- San Martín, E., & De Boeck, P. (2015). What Do You Mean by a Difficult Item? On the Interpretation of the Difficulty Parameter in a Rasch Model. In: Millsap R., Bolt D., van der Ark L., Wang WC. (eds) *Quantitative Psychology Research*. Springer Proceedings in Mathematics & Statistics, vol 89. Springer, Cham.
- San Martín, E., Del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, 30(3), 183–203.
- San Martín, E., Rolin, J., & Castro, L. M. (2013). Identification of the 1PL model with guessing parameter: Parametric and semi-parametric results. *Psychometrika*, 78(2), 341–379.
- Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, 70(4), 629–650.
- Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample, size requirements, and criterion refinement* (p. 130). ETS Research Reports.

Comparing Hyperprior Distributions to Estimate Variance Components for Interrater Reliability Coefficients



Debby ten Hove, Terrence D. Jorgensen, and L. Andries van der Ark

Abstract Interrater reliability (IRR) is often estimated by intraclass correlation coefficients (ICCs). Using Markov chain Monte Carlo (MCMC) estimation of Bayesian hierarchical models to estimate ICCs has several benefits over traditional approaches such as analysis of variance or maximum likelihood estimation. However, estimation of ICCs with small sample sizes and variance parameters close to zero, which are typical conditions in studies for which the IRR should be estimated, remains problematic in this MCMC approach. The estimation of the variance components that are used to estimate ICCs can heavily depend on the hyperprior distributions specified for these random-effect parameters. In this study, we explore the effect of a uniform and half- t hyperprior distribution on bias, coverage, and efficiency of the random-effect parameters and ICCs. The results indicated that a half- t distribution outperforms a uniform distribution but that slightly increasing the number of raters in a study is more influential than the choice of hyperprior distributions. We discuss implications and directions for future research.

Keywords Bayesian hierarchical modeling · Hyperprior distributions · Interrater reliability · Intraclass correlation coefficients · Markov chain Monte Carlo estimation · Random effects · Variance components

1 Introduction

In an ongoing research project (Ten Hove, Jorgensen, & Ten Hove et al. 2018; Ten Hove et al. 2019), we propose to estimate interrater reliability (IRR) with different types of intraclass correlation coefficients (ICCs) using Markov chain Monte Carlo (MCMC) estimation of Bayesian hierarchical models. MCMC estimation has several benefits over more traditional approaches, such as analysis

D. ten Hove (✉) · T. D. Jorgensen · L. A. van der Ark
Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, the Netherlands
e-mail: D.tenHove@uva.nl

of variance (ANOVA) or maximum likelihood estimation (MLE). For example, MCMC estimation can easily accommodate missing at random data, which is a pitfall of ANOVA (Brennan 2001); for small sample sizes and parameters close to a boundary, it typically outperforms MLE (Gelman et al. 2013); and it provides Bayesian credible intervals, which quantify the uncertainty of the estimated ICCs (Hoekstra et al. 2014), whereas for both ANOVA and MLE, estimating confidence intervals is troublesome (Brennan 2001). However, we found that when using MCMC some ICCs were severely underestimated and inefficient when the number of raters was small or one of the variance components involved in the ICCs was close to zero. As a solution to these estimation difficulties, we proposed a planned missing data design, in which a subset of randomly drawn raters was assigned to each subject. This improved the estimation of ICCs using MCMC vastly, but some ICCs were still biased and inefficient due to biased and inefficiently estimated rater variances (Ten Hove et al. 2019).

The estimation difficulties in conditions with few raters and low variability are consistent with several studies on the performance of MCMC estimation for hierarchical models (Gelman and Hill 2006; McNeish and Stapleton 2016; Polson and Scott 2012; Smid et al. 2019). In the MCMC approach to estimating the ICCs, priors should be specified for the distribution of random effects. Because the variance of these random effects is itself estimated, a prior distribution should be specified for that parameter, called a hyperprior distribution (Gelman et al. 2013, p. 107–108). The performance (e.g., bias and efficiency) of the parameter estimates depends on the specification of these hyperpriors (Gelman and Hill 2006; Gelman et al. 2013; McNeish and Stapleton 2016; Polson and Scott 2012; Smid et al. 2019). The specification of hyperpriors thus provides an opportunity to improve the performance of parameter estimates of random effects. In our current research project, we followed Gelman's (2006, p. 527) advice to start with weakly informative uniform prior distributions on the random effects *SDs*. Several researchers (including Gelman himself) debated the use of these hyperpriors when the data provide little information about clusters (here raters) because uniform distributions may put too much probability mass on unreasonably high values. Various alternatives to these uniform hyperprior distributions were proposed and tested (Gelman 2006; McNeish and Stapleton 2016; Polson and Scott 2012; Spiegelhalter et al. 2004; Van Erp et al. 2019).

This study informs researchers which hyperprior distributions should be used to estimate ICCs. We investigated the effect of different hyperprior distributions on the bias, coverage rates, and efficiency of random-effect parameters and ICCs. The remainder of this paper is structured as follows. First, we briefly discuss the definition of IRR in terms of ICCs and their MCMC estimation. Second, we provide a short overview of (properties of) hyperprior distributions for random effects. Third, we present the results of a simulation study that tested the performance of the random-effect parameters and ICCs using different hyperprior distributions. We focus on conditions with very few raters, to draw attention to difficulties in estimating IRR in conditions that are typical for observational studies. Finally, based on the simulation results, we discuss implications for applied research and directions for future methodological research.

2 Interrater Reliability

2.1 Definition

Bartko (1966), Shrout and Fleiss (1979), and McGraw and Wong (1996) defined IRR in terms of ICCs. They identified raters and subjects as the main sources of variance in a rating process and decomposed each observation into the main and interaction effects of these raters and subjects. Let Y_{sr} be the score of subject s as rated by rater r on attribute Y . Y_{sr} is then decomposed into a grand mean (μ), a subject effect (μ_s), a rater effect (μ_r), an interaction effect, and random measurement error. In practice, the subject \times rater interaction and random measurement error cannot be disentangled, so let μ_{sr} denote a combination of both these elements. The decomposition of Y_{sr} equals

$$Y_{sr} = \mu + \mu_s + \mu_r + \mu_{sr}. \quad (1)$$

If the raters are nested within subjects (i.e., a unique set of raters is used to rate each subject's attribute), μ_r and μ_{sr} cannot be disentangled. For simplicity, we ignore this situation in this paper. Each of the effects in Eq. 1 are assumed to be uncorrelated (Brennan 2001). The variance of Y can therefore be decomposed into the orthogonal variance components of each of these effects, resulting in

$$\sigma_Y^2 = \sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2. \quad (2)$$

If it is assumed that raters and subjects are randomly drawn from a larger population of raters and subjects, respectively, the variances in Eq. 2 are modeled as random-effect variances components.

The variances components in Eq. 2 are used for several definitions of IRR. Each of these definitions is an ICC and defines IRR as the degree to which the ordering (consistency: C) or absolute standing (agreement: A) of subjects is similar across raters. In other words, the IRR is the degree to which subject effects can be generalized over raters. Assume we have k raters rating each subject. The most elaborated ICC (agreement based on the average rating of k raters) is defined as

$$\text{ICC}(A, k) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_r^2 + \sigma_{sr}^2}{k}}. \quad (3)$$

Other definitions of IRR are obtained by removing terms from Eq. 3, as is displayed in Table 1. For more information about these ICCs and the underlying variance decomposition, we refer to Bartko (1966), McGraw and Wong (1996), and Shrout and Fleiss (1979).

Table 1 Cross classification of ICCs in terms of type (agreement and consistency) and number of raters (single rater, $k > 1$ raters)

	Agreement	Consistency
Single rater	$ICC(A, 1) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2}$	$ICC(C, 1) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{sr}^2}$
Average of k raters	$ICC(A, k) = \frac{\sigma_s^2}{\sigma_s^2 + (\sigma_r^2 + \sigma_{sr}^2)/k}$	$ICC(C, k) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{sr}^2/k}$

2.2 MCMC Estimation

The ICCs from Table 1 can be estimated using MCMC estimation of a Bayesian hierarchical model. Let θ denote a model's vector of parameters, and Y denote the data. In the MCMC approach, the posterior distribution of the model parameters given the data, $P(\theta|Y)$, is estimated as proportional to the product of the prior probability distribution of the parameters, $P(\theta)$, and the likelihood of the data conditional on the parameters, $P(Y|\theta)$, that is, $P(\theta|Y) \propto P(\theta)P(Y|\theta)$ (Gelman et al. 2013, p. 6–7).

The MCMC approach thus requires the specification of a prior probability distribution for each parameter. Because MCMC treats each of the random effects in Eq. 1 as parameters to estimate, their variance components in Eq. 2 are so-called hyperparameters (i.e., they are parameters that describe the distribution of other parameters). These hyperparameters require their own prior distribution (named hyperprior distribution), which we discuss in more detail in the following section. Depending on the software, the hyperparameters can be estimated in terms of either random-effect SDs (which should be squared to obtain the random-effect variances for the ICCs) or random-effect variances. For simplicity, we ignore the terms hyperparameters and variance components in the remainder of this paper and consistently use *random-effect variances* to refer to σ_s^2 , σ_r^2 , and σ_{sr}^2 or *random-effect SDs* to refer to the square roots of these random-effect variances.

MCMC estimation repeatedly samples from the posterior distributions, resulting in an empirical posterior distribution for each (hyper)parameter. When deriving ICCs, the posterior distributions of the random-effect variances are combined, yielding an empirical posterior distribution for each of the ICCs. From these empirical posterior distributions, Bayesian credible intervals (BCIs) can be derived that quantify the uncertainty about the random-effect variances and the ICCs that are calculated from these random-effect variances, for example, using percentiles as limits or kernel density estimators to obtain highest posterior density (HPD) limits.

The main difficulty in estimating the ICCs from Table 1 is rooted in the estimation of the random-rater effect variance, σ_r^2 (Ten Hove, Jorgensen, & Ten Hove et al. 2018; Ten Hove et al. 2019). Observational studies often involve few raters, and, when these raters have been trained well, they vary little in the average ratings that they provide. The data thus provide little information about, σ_r^2 . As a result,

its posterior is overwhelmingly influenced by the specified hyperprior distribution. This typically results in an over- and inefficiently estimated random-effect variance. Estimation difficulties of σ_r^2 due to influential hyperprior distributions can, in turn, result in under- and inefficiently estimated ICCs.

3 Hyperprior Distributions

The choice among hyperprior distributions for random-effect variances is frequently discussed (see e.g., Gelman 2006; Gelman et al. 2013; Smid et al. 2019; Van Erp et al. 2019). Prior and hyperprior distributions can be classified into informative or uninformative distributions, proper or improper distributions, and default, thoughtful or data-dependent distributions. For more information on these classifications, we refer to Gelman (2006), Gelman et al. (2013, chapter 2), and Smid et al. (2019).

When raters are skilled and the subjects can be scored objectively, it is reasonable to assume that raters differ little in their average ratings. We therefore believe that the hyperprior distribution for σ_r^2 should be weakly informative and put a relatively large weight on small values compared to large values. The other random-effect variances, σ_s^2 and σ_{sr}^2 , are typically obtained from larger sample sizes (i.e., N (subjects) for σ_s^2 and N (subjects) $\times k$ (raters) for σ_{sr}^2). The data thus provide more information about these random-effect variances, making their hyperprior distributions less influential on the posterior compared to the hyperprior distribution of σ_r^2 . Because σ_s^2 and σ_{sr}^2 are expected to be larger than σ_r^2 , their hyperprior distributions should allow for large values.

Given these considerations, we prefer weakly informative, thoughtful hyperprior distributions. Moreover, we prefer hyperprior distributions that yield proper posterior distributions. We take these criteria into account while discussing three popular hyperprior distributions for variance parameters: a uniform distribution (Gelman 2006; McNeish and Stapleton 2016), the inverse-gamma distribution (Spiegelhalter et al. 2004), and the half- t or half-Cauchy distributions (Gelman 2006; McNeish and Stapleton 2016; Polson and Scott 2012).

3.1 Uniform Distribution

The uniform distribution is a popular hyperprior distribution with two parameters: a lower bound and an upper bound. This distribution implies a researcher's believe that all values within a specified range are equally likely. For random-effect SDs, the uniform hyperprior distribution can be specified as weakly informative by using the range $[0, \frac{\max_Y - \min_Y}{2}]$ (i.e., the smallest and largest possible SD), where \max_Y and \min_Y are the maximum and minimum value of Y , respectively. If \max_Y and \min_Y are estimated from the data, the uniform distribution is data dependent; if \max_Y and \min_Y are specified as the theoretically maximum and minimum values of Y ,

respectively (e.g., using the minimum and maximum possible scores, such as anchor points on a Likert scale), the uniform distribution is data independent. In practice, it is unlikely to find a random-effect SD near the upper bound of $\left[\frac{\text{maxy}-\text{miny}}{2}\right]$. Such a large upper bound is unintentionally influential on the posterior when the data contain too little information about a parameter. Examples of little information include small sample sizes but also when the random-effect variance is nearly zero. However, it may be difficult to defend the choice of a hard upper bound of the uniform posterior distribution that is smaller than the maximum possible SD . A uniform hyperprior distribution performs best when it is specified for random-effect SD s (e.g., σ_r), rather than for random-effect variances (e.g., σ_r^2). To yield proper posteriors, a hyperprior distribution requires at least three clusters for random-effect SD s, or at least four clusters for random-effect variances (Gelman 2006).

3.2 *Inverse-Gamma Distribution*

The inverse-gamma distribution is another popular hyperprior distribution for random-effect variances, which is defined on a positive scale and has two parameters: a shape and scale parameter. This distribution is very sensitive to its specified shape and scale parameters when the estimated σ^2 is small, and its specification is therefore too influential for typical IRR studies in which σ_r^2 is expected to be low (Gelman et al. 2013, p. 315–316). Moreover, the inverse-gamma hyperprior distribution yields improper posteriors when the shape and scale parameters are specified as very uninformative (Gelman 2006). Therefore, although it is a commonly applied prior in many other settings (and potentially required as a conjugate prior for Gibbs sampling), we consider the inverse-gamma hyperprior inappropriate for our purpose.

3.3 *Half- t or Half-Cauchy Distribution*

The half- t and half-Cauchy hyperprior distributions were also proposed as hyperprior distributions for random-effect variances and are defined on a positive scale (Gelman 2019; Polson and Scott 2012). The half- t distribution has three parameters: a shape, location, and scale parameter. The half-Cauchy distribution is equivalent to a half- t distribution for $df = 1$ and thus only has a location and scale parameter. The half-Cauchy distribution has more kurtosis than t distributions having $df > 1$, allowing the greatest probability density for extreme values while still placing most probability density near the center of the distribution. If a wide range of possible values is specified for the random-effect variances, these distributions are specified as data independent and weakly informative. Especially for σ_r^2 , we expect values near zero, so a half- t distribution with higher $df = 4$ is slightly more informative and is recommended for variance parameters that are expected to have values near

the lower bound of zero (Gelman 2019). This may, however, be less beneficial for the other random-effect variances in Eq. 2.

4 Simulation Study

4.1 Methods

4.1.1 Data Generation

We generated data from Eq. 1 using the parameters in Eq. 2. We fixed μ to 0 and drew $N = 30$ values of μ_s from $\mathcal{N}(0, \sigma_{sr}^2 = \frac{1}{2})$, k values of μ_r from $\mathcal{N}(0, \sigma_r^2)$, and $30 \times k$ values from from $\mathcal{N}(0, \sigma_{sr}^2 = \frac{1}{2})$, and used Eq. 1 to obtain scores Y_{sr} . The choice to keep N , σ_s^2 , and σ_{sr}^2 constant were arbitrary but believed to be realistic for observational studies with a small number of subjects.

4.1.2 Independent Variables

We varied the number of raters, the variance in the random-rater effects, the hyperprior distributions of the random-effect SDs, the type of estimator, and the type of BCI. The number of raters (k) had three levels: $k = 2, 3$, and 5 . We selected $k = 2$ because two is the minimum number of raters required to estimate the IRR. And $k = 2$ is often used in the applied literature to estimate IRR. We also specifically incorporated this low number to draw attention to estimation difficulties. We selected $k = 3$ because a sample size of at least three is required to yield proper posteriors for uniform distributions. We used $k = 5$ to see how the results of different priors differed for slightly higher sample sizes.

The random-rater effect variance (σ_r^2) had two levels: $\sigma_r^2 = .01$ and $\sigma_r^2 = .04$. Random-effect variances extremely close to the lower bound of zero are typically poorly estimated but are less influential in the ICCs. Therefore, we used a value extremely close to zero and a slightly higher value to test whether increasing σ_r^2 improved the estimations. The population ICCs of $\text{ICC}(A, 1)$ and $\text{ICC}(A, k)$ ranged from 0.48 to 0.83. $\text{ICC}(C, 1)$ and $\text{ICC}(C, k)$ do not include σ_r^2 (Table 1) and are thus identical across the levels of σ_r^2 . Therefore, we further ignored the $\text{ICC}(C, 1)$ and $\text{ICC}(C, k)$ in this study and focused on the estimation $\text{ICC}(A, 1)$ and $\text{ICC}(A, k)$.

The hyperprior distributions had three levels: uniform, half- t , and mixed. We specified each of these hyperprior distributions for the random-effect SDs (i.e., σ_s , σ_{sr} , and σ_r) rather than the random-effect variances. In the uniform condition, we specified uniform hyperprior distributions over the range $[0, \frac{\text{max}_Y - \text{min}_Y}{2}]$ for all random-effect SDs, with max_Y and min_Y being estimated from the data. A specified upper bound should not be data dependent, but we could not specify a reasonable data-independent upper bound because our simulated data have no natural boundaries such as Likert scales do. Nonetheless, this data-dependent upper

bound will probably behave comparable to natural upper bounds for the random-effect SDs , because the specified range still places too much probability mass on unreasonably high values compared to the expected lower values. In the half- t conditions, we specified a half- $t(4,0,1)$ hyperprior distributions for all random-effect SDs . In the mixed conditions, we used a uniform hyperprior distribution for σ_s and σ_{sr} and a half- t hyperprior distribution for σ_r . We refer to Sect. 3 for the justification of these choices.

We obtained point estimates of σ_r and the ICCs using two approaches: posterior means (i.e., expected a posterior; EAPs) and posterior modes (i.e., maximum a posteriori; MAPs). Small sample sizes, such as the small number of raters, can result in skewed posterior distributions, especially for random-effect parameters near the lower bound of zero. Modal estimates resemble their MLE counterparts and are, especially for skewed distributions, preferred over the mean and median as a measure of central tendency (Gelman 2006).

We obtained 95% BCIs using two approaches: Using 2.5% and 97.5% percentiles as limits, and using the highest posterior density intervals (HPDIs). Percentiles are readily provided by most MCMC software but are only appropriate for symmetric unimodal distributions. HPDIs can be obtained from the empirical posterior distribution using kernel densities and accommodate skewness in the posterior distributions. When the posterior is bimodal or non-symmetrical, these approaches may thus yield very different results (for a brief discussion, see Gelman et al. 2013, p. 33).

The simulation design was fully crossed, resulting in $3(k) \times 2(\sigma_r^2) \times 3$ (hyperprior) = 18 between-replications conditions, for each of which we simulated 1000 replications. Within each of the 18 between-replication conditions, we investigated bias for the two types of point estimate (EAPs and MAPs) and coverage rates for the two types of interval estimate (percentiles and HPDI).

4.1.3 Parameter Estimation

We used MCMC estimation of Bayesian hierarchical models and specified the hyperpriors of the random-effect SDs as discussed in the previous paragraph. We used three independent chains of 1000 iterations. The first 500 iterations per chain served as burn-in iterations, and the last 500 iterations of each chain were saved in the posterior. This resulted in a posterior of 1500 iterations to estimate each parameter. We checked convergence using the potential scale reduction factor, \hat{R} , and the effective sample size, N_{eff} , of each parameter (Gelman 2006). If any of the $\hat{R} < 1.10$, we doubled the number of burn-in iterations. This was repeated until the model converged, or did not converge after the limit of 10,000 burn-in iterations was reached, in which case the replication was discarded. Thereafter, we checked whether each parameter's N_{eff} exceeded 100. If a parameter or ICC showed an effective sample size that was too low, we increased the number of post burn-in iterations based on the lowest N_{eff} with a factor of $120/\min(N_{\text{eff}})$.

4.1.4 Software

We used the R software environment (R Core Team 2019) for data generation and analyses, and the Stan software (Stan Development Team 2017) with the R package `rstan` (Stan Development Team 2018) to estimate the Bayesian hierarchical models and the ICCs. We obtained the MAP estimates using the `modeest` package (Poncet 2019) and 95% HPDIs using the `HDInterval` package (Meredith and Kruschke 2018). Our software code is available on the Open Science Framework (OSF): <https://osf.io/shkqm/>

4.1.5 Dependent Variables

We evaluated the quality of the estimated ICCs and the random-rater effect SD , σ_r ,¹ using four criteria: convergence, relative bias, 95% BCI coverage rates, and relative efficiency. We calculated the percentage of converged solutions per condition, which was preferably 100%. Let $\bar{\theta}$ denote the average EAP or MAP estimate of σ_r or the derived ICC across replications in a condition, and let θ denote the population parameter in that condition. We computed relative bias as $\frac{\bar{\theta}-\theta}{\theta}$, and we used relative bias $> .05$ as indicating minor bias and $> .10$ as indicating substantial bias. We tested the 95% BCI coverage rates of both percentiles and HPDIs, using a coverage rate $< 90\%$ and $> 97\%$ as a rule of thumb for defining the width of BCIs as, too narrow or wide BCIs, respectively. We calculated relative efficiency as the ratio of the average posterior SD of σ_r and the ICCs, relative to the SD of their posterior means.² A ratio of 1 indicates accurate estimates of variability. We used relative efficiency $< .90$ or > 1.10 as indicating minor under- or overestimation of the posterior SD s and relative efficiency $< .80$ or > 1.20 as indicating substantial under- or overestimation of the posterior SD s.

4.2 Results

We provide a summary of the simulation results and diverted the complete results to the authors' OSF account. The results for the conditions with mixed hyperprior distributions and the conditions with half- t hyperprior distributions for each random-effect SD were very similar. Therefore, we do not discuss the results

¹We focused on σ_r instead of σ_r^2 because our Stan program estimated random-effect SD s, from which we derived the random-effect variances.

²We want to emphasize the difference between random-effect SD s, which quantify the variability of the random effects, and the posterior SD s, which quantify the uncertainty about the estimated parameters (the random-effect SD s and the ICCs).

of conditions with mixed hyperprior distributions. Similarly, the results for the estimated $ICC(A,k)$ resembled the results for $ICC(A,1)$, so we only present the results for the $ICC(A,1)$.

4.2.1 Convergence

All replications in all conditions converged to a solution. The following results are therefore based on $18 \times 1000 = 18,000$ replications.

4.2.2 Relative Bias

Figures 1 and 2 show the relative bias of the estimated σ_r and $ICC(A,1)$ across conditions. Both σ_r and $ICC(A,1)$ showed less bias in conditions with a half- t hyperprior distribution than in those with a uniform hyperprior distribution. EAPs severely overestimated σ_r , whereas the MAP was an unbiased estimator of this parameter in all conditions with $k > 2$. The MAP and EAP estimates of $ICC(A,1)$ were comparable. Neither σ_r or $ICC(A,1)$ resulted in unbiased estimates in any condition with $k = 2$. MAPs of both σ_r and $ICC(A,1)$ were unbiased in all conditions with $k = 5$.

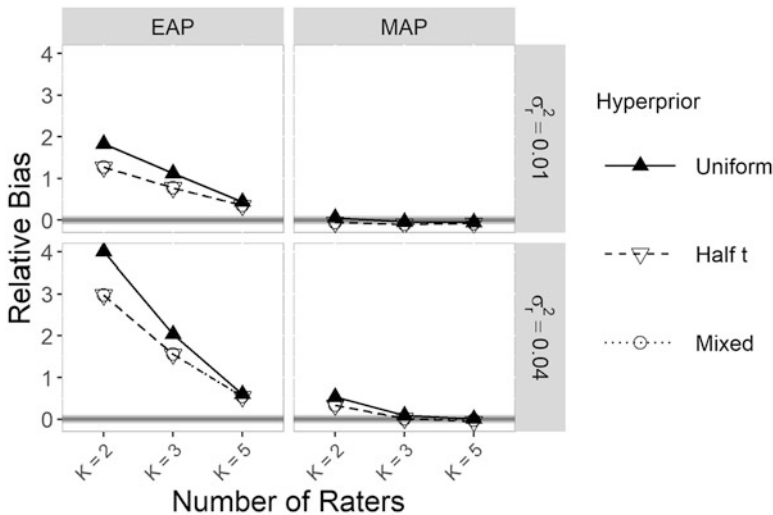


Fig. 1 Relative bias of σ_r under different conditions. White areas, large bias ($>10\%$); light-gray areas, substantial bias ($5\text{--}10\%$); dark-gray areas, minor bias ($< 5\%$)

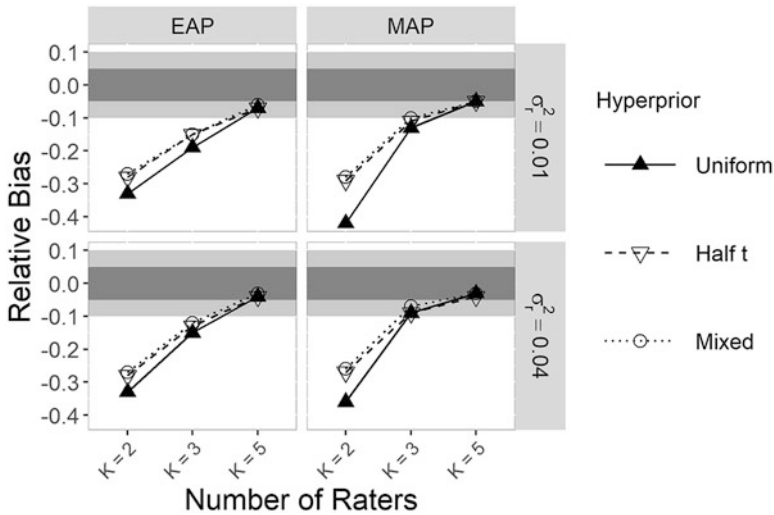


Fig. 2 Relative bias of ICC(A,1) under different conditions. White areas, large bias (>10%); light-gray areas, substantial bias (5–10%); dark-gray areas, minor bias (< 5%)

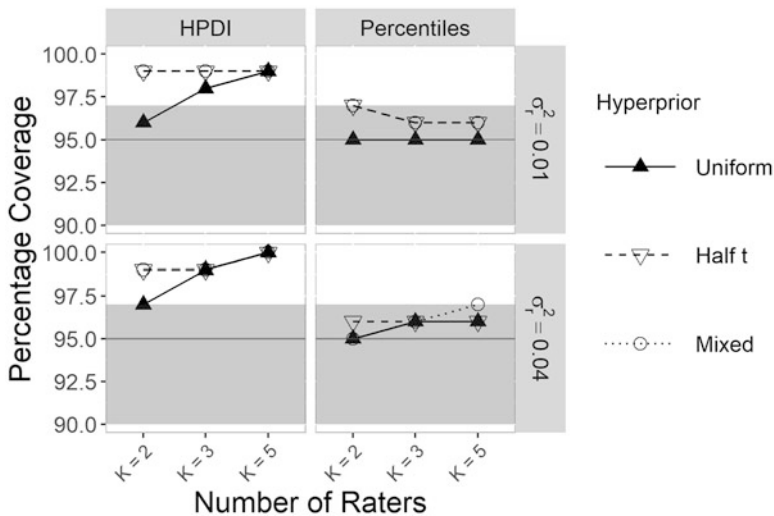


Fig. 3 95% BCI coverage rates of σ_r under different conditions. White areas, substantially too narrow (< 90%) or too wide BCIs (> 97%); light-gray areas, slightly too narrow (90 ≤ 95%) or too wide BCIs (95 > 97%)

4.2.3 95% BCI Coverage

Figures 3 and 4 show the 95% BCI coverage rates of σ_r and ICC(A,1), respectively, across conditions. HPDIs were too wide for σ_r but yielded nominal coverage rates

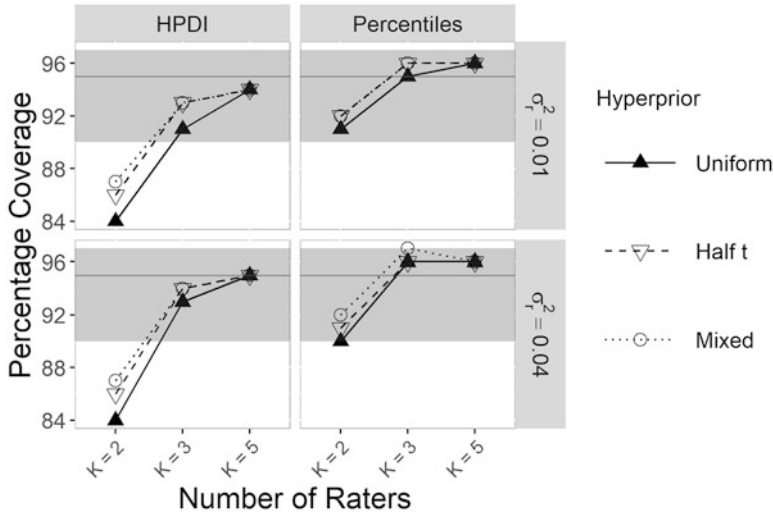


Fig. 4 95% BCI coverage rates of ICC(A,1) under different conditions. White areas, substantially too narrow (< 90%) or too wide BCIs (> 97%); light-gray areas, slightly too narrow (90 ≤ 95%) or too wide BCIs (95 > 97%)

for the ICC(A,1) for more than two raters. Percentiles yielded nominal coverage rates for σ_r and for the ICC(A,1) but only for $k > 2$.

4.2.4 Relative Efficiency

Figure 5 shows the relative efficiency of the estimated σ_r and ICC(A,1) across conditions. Both hyperprior distributions yielded posterior SDs of both σ_r and ICC(A,1) that were considerably larger than the actual sampling variability of these estimates. The overestimation of posterior SDs decreased when k increased but remained severe even in conditions with $k = 5$. Overestimation of the posterior SDs was more severe for σ_r than for ICC(A,1) and comparable for both hyperprior distributions.

5 Discussion

The results of this study indicate that half- t hyperprior distributions have a slight advantage over uniform hyperprior distributions for estimating IRR with ICCs. The best performing condition combined MAP point estimates, percentiles based BCIs, half- t hyperprior distributions, and $k > 2$ raters. For $k = 2$, ICCs were underestimated and inefficient. This bias and inefficiency decreased as k

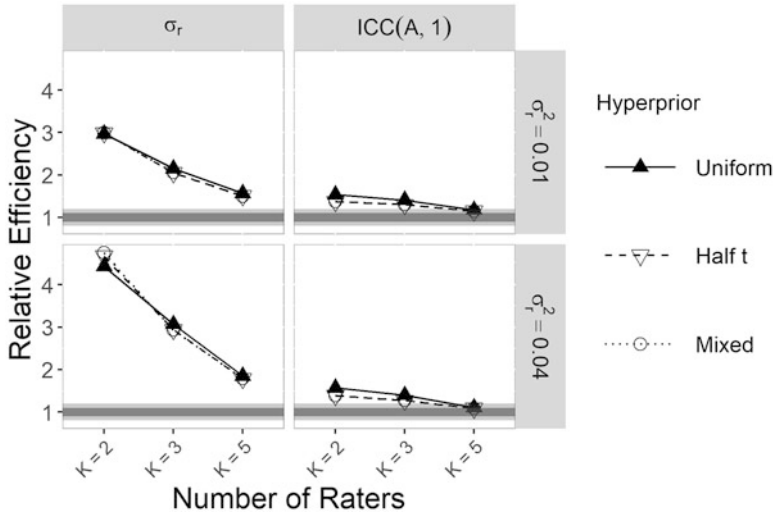


Fig. 5 Relative Efficiency of σ_r and ICC(A,1) under different conditions. White areas: highly inefficient (>20%); Light-gray areas: substantial inefficient (10–20%); Dark-gray areas: slightly inefficient (< 10%)

increased. For $k > 2$ (the conditions with unbiased estimates), the combination of a half- t hyperprior distribution with percentile BCIs yielded nominal coverage rates. Overall, the number of raters used to estimate IRR had a larger effect on the performance of the MCMC estimates than the choice of hyperprior distributions.

The results of this study are in line with earlier research indicating that random-effect variances cannot be properly estimated when the number of clusters (here raters) is as small as two (Gelman 2006). This should discourage researchers from estimating the IRR with ICCs when data are collected from as few as two raters, a situation that we observed frequently in the applied literature. Using $k > 2$ raters in an observational study may sound like a high burden for researchers. Fortunately, estimation of the IRR in conditions with scarce resources could already be improved by randomly sampling a subset of raters for each subject from a larger rater pool (Ten Hove et al. 2019). This would result in a larger rater-sample size, with missing at random data. This resembles an often seen practice (Viswesvaran et al. 2005), which diminishes the burden per rater and allows to keep the total number of observations at the same level as a fully crossed design in which each of two raters rates each subject. It would be interesting to test the combination of the half- t hyperprior distribution with such a planned missing data design in a future study.

Our simulation study was not comprehensive concerning the number of conditions. The performance of the ICCs in our simulation study may thus, for example, depend the population values of the other random-effect variances in the ICCs. Our statements about obtaining (in)appropriate estimates for these ICCs can

therefore not readily be generalized to conditions with differing variability in each of the involved effects. However, the results on σ_r itself are promising, because its estimation seems to improve when the variability in the rater effects increases only slightly (at least for $k > 2$). With increasing magnitude, the rater variance had a larger effect on the ICCs, and arguably, the quality of its estimates has a larger influence on the quality of the ICC estimates. Presumably, the ICCs will thus be estimated more accurately and efficiently when the variability in rater effects increases.

In conclusion, we advise researchers to use an half- t hyperprior distribution for the random-rater effect SD , MAP point estimates, percentiles based BCIs, and, most importantly, at least three raters to estimate the IRR using an MCMC approach. However, we want to highlight Gelman's (2006) advice that every noninformative or weakly informative (hyper)prior distribution is inherently provisional, implying that researchers should always inspect whether their posterior forms a proper distribution. He argued that, if an approach yields improper posteriors, there is more prior information available that needs to be incorporated in the estimation procedure.

References

- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3–11. <https://doi.org/10.2466/pr0.1966.19.1.3>
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1, 515–534. Retrieved from <https://projecteuclid.org/euclid.ba/1340371048>
- Gelman, A. (2019). *Prior choice recommendations*. Retrieved from <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). New York: Chapman and Hall/CRC.
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21, 1157–1164. <https://doi.org/10.3758/s13423-013-0572-3>
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>
- McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28, 295–314. <https://doi.org/10.1007/s10648-014-9287-x>
- Meredith, M., & Kruschke, J. (2018). *HDInterval: Highest (posterior) density intervals*. Retrieved from <https://CRAN.R-project.org/package=HDInterval> (Computer software)
- Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4), 887–902. <https://doi.org/10.1214/12-BA730>
- Poncet, P. (2019). *modeest: Mode estimation*. Retrieved from <https://CRAN.R-project.org/package=modeest> (Computer software)
- R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/> (Computer software)

- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2019). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 169–191. <https://doi.org/10.1080/10705511.2019.1604140>
- Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation* (Vol. 13). New York: Wiley.
- Stan Development Team. (2017). *Stan modeling language: User's guide and reference manuals*. Retrieved from <https://mc-stan.org/users/interfaces/stan.html> (Computer software)
- Stan Development Team. (2018). *RStan: The R interface to Stan*. Retrieved from <https://mc-stan.org/users/interfaces/rstan.html> (Computer software)
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2018). *Interrater reliability for dyad-level predictors in network data*. (Paper presented at the XXXVIII Sunbelt 2018 Conference, Utrecht)
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2019). *Interrater reliability for multilevel data: A generalizability theory approach*. (Paper presented at the 84th annual International Meeting of the Psychometric Society, Santiago, Chile).
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2019). *Interrater reliability for multilevel data: A generalizability theory approach*. (Manuscript submitted for publication)
- Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. <https://doi.org/10.1016/j.jmp.2018.12.004>
- Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*, 108–131. <https://doi.org/10.1037/0021-9010.90.1.108>

A Hierarchical Joint Model for Bounded Response Time and Response Accuracy



Sandra Flores, Jorge Luis Bazán, and Heleno Bolfarine

Abstract Response time (RT) models traditionally consider positive continuous distributions, for instance, the lognormal, gamma, exponential, and Weibull distributions, with support $< 0, \infty >$. However, usually in Assessment, the time that an examinee takes to complete a test is limited and not infinite as the previous models assume. By considering this fact, the purpose of this article is to model RT following a bounded distribution and then in combination with response accuracy to obtain a joint model. Specifically, the use of the simplex distribution is proposed to model RT adopting the Bayesian inference. Performance of the proposed model is evaluated in a simulation study and the PISA 2015 computer-based reading data is used to apply the model.

Keywords Bounded distribution · Hierarchical model · Limited variable · Response time

1 Introduction

Currently, with the use of personal computers (PCs) in Assessment as, for example, using PCs by the Programme for International Student Assessment (PISA), the time that an individual spends in resolving an item becomes easily available. Several proposals of models are using those Response Times (RTs) as additional information estimating the ability of examinees. For instance, see van der Linden (2007), Fox et al. (2007), Klein Entink (2009), Im (2015), and Zhan et al. (2018).

S. Flores (✉) · H. Bolfarine

Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, SP, Brazil
e-mail: sefari@usp.br

J. L. Bazán

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brazil

Studies of RT to investigate mental activity were developed in literature since a long time; they were associated with the study of the speed to answer a Test. A review of those studies is presented in the work of Schnipke and Scrams, see Schnipke and Scrams (2002). At the beginning, based in the number of correct responses in a Test, it was thought that RTs and accuracy measured the same construct, but further investigations did not support this statement. From this definition and also from the increasing availability of RTs data, several models for speed were suggested in literatures which use RT models for the items of a Test.

A classification for this time models, suggested by Fox et al. (2007), is given in three ways: the first one models RT in the IRT context, the second one models RT separately, and the third one models RT and response accuracy in a hierarchical way, with a trade-off relation between speed (measurement of response time) and accuracy (measurement of correct response). The last one permits to empirically understand the relation between speed and accuracy.

Because of the lower bound at zero, a classical distribution assumed for RT is the lognormal distribution. Several works, for instance, Schnipke and Scrams (1999), show the best performance of this selection. However, in a more realistically thinking, time is not infinite. It has an upper bound, since the Test has a time set to complete all responses. This work investigates the use of a limited distribution to model the proportion of time spent in an item. Specifically, the simplex distribution (Barndorff-Nielsen and Jorgensen 1991; Jorgensen 1997) is considered in order to model RTs. After that, an evaluation of the performance of this model is presented.

Additionally, the model of proportion of time previously suggested in the hierarchical framework proposed by van der Linden (2007) is applied. That proposal suggests modeling responses (as accuracy) and RTs (as speed) in a jointly way. Accuracy model is named in this work as IRT part and speed model as RT part. The two constructs (speed and accuracy) are specified in two levels: the first one, for individual level, implements the trade-off relation between accuracy and speed, and the second one is at a population level, it means for the complete group of examinees.

The remainder of this paper is organized as follows: Sect. 2 defines the simplex distribution that will be used to model RTs; in Sect. 3 the model for limited or bounded response time is presented, which is used in an hierarchical model in Sect. 4. Bayesian estimation of the proposal is studied in Sect. 5. Simulation studies comparing the performance of the suggested model are shown in Sect. 6, an application of the model is presented in Sect. 7, and final comments are discussed in Sect. 8.

2 The Simplex Distribution

Simplex distribution was introduced by Barndorff-Nielsen and Jorgensen (1991) and Jorgensen (1997) being part of a general family named as dispersion models. The univariate simplex distribution $S(\mu, \sigma^2)$, with parameters $\mu \in (0, 1)$ as a

position parameter and $\sigma^2 > 0$ as a dispersion parameter, is given by the following probability density function:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2\{x(1-x)\}^3}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2x(1-x)\mu^2(1-\mu)^2}\right\} \quad (1)$$

where mean and variance are given by the following expressions:

$$E(X|\mu, \sigma^2) = \mu$$

$$Var(X|\mu, \sigma^2) = \mu(1-\mu) - \frac{1}{\sqrt{2\sigma^2}} \exp\left\{\frac{1}{\sigma^2\mu^2(1-\mu)^2}\right\} \Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\}$$

with $\Gamma(a, b) = \int_b^\infty t^{a-1} e^{-t} dt$ being the incomplete gamma function.

Some shapes of simplex density are depicted in Fig. 1. It is possible to see how simplex distribution handle different forms of the density. The simplex distribution is more flexible than other distributions, for example, it handles bimodality when σ^2 increases Quintero (2017). Note, expectation value is the mean parameter μ and variance depends jointly on mean and dispersion parameters (μ, σ) . This distribution was used as a generalized linear model in Song and Tan (2000), and an R package for simplex regression analysis was developed by Zhang et al. (2016).

3 The Bounded Response time (BRT) Model

The present section suggests a model to analyse RT as a bounded variable. An examinee, being evaluated to determine his or her ability, typically takes a test with a fixed period of time. Then, he or she can distribute this fixed time in all items according to his or her ability and to strategies applied during this fixed period of time. For this reason, a model for limited response time is proposed.

Considering a RT data set for I examinees in J items, each RT is the realization of a random variable T_{ij} , where i denotes an examinee and j an item. Defining d_j as the greater time that some examinee spends in an item j and c_j as the smaller time, it is possible to identify the proportion of time (PT), Z_{ij} , that an examinee i spends in item j as follows:

$$Z_{ij} = \frac{T_{ij} - c_j}{d_j - c_j}$$

Thus, Z_{ij} is in $< 0, 1 >$ interval and could be modeled with the simplex distribution having parameters defined as follows:

$$z_{ij}|\tau_i, \beta_j, \alpha_j \sim S(\mu_{ij}, 1/\alpha_j^2) \quad (2)$$

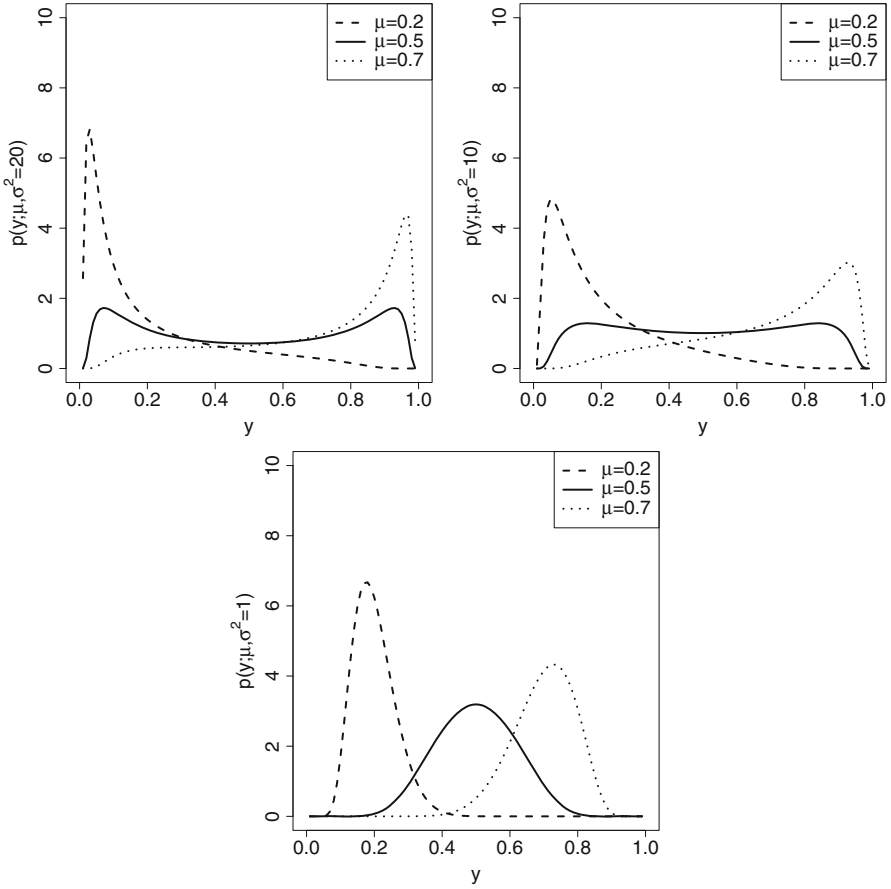


Fig. 1 Simplex density functions for different values of μ and σ

$$g(\mu_{ij}) = \beta_j - \tau_i$$

where $S(\cdot, \cdot)$ denotes the simplex distribution as defined in Eq. (1); note that $\alpha = 1/\sigma$ gives a precision parameter interpretation. $\beta_j \in \mathbb{R}$ is the position parameter for item j , $\tau_i \in \mathbb{R}$ is the position parameter for examinee i , and $\alpha_j > 0$ is the precision parameter and could be interpreted in a similar way as the discrimination power for the item j . Link function g is defined as the logit function, meaning $g(x) = \log(\frac{x}{1-x})$. Simplex distribution preserves and models the asymmetric behavior of RTs.

4 Hierarchical Joint Model for Bounded Response Time and Response Accuracy (HBRT)

This section suggests a model for speed, which has bounded time and accuracy. This proposal follows the work of van der Linden, see van der Linden (2007). For that purpose, consider a data set which contains response accuracy, $y_{ij} \in \{0, 1\}$, and proportion of response time, $z_{ij} \in (0, 1)$, from an examinee i and an item j , which is jointly modeled in two levels as follows:

In the first level, speed and accuracy are modeled using two distributions. Response accuracy, correct and incorrect responses for the items, is modeled using the two-parameter IRT probit model. In other words,

$$y_{ij} \mid \theta_i, a_j, b_j \sim \text{Bernoulli}(p_{ij}) \tag{3}$$

$$p_{ij} = p(Y_{ij} = y_{ij} \mid \theta_i, a_j, b_j) = \Phi(a_j(\theta_i - b_j))^{y_{ij}} (1 - \Phi(a_j(\theta_i - b_j)))^{1-y_{ij}}$$

where variable Y_{ij} follows a Bernoulli probability function with probability of correct response $p(Y_{ij} = 1 \mid \theta_i, a_j, b_j) = \Phi(a_j(\theta_i - b_j))$. The ability parameter for examinee i is given by $\theta_i \in \mathbb{R}$. $a_j > 0$ and $b_j \in \mathbb{R}$ denote parameters for item j which are discrimination and difficulty, respectively.

On the other hand, proportion of RT, z_{ij} , is modeled according to proposal given in Eq. (2), where the probability density function of variable Z is given as follows:

$$p(z_{ij} \mid \tau_i, \alpha_j, \beta_j) = \frac{\alpha_j}{\sqrt{2\pi \{z_{ij}(1 - z_{ij})\}^3}} \exp\left\{-\frac{\alpha_j^2 \left(z_{ij} - \frac{1}{1+e^{-(\beta_j - \tau_i)}}\right)^2}{2z_{ij}(1-z_{ij}) \left(\frac{1}{1+e^{-(\beta_j - \tau_i)}}\right)^2 \left(\frac{1}{1+e^{\beta_j - \tau_i}}\right)^2}\right\} \tag{4}$$

with $\tau_i \in \mathbb{R}$ being the speed parameter of the examinee i and item parameters, as described in Sect. 3, $\alpha_j > 0$ and $\beta_j \in \mathbb{R}$ are discrimination and time intensity, respectively.

As a second level, assuming conditional independence, for each examinee and item responses, considering ability and speed parameters $\xi_i = (\theta_i, \tau_i)$, the joint distribution takes the form:

$$p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\xi}, \mathbf{v}) = \prod_{i=1}^I \prod_{j=1}^J p(y_{ij} \mid \theta_i, a_j, b_j) p(z_{ij} \mid \tau_i, \alpha_j, \beta_j) \tag{5}$$

where $p(y_{ij} \mid \cdot)$ and $p(z_{ij} \mid \cdot)$ are defined as in Eqs. 3 and 4, respectively.

5 Bayesian Estimation

In order to estimate the proposed joint model under a Bayesian approach, likelihood is specified, then the priors for the parameters in the model and finally the posterior distribution is obtained. Likelihood function is given by the following equation:

$$L(\boldsymbol{\xi}, \mathbf{v} | \mathbf{y}, \mathbf{z}) = \prod_{i=1}^I \prod_{j=1}^J p(y_{ij} | \theta_i, a_j, b_j) p(z_{ij} | \tau_i, \alpha_j, \beta_j) p(\xi_i | \mu_\xi, \Sigma_\xi) p(v_j | \mu_v, \Sigma_v) \quad (6)$$

where $\boldsymbol{\xi} = (\xi_1, \dots, \xi_I)$ represents the vector of abilities and speed parameters for each one of the examinees and $\mathbf{v} = (v_1, \dots, v_J)$ is the vector of item parameters, $v_j = (a_j, b_j, \alpha_j, \beta_j)$.

Assuming that $p(\xi_i | \mu_\xi, \Sigma_\xi)$ follows a bivariate normal prior distribution, two marginal normal distributions are defined as $p(\theta_i | \mu_\theta, \sigma_\theta^2)$ and $p(\tau_i | \mu_\tau, \sigma_\tau^2)$, where $\mu_\tau = \rho_{\theta\tau} \theta_i$ and $\rho_{\theta\tau}$ provides the relation between θ and τ , also $\sigma_\tau^2 = \sigma_\tau^{2c} + \rho_{\theta\tau}^2 \sigma_\theta^2$ (Fox et al. 2007). Additionally, it is assumed that v_j follows a multivariate normal distribution of dimension 4, with density function given by

$$p(v_j | \mu_v, \Sigma_v) = \frac{|\Sigma_v^{-1}|^{1/2}}{(2\pi)^{5/2}} \exp \left\{ -\frac{1}{2} (v_j - \mu_v)^T \Sigma_v^{-1} (v_j - \mu_v) \right\}$$

which has a mean vector $\mu_v = (\mu_a, \mu_b, \mu_\theta, \mu_\tau)$ and a covariance matrix

$$\Sigma_v = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{bmatrix}.$$

In order to identify the most suitable joint model structure, following van der Linden (2006, 2007), the mean of speed is fixed as $\mu_\tau = 0$, the mean of accuracy as $\mu_\theta = 0$, and the variance of accuracy as $\sigma_\theta^2 = 1$. Considering that responses give the estimation of item parameters, no constraint is defined for item parameters.

This section adopts the Bayesian approach to estimate the parameters of the model previously proposed. In this approach the conclusion about a parameter is given in terms of probabilities which are conditional regarding the observed values (Gelman et al. 1995). Models for responses and response times usually adopt the Bayesian approach, as for example, van der Linden (2007); Klein Entink (2009); Im (2015); Zhan et al. (2018).

Since the distribution of the parameter depends on hyper-parameters, the following prior distributions are defined in order to complete the model:

$$\rho_{\theta\tau} \sim \mathcal{N}(0, \sigma_\rho^2) \quad (7)$$

$$\sigma_\tau^2 \sim \text{Inverse} - \text{Gamma}(v_1, v_2) \quad (8)$$

$$\Sigma_v \sim \text{Inverse - Whishart}(\Sigma_{v0}^{-1}, v_{v0}) \quad (9)$$

$$\mu_v | \Sigma_v \sim \text{MVN}(\mu_{v0}, \Sigma_v) \quad (10)$$

σ_ρ^2 , v_1 , v_2 take 0.1 value, giving little information for the model. Σ_{v0} is an identity matrix, $v_{v0} = 4$ and $\mu_{v0} = (0, 0, 0, 0)$ (van der Linden 2007).

The posterior distribution of the parameters is given by the following equation:

$$p(\xi, \nu, \sigma_\tau^2, \rho_{\theta\tau}, \mu_\nu, \Sigma_\nu | \mathbf{y}, \mathbf{z}) \propto p(\mathbf{y}, \mathbf{z} | \xi, \nu) p(\xi, \nu) p(\sigma_\tau^2) p(\rho_{\theta\tau}) p(\mu_\nu | \Sigma_\nu) p(\Sigma_\nu) \quad (11)$$

The implementation of the proposed model uses R2WinBUGS R package. This package runs under WinBUGS software, which is an interactive Windows version of the BUGS program for Bayesian analysis of complex statistical models using Markov Chain Monte Carlo (MCMC) techniques (Lunn et al. 2000). The code for the model is released in the appendix.

6 Simulation Studies

A simulation study comparing the performance of the suggested model for the proportion of RTs using the simplex distribution comparing the estimation parameters for the lognormal model is developed in this section. A second study for evaluating parameter recovery in the hierarchical limited response time model is also presented in this section.

6.1 Bounded Response Time (BRT) Model Regarding the Classical Lognormal Model

The first study has the purpose of knowing the performance of the simplex model in recovering parameters and comparing this with the lognormal model. In order to compare application data parameters, where response times are defined as values inside the interval $(0, 1)$, 50 replicated data sets from the simplex model were generated. Each one contains 30 items and 1000 examinees with fixed values for α and β . Those fixed values are similar with results of the application data using limited response times. τ values were drawn from a normal distribution with zero mean and standard deviation of 0.5. The replicated data sets were fitted with the simplex model.

In the case of lognormal model, 50 replicated data sets were generated. Values for α' and β' were fixed as similar values from results of the application data using response times in minutes. τ' values were drawn from a normal distribution with

Table 1 Mean of percentage of times where CI contains the parameter (PCI) in the simplex RT model and the lognormal RT model

model	Simplex		Lognormal	
	precision $\hat{\alpha}$	position $\hat{\beta}$	precision $\hat{\alpha}'$	position $\hat{\beta}'$
PCI	0.951	0.975	0.945	0.999
Time	41 min		7 min	

zero mean and standard deviation of 0.5. Each one of the replicated data sets was fitted with the lognormal model.

As the responses in both models have different scales, the percentage of times in which the Credible Interval (CI) of the estimated parameters contain the population parameter, were calculated for each model. Table 1 shows the mean of these percentages and also the mean time for fitting each model is presented. The posterior mean of position parameter β and of precision parameter α for each replica are shown, for the simplex model, in Fig. 2. In the case of the lognormal model, as parameters analogous to the ones of simplex model are shown in Fig. 3.

The simplex model has a good performance, suggesting the use of it as an alternative model for RTs.

6.2 *Parameter Recovery in the Hierarchical Joint Model for Bounded Response Time and Response Accuracy*

In order to know characteristics regarding parameter recovery of the Hierarchical joint model, 25 replicated data sets from this model were generated and fitted with it. Population values from item parameters $\mathbf{v} = (\log(\mathbf{a}), \mathbf{b}, \log(\boldsymbol{\alpha}), \boldsymbol{\beta})$ were drawn from a multivariate normal distribution with vector of means given by

$$\mu_{\mathbf{v}} = (-0.599, -0.616, -2.826, -2.060)$$

and covariance matrix $\Sigma_{\mathbf{v}}$; those values are fixed, saving comparisons with application data.

$$\Sigma_{\mathbf{v}} = \begin{bmatrix} 0.15 & 0 & 0 & 0 \\ 0 & 0.5 & -0.2 & 0.2 \\ 0 & -0.2 & 0.5 & 0 \\ 0 & 0.2 & 0 & 0.3 \end{bmatrix}$$

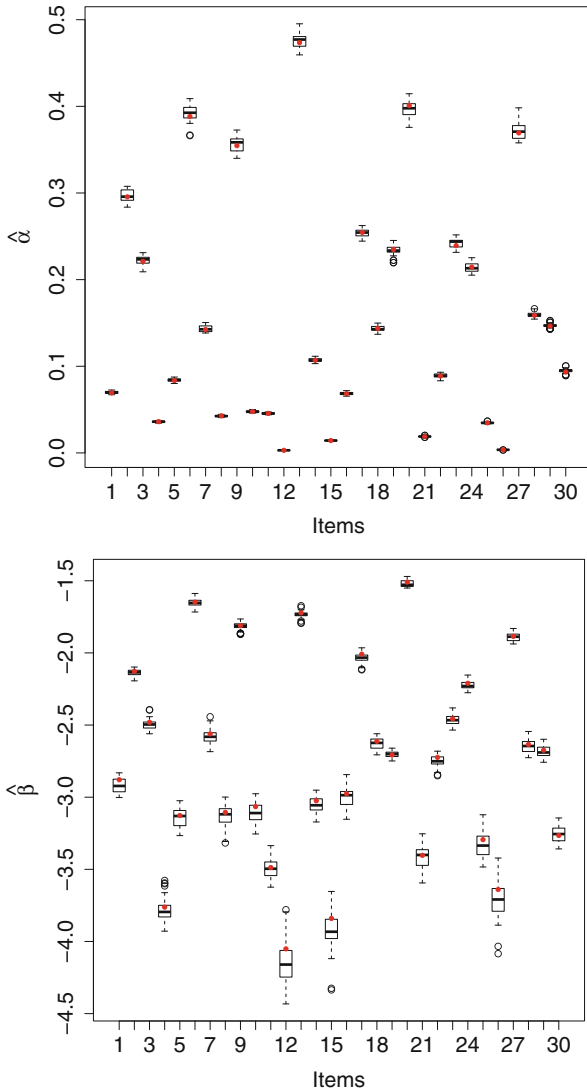


Fig. 2 Boxplots of the posterior mean of item estimates in the simplex RT model

In the case of parameters of examinees $\xi = (\theta, \tau)$, they were drawn from a bivariate normal distribution with vector zero mean and covariance matrix of Σ_{ξ}

$$\Sigma_{\xi} = \begin{bmatrix} 1 & 0.05 \\ 0.05 & 0.15 \end{bmatrix}$$

giving $\rho_{\theta\tau} = 0.05$ and $\sigma_{\tau}^2 = 0.15$.

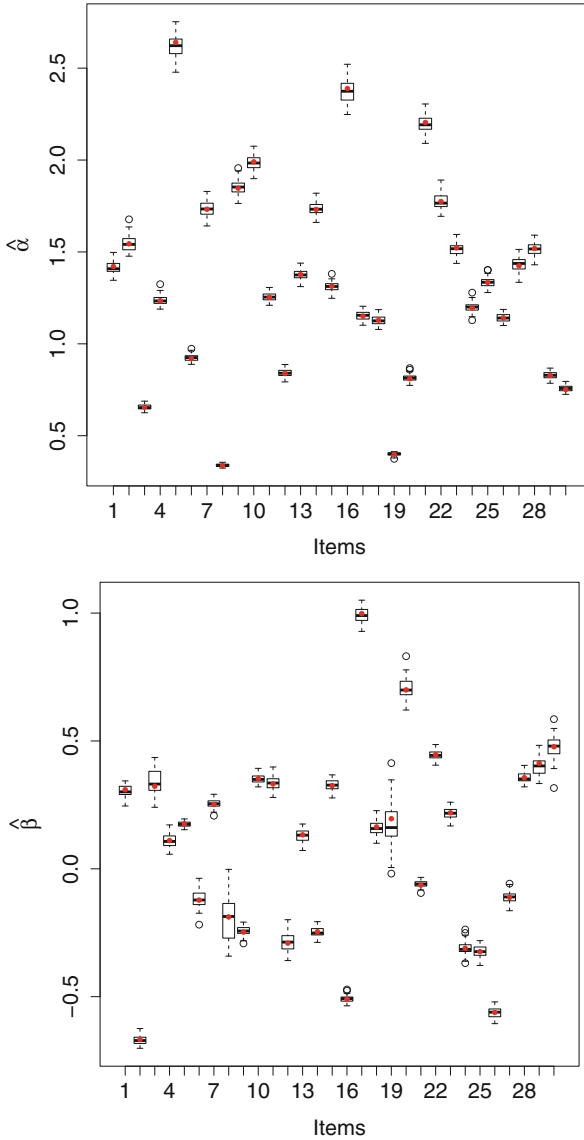


Fig. 3 Boxplots of the posterior mean of item estimates in the lognormal RT model

Results from this study arrive to the following posterior mean estimates: $\rho_{\hat{\theta}_\tau} = 0.05017$, $\sigma_{\hat{\tau}}^2 = 0.1919$. Regarding item parameters, Table 2 summarizes the root-mean-square error (RMSE) and the mean average error (MAE) for the posterior mean of item estimates.

Table 2 RMSE and MAE in parameter recovery for the HBRT model

	Parameter	RMSE	MAE
IRT part	a	0.1999	0.1983
	b	0.4612	0.4605
RT part	α	0.0009	0.0007
	β	0.0267	0.022

7 Application

In this section the model suggested in the previous section is applied using the PISA 2015 computer-based data for 28 reading items. Those items belong to the R1 and R3 clusters in bookid 37; see OECD (2017) page 38. Completed responses for clusters R1 and R3 were selected to avoid missing data. Each cluster has been designed to complete in 30 min, and after two clusters, a break is given for the students. Thus, in theory, the total time spent in the two clusters should not be more than an hour. Some cases which are higher than 70 min were considered as problems in the application and dropped from the analysis. A total of 4960 RTs and responses for 28 items from 53 countries or economies were used to apply the model.

The proportion of RT for an examinee i in the item j , z_{ij} , was calculated using the transformation suggested in Sect. 3, that is, $z_{ij} = \frac{t_{ij}}{d_j}$, where d_j is the greater time that an examinee spends in the item j . One valuable feature of this transformation is that it preserves the asymmetric behavior of the time, as is possible to see in the Fig. 4.

Three items, out of the 28, had partial credit. A transformation as correct and incorrect responses was used for those items, making partial credit an incorrect response.

The HBRT model proposed in Sect. 4 was fitted to this data following the estimation method described in Sect. 5. For this application, the implementation used 16,000 iterations; 4,000 iterations were as burning, and to avoid autocorrelation, a thin of four was selected. The posterior average of different item parameters regarding the IRT and RT parts of the model depicted in the Fig. 5, suggests interpretation of items' characteristics and can be used in posterior analysis. Descriptive analysis of posterior distributions of personal parameters θ_i and τ_i is not showed here.

8 Final Comments

In this work, the simplex distribution is used to model the proportion of response time as one alternative formulation to traditional RT models which assume that the time to answer a test is unlimited and then a lognormal distribution is used. This model's handling bounded RT performs satisfactorily. In addition, a hierarchical bounded response time (HBRT) model was also formulated. HBRT model uses RTs

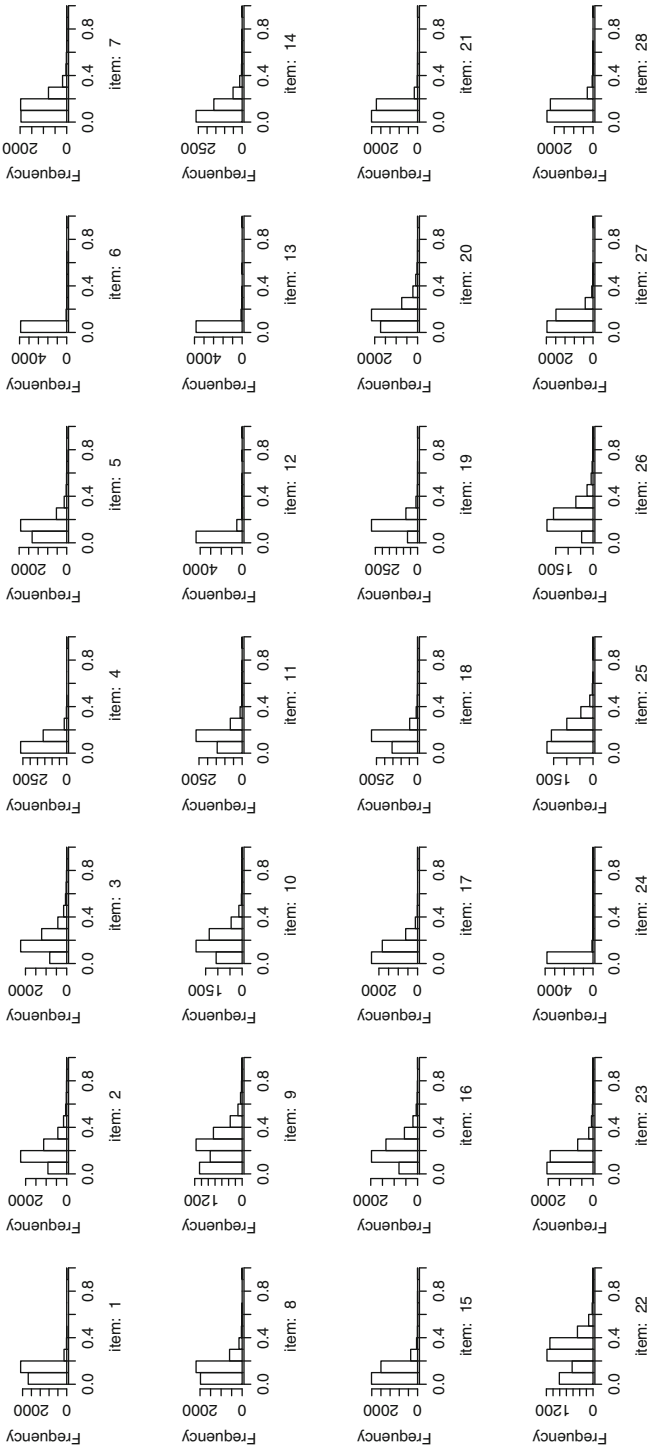


Fig. 4 Distribution of the proportion of RTs of examinees for 28 selected Reading items from PISA 2015

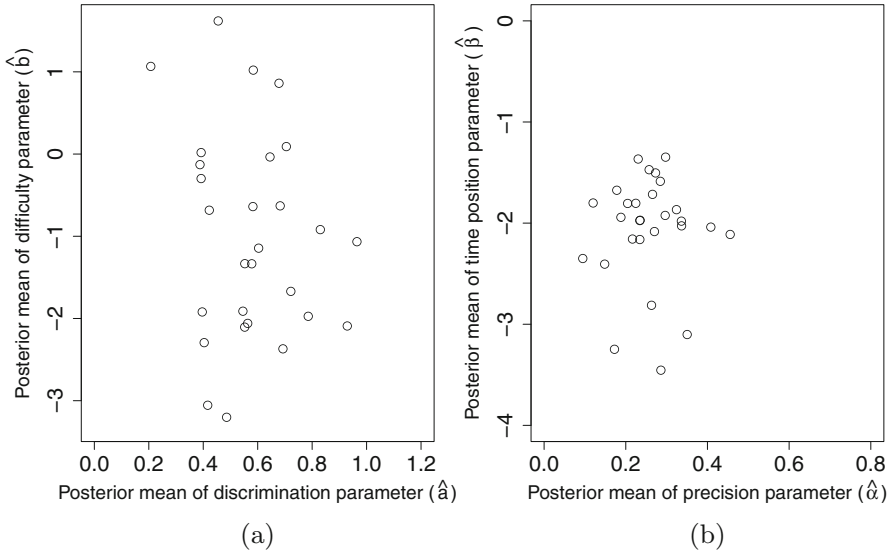


Fig. 5 (a) IRT item parameters and (b) Bounded RT item parameters for each one of the 28 selected Reading items from PISA 2015

and responses jointly, and a relation for the two constructs (accuracy and bounded speed) is also modeled.

Simulation studies suggest good performance of the HBRT model as an alternative hierarchical model proposed by van der Linden (2007). Also, the application developed shows several possibilities for the model proposed to Assessment. Future works could be developed using asymmetric distributions for accuracy and speed.

Appendix

This appendix has the BUGS code for the HBRT model. Since the list of distributions available in WinBUGS does not contain the simplex distribution, the “zero poisson” method was implemented in order to simulate from the simplex distribution as suggested in the WinBUGS manual (Spiegelhalter et al. 2003).

```
# Hierarchical joint simplex for bounded response
time and responses (HBRT) model
{
  for (j in 1:J) {
    psi[j,1:4] ~ dnorm(mu_i[],Omega_i[,])
    a[j] <- exp(psi[j,1])
    phi[j]<-exp(psi[j,3])
    alpha[j]<-sqrt(phi[j])
  }
}
```

```

}

  for (i in 1:I){
theta[i] ~ dnorm(0,1)

cm[i]<-rho*theta[i]
tau[i] ~ dnorm(cm[i],ctau)

for (j in 1:J){
  u[i,j] ~ dbern(p[i,j])
  m[i,j] <-a[j]*(theta[i]-psi[j,2])
  p[i,j] <- phi(m[i,j])

zeros[i,j]<-0
zeros[i,j]~dpois(lik[i,j])
lik[i,j]<- log(1000*3.141593) -.5*log(phi[j])+1.5*log
  (z[i,j]*(1-z[i,j]))+0.5*phi[j]*dev[i,j]
dev[i,j]<-pow((z[i,j]-mu[i,j]),2)/((z[i,j]*
  (1-z[i,j]))*
  pow((mu[i,j]*(1-mu[i,j])),2))
logit(mu[i,j])<-psi[j,4]-tau[i]
}
}

ctau ~ dgamma(0.1,0.1)
rho ~ dnorm(0,0.1)

  mu_i[1:4] ~ dmnorm(mm[],Omega_m[,])
  Omega_i[1:4,1:4] ~ dwish(Ri[,],4)
  Sigma_i[1:4,1:4] <- inverse(Omega_i[,])
  Omega_m[1:4,1:4] ~ dwish(Ri[,],4)
}

```

Acknowledgments This study was financed in part by the “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil” (CAPES) - Finance Code 001. The second author was partially supported by FAPESP-Brazil 2017/15452-5.

References

- Barndorff-Nielsen, O., & Jorgensen, B. (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis*, 39(1), 106–116.
- Fox, G. J., Klein Entink, R., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, 20(7), 1–14.

- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Raton: Chapman & Hall.
- Im, S. K. (2015). *The hierarchical testlet response time model: Bayesian analysis of a testlet model for item responses and response times*. Ph.D. thesis, University of Kansas.
- Jorgensen, B. (1997). *The theory of dispersion models*. Boca Raton: CRC Press.
- Klein Entink, R. H. (2009). *Statistical models for responses and response times*. Ph.D. thesis, University of Twente.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4), 325–337.
- OECD. (2017). *PISA 2015 Technical Report*. Technical report, OECD.
- Quintero, F. O. L. (2017). Sensitivity analysis for variance parameters in bayesian simplex mixed models for proportional data. *Communications in Statistics-Simulation and Computation*, 46(7), 5212–5228.
- Schnipke, D. L., & Scrams, D. J. (1999). *Representing response-time information in item banks. Law school admission council computerized testing report*. LSAC research report series. (LSAC-R-97-09).
- Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In *Computer-based testing: Building the foundation for future assessments* (pp. 237–266).
- Song, P. X.-K., & Tan, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics*, 56(2), 496–502.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS User Manual*. MRC Biostatistics Unit, Cambridge, UK.
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287.
- Zhan, P., Liao, M., & Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Frontiers in Psychology*, 9, 607.
- Zhang, P., Qiu, Z., & Shi, C. (2016). simplexreg: An r package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Software*, 71(1):1–21.

Selecting a Presmoothing Model in Kernel Equating



Gabriel Wallin and Marie Wiberg

Abstract In the kernel method of test score equating, the first step of the procedure is to presmooth the score distributions. The most common way of doing so is by fitting a log-linear model to the observed-score distributions. In this way, irregularities in the score distributions are smoothed, yielding a more stable estimated equating transformation. Within the kernel equating framework, an alternative way of presmoothing by using item response theory models has recently been suggested. There are furthermore several model selection criteria available for both of these classes of models. Here the model selection criteria are studied for both log-linear and item response theory models. Specifically, the likelihood ratio, AIC, and BIC measures are compared using real admissions data. Results show that the different model selection criteria result in equated scores that have real impact differences.

Keywords Model selection · Log-linear models · Item response theory

1 Introduction

Test score equating is a family of statistical models and methods that are used to make test scores comparable among different test versions. Post equating, the scores on these different test versions may be used interchangeably (González and Wiberg 2017). In test score equating in general and in the kernel equating (KE) framework (von Davier et al. 2004) in particular, a first step is typically to presmooth the score distributions with a statistical model to reduce irregularities before the score distributions are equated. The underlying assumption is thus that the

G. Wallin (✉)

Maasai research team and Laboratoire J.A. Dieudonné, Université Côte d'Azur, Inria, Sophia-Antipolis, France
e-mail: gabriel.wallin@inria.fr

M. Wiberg

Department of Statistics, USBE, Umeå University, Umeå, Sweden
e-mail: marie.wiberg@umu.se

population score distributions are smooth. Log-linear models are the most common way to presmooth test score distributions and have been shown to have positive effects on equating accuracy (Hanson 1991; Livingston 1993; Moses and Holland 2007; Moses and Liu 2011). Other statistical models which have been used for presmoothing in equating include the beta-4 model (Kim et al. 2005), the cubic B-spline, direct presmoothing (Cui and Kolen 2009), and item response theory (IRT) models (Andersson and Wiberg 2017). To choose a suitable model, different model selection criteria can be used; see Moses and Holland (2010a) for an evaluation for log-linear models. Further, Moses and Holland (2010b) studied the effects of the model selection criteria on traditional equipercentile equating which uses linear interpolation instead of kernel functions to create (piecewise) continuous functions.

The overall aim of this study is to examine the effect of model selection criteria for log-linear and IRT models in the presmoothing step within the KE framework and examine the model sensitivity in the equated scores. We will examine this effect using a real college admissions test for the non-equivalent groups with anchor test (NEAT) design. Specifically, we will study three of the most commonly used model selection criteria, namely, the likelihood-ratio chi-square statistic, the Akaike information criterion (AIC, Akaike 1981), and the Bayesian information criterion (BIC, Schwarz 1978). The study is limited to log-linear modeling and IRT modeling as these are currently implemented in the KE framework. This study is different from previous studies (e.g. Moses and Holland 2010a, b), as no previous study exist which evaluates model selection strategies for log-linear models and IRT models within the KE framework.

The rest of the paper is structured as follows. First an introduction to test score equating is given, followed by a brief description of KE. Next, brief presentations of log-linear and IRT presmoothing are given followed by an empirical study and its results. The paper ends with a discussion with some concluding remarks and practical recommendations.

2 Test Score Equating

Let the test scores on test forms X and Y be denoted by X and Y , respectively. Assume that X and Y are random variables from the populations \mathbf{P} and \mathbf{Q} . In the NEAT design, we assume that the test takers belong to different population, i.e., $\mathbf{P} \neq \mathbf{Q}$, and that we have access to a number of common anchor items, which can be used to compare the difficulty level of the test forms and the ability of the test takers. Let \mathbf{T} be the target population of the equating. We can then define the test score distribution of X and Y as $F_X(x) = \Pr(X \leq x | \mathbf{T})$ and $G_Y(y) = \Pr(Y \leq y | \mathbf{T})$. To find an equivalent test score y on test form Y for a test score x on test form X we assume that X and Y are continuous, so that we can use the equipercentile equating transformation

$$y = \varphi_Y(x) = G_Y^{-1}(F_X(x)). \quad (1)$$

2.1 Kernel Equating

KE is an observed-score equating framework which comprises five steps: (i) presmoothing the observed-score distributions, (ii) calculating the score probabilities, (iii) continuing the empirical score distributions, and (iv) equating and (v) computing of accuracy measures. Denote the unknown score probabilities in the target population \mathbf{T} by $r_j = \Pr(X = x_j | \mathbf{T})$, $j \in [1, J]$, and $s_k = \Pr(Y = y_k | \mathbf{T})$, $k \in [1, K]$. The estimated score probabilities $\hat{\mathbf{r}} = (\hat{r}_1, \dots, \hat{r}_J)^t$ and $\hat{\mathbf{s}} = (\hat{s}_1, \dots, \hat{s}_K)^t$ are obtained from the fitted values of the presmoothing model, via a design function that is determined by the data collection design. Let $\Phi(\cdot)$ be the standard normal distribution function and denote the bandwidth by h_X , which determines the smoothness of the function. Further, let $\hat{\mu}_X = \sum_j x_j \hat{r}_j$ and $\hat{\alpha}_X = \sqrt{\hat{\sigma}_X^2 / (\hat{\sigma}_X^2 + h_X)}$, where $\hat{\sigma}_X^2 = \sum_j (x_j - \hat{\mu}_X)^2 \hat{r}_j$. Then, the estimated continuized score distribution of X can be approximated by

$$\hat{F}_{h_X}(x) = \sum_{j=1}^J \left(\frac{x - \hat{\alpha}_X x_j - (1 - \hat{\alpha}_X) \hat{\mu}_X}{h_X \hat{\alpha}_X} \right).$$

The continuous score distribution \hat{G}_{h_Y} for the Y scores can be obtained similarly. For the rest of the paper, only operations connected to the X scores are shown as it is similarly for the Y scores. The continuous distributions are placed into Eq. 1 to obtain the equating transformation

$$\hat{\varphi}_Y(x) = \hat{G}_{h_Y}^{-1}(\hat{F}_{h_X}(x)) = G_{h_Y}^{-1}(F_{h_X}(x; \hat{\mathbf{r}}); \hat{\mathbf{s}}).$$

In the last step, different accuracy measures are examined – especially the standard error of equating (SEE; von Davier et al. 2004) defined as

$$SEE_Y(x) = \sqrt{\text{Var}(\hat{\varphi}_Y(x))}.$$

3 Presmoothing Options

3.1 Log-Linear Models

Log-linear models have been used extensively to estimate the r_j :s and s_k :s. Let n_j and m_k denote the number of test takers scoring $X = x_j$ and $Y = y_k$, respectively, with $\sum_j n_j = N$ and $\sum_k m_k = M$. Denote the probability vectors for n_1, \dots, n_J and m_1, \dots, m_K with \mathbf{p} and \mathbf{q} , respectively. Assume that $\mathbf{n} = (n_1, \dots, n_J)^t \sim \text{Multi-}$

nomial(N, \mathbf{p}) and $\mathbf{m} = (m_1, \dots, m_J)^t \sim \text{Multinomial}(M, \mathbf{q})$ and that \mathbf{n} and \mathbf{m} are independent. The log-likelihood function for the X scores can then be defined as

$$l_r = \sum_{j=1}^J n_j \log(r_j).$$

To estimate the score probabilities, we can use the log-linear model

$$\log(r_j) = \beta_0 + \sum_{i=1}^I \beta_i x_j^i + \sum_{h=1}^H \beta_{a,b} a_k^h + \sum_{d=1}^D \sum_{h=1}^H \beta_{a,b} a_k^h \beta_{x,a,d} x_j^d a_k^e, \quad (2)$$

where β_0 is a normalizing constant, β_i is a parameter to be estimated, and $x_j : s$ and $a_k : s$ are functions of the test scores and anchor scores, respectively. When the parameters of the log-linear model are estimated with maximum likelihood, the moments of the estimated distributions match those of the empirical distributions (Moses and Holland 2010a, b). This means that for the log-linear model in Eq. 2, I and H numbers of moments in the marginal distributions of X and A , respectively, are preserved, and D and E set the number of cross-moments in the joint distribution of X and A that are preserved. In practice, a Poisson regression model is typically used as the frequencies conditional on the sum of the frequencies follow a Poisson distribution.

3.2 Item Response Theory Models

IRT models can be used in the presmoothing step as an alternative to log-linear models, as score probabilities can be obtained by using the Lord and Wingersky (1984) algorithm. Denote test takers' latent ability parameter $\theta \in \{-\infty, \infty\}$, and denote the probability of a randomly chosen test taker answering item $l_X \in \{1, \dots, k_X\}$ from test form X with $P_{X_{l_X}}$ and likewise $P_{Y_{l_Y}}$ for item $l_Y \in \{1, \dots, k_Y\}$ from test form Y. If the three-parameter logistic (3PL) model is used, the probability of a randomly chosen test taker answering item l_X correctly is defined as

$$P_{X_{l_X}} = c_{l_X} + \frac{1 - c_{l_X}}{1 + \exp(-a_{l_X}(\theta - b_{l_X}))},$$

where $a_{l_X} \in [0, \infty\}$ is the discrimination of item l_X , $b_{l_X} \in \{-\infty, \infty\}$ is the difficulty of item l_X , and $c_{l_X} \in [0, 1]$ is the lower asymptote (guessing) parameter. If $c_{l_X} = 0$, we instead have the two-parameter logistic (2PL) model, and if, additionally, $a_{l_X} = 1$, we have the one-parameter logistic (1PL) model. In the empirical study of Sect. 5, the 1-PL, 2-PL, and 3-PL models are the candidate models considered for kernel equating using IRT.

4 Presmoothing Model Selection Criteria

For both classes of models, we will use the likelihood ratio test, the AIC, and BIC measure as model criteria when we choose the presmoothing model. This is due to the widespread use of each of these measures throughout statistics and psychometrics.

The likelihood-ratio chi-square statistic is asymptotically chi-square distributed and is defined as

$$W^2 = 2 \sum_j n_j \log \left(\frac{n_j}{Nr_j} \right),$$

the AIC is defined as

$$W^2 + 2(I + 1),$$

and the BIC is defined as

$$W^2 + [1 + \log N](I + 1).$$

It should be noted that the likelihood-ratio chi-square evaluates fit using significance testing. The procedure for model selection can thus be performed by selecting the simplest model with a nonsignificant chi-square statistic. The AIC and BIC measures on the other hand belong to the class of parsimonious model selection criteria. The general idea for these types of measures is to balance the fit of the model with the parameterization used to achieve that fit.

5 Empirical Study

In this section, the sensitivity of the equated values to changes of the presmoothing model is investigated using admission data from the Swedish Scholastic Aptitude Test (SweSAT). SweSAT is a paper and pencil test used in the application process to higher education in Sweden. Test results are valid for 5 years, and there is no upper limit on how many times one is allowed to take the test. It is always the best, valid result that is used when a test taker applies to a university program. SweSAT consists of two parts, and here a sample of the quantitative part of the spring administration of 2015 will be examined. Previous studies have shown that test groups of the SweSAT typically are non-equivalent (Lyrén and Hambleton 2011), meaning that the NEAT design is more appropriate than, for example, the equivalent group design. We will use the NEAT design with chained equating, meaning that the test forms are equated in a link, from the old test form via the anchor test to the new test form.

5.1 *Evaluation Measure and Study Design*

Within statistics and psychometrics research, it is a standard practice to evaluate an estimator by calculating its bias and mean squared error under a controlled setting such as a simulation study. This would require the possibility to define the true parameter value. For equating studies however, this is complex. This is partly due to the difficulty of generating simulated data that does not favor any particular method. Instead, equating-specific measures and summary indices have normally been employed. One example of the former is the difference that matters (DTM), defined as the difference between scale scores and equated scores that are larger than half a score unit (Dorans and Feigenbaum 1994), and an example of the latter is the discrepancy between the equating estimator and a reference equating transformation (e.g., Han et al. 1997). Recently, Wiberg and González (2016) proposed traditional statistical evaluation measures for the equating transformation such as bias, standard error, and mean squared error. This has partially inspired the method of evaluation in this study.

We follow the approach of Lord (1977) and Leôncio and Wiberg (2018) by splitting the X test form into two matrices of equal size. The original data contained 80 items and 40 anchor items given to 2826 test takers. After the split, we had two forms containing 80 items and 40 anchor items each, administered to 1413 test takers. One of the matrices will be treated as test form X, and the other matrix will be treated as test form Y. By doing so, the two test forms are equal, and so it follows that $\varphi(x) = x$. We will thus be able to calculate the error of the equating transformation, which we define as

$$e(x) = \hat{\varphi}(x) - \varphi(x). \quad (3)$$

The term $e(x)$ will reflect the error of each method, but not sampling variability since it is calculated using only one sample. This is the reason why we avoid calling $e(x)$ “bias.” Furthermore, the DTM and SEE will be presented.

5.2 *Model Selection*

For all models, a forward stepwise selection procedure was utilized for each model selection criterion. The three different criteria thus could select three potentially different parameterizations for each class of models and a total of $2 \times 3 = 6$ different presmoothing models. The three IRT models (1PL, 2PL, and 3PL) and the log-linear models were chosen from the selection criteria and can be seen in Table 1. The selected models were then plugged into the first step of the KE framework, yielding six different KE estimators.

5.3 Results

In Table 1, the selected models are presented for both model classes and for each model selection criterion. Only the highest power moment is presented, so, for example, X^6 means that the first six moments are included, and $X^2 : A^2$ means that $X : A, X^2 : A, X : A^2, X^2 : A^2$ are included. It is apparent that the AIC and likelihood-ratio chi-square criteria selected the same presmoothing model for both the log-linear and IRT models. These two cases will thus result in the same KE estimator. Table 1 also shows that the BIC criterion selected a more parsimonious model for both the log-linear and IRT model compared to the AIC and Likelihood-ratio chi-square criterion.

In Fig. 1, the error, as defined in Eq. 3, is plotted for each KE estimator. As the AIC and likelihood-ratio chi-square criteria selected the same presmoothing model for both the log-linear and the IRT model, they are represented by the same respective lines. What is noteworthy is that the two classes of models follow each other closely in terms of equated scores; the models are homogeneous within its presmoothing model class and heterogeneous between classes. It is also apparent that the KE estimators using IRT presmoothing models exhibit a much smaller error, especially in the low and high end of the score scale.

In Fig. 2, the SEE for each KE estimator is displayed. As for the error, the KE estimators based on log-linear models yield very similar results. The IRT KE estimators are not equally similar, and the IRT KE estimator using the BIC criterion shows the by far lowest SEE among the whole score scale.

Figure 3 shows the difference, in equated scores, between the KE estimator using a log-linear model selected by the AIC/likelihood-ratio chi-square criterion and every other KE estimator. The solid black lines represent the DTM. The differences, with exception of the KE estimator using a log-linear model selected by the BIC

Table 1 Log-linear and IRT model selected by each respective criterion

Model	AIC	BIC	χ^2
Log-linear	$X^6, A^4, X^2 : A^2$	$X^6, A^4, X : A$	$X^6, A^4, X^2 : A^2$
IRT	3PL	2PL	3PL

Fig. 1 The error of each respective KE estimator

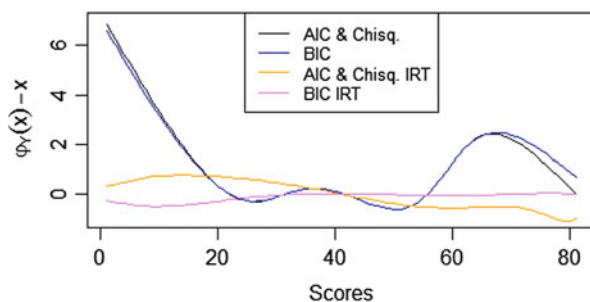


Fig. 2 The SEE of each respective KE estimator

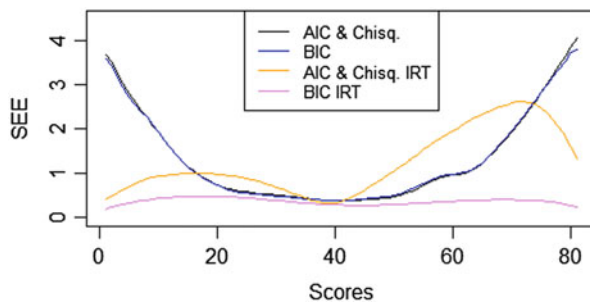
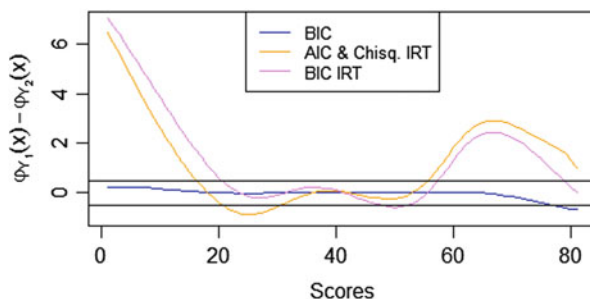


Fig. 3 The difference in equated scores between the KE estimator using the AIC (and likelihood-ratio chi-square) model selection criterion and each other KE estimator



criterion, exceed the DTM bounds for large parts of the score scale, meaning that the different KE estimators result in practically significant differences for the test takers.

6 Conclusions

This study has evaluated different model selection criteria for log-linear and IRT models for the presmoothing of test score distributions with the purpose of equating test forms. The comparison between the criteria was made under the KE framework and using real admission data. This study was motivated by the fact that it is important to compare several equatings to see how sensitive the equated scores are. The results showed that depending on the model selection criteria (likelihood ratio/AIC/BIC) and depending on the presmoothing model (log-linear/IRT), the resulting equated scores can differ to a degree where it will have real life implications for the test takers. The Empirical study suggests that when IRT is used in the presmoothing step, the BIC measure is preferred as it produces the lowest error. If log-linear models instead are used to presmooth the score distributions, there is practically no difference between the BIC measure and the AIC/likelihood ratio measures in terms of equated scores. These conclusions should however be taken with a large portion of caution as they are only based on an empirical study. This study thus motivates why future research should conduct a

rigorous simulation study where model selection criteria is further investigated. In such a study, it would be possible to get further insight on when each criterion works best, by, e.g., varying test length, sample size, and data collection design. The last point should also be mentioned as a limitation of this study, as only the NEAT design has been considered. As previous studies have only focused on the estimation score distributions, or on traditional equipercentile equating, future research should investigate these issues for the KE framework, considering more data collection designs and test scenarios.

Acknowledgments The research was funded by the Swedish Research Council grant number 2014-578.

References

- Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*, 3–14.
- Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika*, *82*, 48–66.
- Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic B-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement*, *46*, 135–158.
- Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT (Research Memorandum No. RM-94-10)*. Princeton: Educational Testing Service.
- González, J., & Wiberg, M. (2017). *Applying test equating methods – Using R*. Cham: Springer.
- Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true- and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, *10*, 105–121.
- Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement*, *15*, 391–408.
- Kim, D.-I., Brennan, R., & Kolen, M. (2005). A comparison of IRT equating and beta 4 equating. *Journal of Educational Measurement*, *42*, 77–99.
- Leôncio, A., & Wiberg, M. (2018). Evaluating equating transformations from different frameworks. In M. Wiberg, S. A. Culpepper, R. Jansen, J. González, & D. Molenaar (Eds.), *Quantitative psychology – 82nd annual meeting of the psychometric society, Zurich, Switzerland, 2017* (pp. 101–110). New York: Springer.
- Livingston, S. (1993). Small-sample equatings with log-linear smoothing. *Journal of Educational Measurement*, *30*, 23–39.
- Lord, F. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, *14*, 117–138.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings”. *Applied Psychological Measurement*, *8*, 452–461.
- Lyrén, P.-E., & Hambleton, R. K. (2011). Consequences of violated the equating assumptions under the equivalent group design. *International Journal of Testing*, *36*, 308–323.
- Moses, T., & Holland, P. (2007). *Kernel and traditional equipercentile equating with degrees of presmoothing*. ETS research report RR-07-15.

- Moses, T., & Holland, P. (2010a). A comparison of statistical selection strategies for univariate and bivariate log-linear models. *British Journal of Mathematical and Statistical Psychology*, *63*, 557–574.
- Moses, T., & Holland, P. (2010b). Selection strategies for univariate loglinear smoothing models and their effect on equating function accuracy. *Journal of Educational Measurement*, *46*, 159–176.
- Moses, T., & Liu, J. (2011). *Smoothing and equating methods applied to different types of test score distributions and evaluated with respect to multiple equating criteria*. ETS research report RR-11-20.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.
- Wiberg, M., & González, J. (2016). Statistical assessment of estimated transformations in observed-score equating. *Journal of Educational Measurement*, *53*, 106–125.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.

Practical Implementation of Test Equating Using R



Marie Wiberg and Jorge González

Abstract Test equating methods are widely used in order to make comparable different test forms administered at different occasions to different test takers. Although software for test equating is currently available, in this paper we focus the attention on four different R packages which can facilitate test equating for researchers and test developers. This paper list the different R packages which are available at the moment. Examples are provided for the `equate`, `equateIRT`, `kequate`, and the `SNSEquate` packages. Additional features of these packages are discussed as well.

Keywords R packages for test equating · Equate · EquateIRT · Kequate · SNSEquate

Test equating is a statistical process used to transform scores on two test forms so that they are placed on a common scale and are thus comparable (González and Wiberg 2017). Let x and y be the quantiles in the cumulative distribution functions (CDF) F_X and F_Y of test scores X and Y , respectively. Then an equivalent score y on test Y for a score x on test X can be obtained using the so-called equipercentile transformation φ (Braun and Holland 1982) which is defined as

$$y = \varphi(x) = F_Y^{-1}(F_X(x)). \quad (1)$$

Note, linear versions of the equipercentile transformation (1) can also be used for equating. In test equating, different data collection designs are used to obtain test

M. Wiberg (✉)

Department of Statistics, USBE, Umeå University, Umeå, Sweden

e-mail: marie.wiberg@umu.se

J. González

Facultad de Matematicas, Pontificia Universidad Católica de Chile, Santiago, Chile

e-mail: jorge.gonzalez@mat.uc.cl

score data. The data collection designs depend on how and to whom the test forms are distributed. To be able to make score scales comparable, one needs some kind of common or equivalent information such as common or equivalent test takers, common items, or common background information about the test takers. Data collection designs described in the literature include the single group (SG) design, counterbalanced (CB) design, equivalent groups (EG) design, nonequivalent groups with anchor test (NEAT) design, and the nonequivalent groups with covariates (NEC) design. Independent of the design selected for the administration of the tests, the observed data can be accommodated in a matrix in which each row contains the response pattern of each test taker. Scores can then be obtained either as an estimation of a person parameter by fitting an item response theory (IRT) model or as observed scores, most often, sum scores, from the data matrix.

Test equating methods can be categorized in several ways. In this paper we make the distinction between four groups: (i) IRT equating methods (Lord 1980), which are based on IRT models; (ii) traditional equating methods (Kolen and Brennan 2014) which include linear and equipercentile equating; (iii) kernel equating (KE) methods (von Davier et al. 2004), which use a five-step approach in which kernels are used for the continuous approximations of the score CDFs in (1); and (iv) IRT KE methods, which are a combination of both (i) and (iii). Within the first group, IRT true-score equating (IRTTSE) and IRT observed-score equating (IRTOSE) (Lord 1980; Lord and Wingersky 1984) have been two widely used methods for test equating. Both are based on conditional distributions of tests scores given the ability (González et al. 2016). While in the former the equating transformation is based on the mean of the conditional score distributions (the true score), the latter is based on the marginal score distributions and uses the IRT model to define the conditional score probabilities involved in the computation of φ . In addition, IRT equating requires IRT parameter linking (von Davier and von Davier 2011) as a necessary preliminary step to correct for differences on the scales of the parameters.

The equating methods categorized in these four groups are currently implemented in several R packages, and we will summarize their features here and give a couple of short examples.

1 Equating Test Scores with R

Table 1 shows a list of several packages which are available to perform test equating using R, including `irtoys` (Partchev 2014), `lordif` (Choi et al. 2011), `sirt` (Robitzsch 2016), `plink` (Weeks 2010), `equateIRT` (Battaaz 2015), `equateMultiple` (Battaaz 2017), `equate` (Albano 2016), `kequate` (Andersson et al. 2013), and `SNSequate` (González 2014). We will shortly review the most commonly used packages: `equate`, `equateIRT`, `kequate` and `SNSequate`. We will also give example codes by using data files described in González and Wiberg (2017). More specifically, we will use data from a college admissions test given twice a year. The admissions test is composed of a verbal

Table 1 Equating methods implemented in different R packages

R package	IRT equating			Traditional equating		KE	IRT KE
	True-score	Observed -score	parameter linking	Equipercentile	linear		
irtoys			✓				
lordif			✓				
sirt			✓				
plink	✓	✓	✓				
equateIRT	✓	✓	✓				
equateMultiple	✓	✓	✓				
equate				✓	✓		
kequate						✓	✓
SNSEquate	✓	✓	✓	✓	✓	✓	

and a quantitative section, each containing 80 multiple-choice binary scored items. The sections are equated separately. For the EG design, we will use 2 samples of 10,000 test takers who took 2 different administrations of the quantitative section, and they are stored in the ADM1 and ADM2 data sets. For the NEAT design, we will use 2 samples of 2,000 test takers who took 2 different administrations of the verbal section of the admissions test and whose results are stored in the ADMneatX and ADMneatY data sets which additionally contains information for a 40-item verbal anchor test. The data can be obtained from the following website <http://www.mat.uc.cl/~jorge.gonzalez/EquatingRbook>.

1.1 Traditional Methods

Traditional methods utilize two types of equating transformations: equipercentile and linear equating, where mean equating is a particular case of linear equating. Under the NEAT design, different methods have been developed including Tucker, Levine observed-score, Levine true-score, Braun-Holland, Nominal weights, Chained equating, Frequency estimation, etc. (Kolen and Brennan 2014). Most of the traditional equating methods are included in the package `equate` (Albano 2016).

To perform traditional equating methods in R, the following code can be used for loading the data, creating score vectors (`quant.x`, `quant.y`) and creating score frequency distributions using the function `frequstab()`. Then, we can perform mean, linear, and equipercentile equating using the `equate()` function:

```
> load(url("http://www.mat.uc.cl/~jorge.gonzalez/
+ EquatingRbook/ADM1.Rda"))
> load(url("http://www.mat.uc.cl/~jorge.gonzalez/
+ EquatingRbook/ADM2.Rda"))
```

```

> quant.x <- apply(ADM2[,1:80],1,sum)
> quant.y <- apply(ADM1[,1:80],1,sum)
> library(equate)
> egADM.x<-freqtab(quant.x,0:80)
> egADM.y<-freqtab(quant.y,0:80)
> eq.mean <- equate(egADM.x, egADM.y, type = "mean")
> eq.linear <- equate(egADM.x, egADM.y,
+ type = "linear")
> eq.equipercent <- equate(egADM.x, egADM.y,
+ type = "equipercentile")
> eq.equipercent

```

The output for the equipercentile equating is as follows:

Equippercentile Equating: egADM.x to egADM.y

Design: equivalent groups

Smoothing Method: none

Summary Statistics:

	mean	sd	skew	kurt	min	max	n
x	38.32	12.72	0.34	2.47	10	77.00	10000
y	36.70	12.25	0.57	2.82	7	79.00	10000
yx	36.70	12.25	0.57	2.82	8	78.75	10000

The output states that we have performed equating under the EG design, and it shows summary statistics for the observed scores in tests X and Y (first and second row) and also for the equated scores (third row). The actual equated values are stored in the equating object `eq.equipercent` and can be obtained typing `eq.equipercent$concordance`. The `equate` package can handle the SG, NEAT, and EG designs. Other features in this package include nominal weighting, log-linear presmoothing, and bootstrap standard errors.

1.2 IRT Equating Methods

IRT parameter linking must be conducted when equating is performed under the NEAT design or when different scaling conventions for ability are used under other designs. IRT parameter linking is implemented in the packages `equateIRT` and `SNSequate` through the functions `direct()` and `irt.link()`, respectively. If sum scores rather than IRT scores are to be reported, then they can be equated using either IRT true-score or IRT observed-score equating (Lord 1980). These methods are implemented through the functions `irt.eq()` and `score()` in the `SNSequate` and `equateIRT` packages, respectively.

The following code is used for loading the data, fits a three parameter logistic model (3PL) to both X and Y test score data using the `tpm()` function from the `ltm` package (Rizopoulos 2006), and extracts the item parameter estimates from both runs:

```
> load(url("http://www.mat.uc.cl/~jorge.gonzalez/
+ EquatingRbook/ADMneatX.Rda"))
> load(url("http://www.mat.uc.cl/~jorge.gonzalez/
+ EquatingRbook/ADMneatY.Rda"))

> library(ltm)
> mod3pl.x<-tpm(ADMneatX)
> mod3pl.y<-tpm(ADMneatY)

> a.x<-coef(mod3pl.x) [,3]
> b.x<-coef(mod3pl.x) [,2]
> c.x<-coef(mod3pl.x) [,1]

> a.y<-coef(mod3pl.y) [,3]
> b.y<-coef(mod3pl.y) [,2]
> c.y<-coef(mod3pl.y) [,1]

> param_x <- list(a=a.x,b=b.x,c=c.x)
> param_y <- list(a=a.y,b=b.y,c=c.y)
> parm = as.data.frame(cbind(a.y,b.y,c.y,
+ a.x,b.x,c.x))
```

1.2.1 IRT Parameter Linking

The following code can be used to perform IRT parameter linking using `SNSequate`:

```
> library(SNSequate)
> comitems = 1:40
> irt.link(parm, comitems, model = "3PL",
+ icc = "logistic", D = 1.7)
```

Call:

```
irt.link.default(parm = parm, common = comitems,
  model = "3PL",
  icc = "logistic", D = 1.7)
```

IRT parameter-linking constants:

	A	B
Mean-Mean	0.8960919	0.093921946
Mean-Sigma	0.8722422	0.111956630
Haebara	0.9022729	0.050350133
Stocking-Lord	0.9644684	0.002433244

The `irt.link()` function receives as inputs a list containing the item parameter estimates, a vector indicating which are the common items, the IRT model from which the item parameters were obtained, and the value of the scaling constant D . It returns the equating coefficients A and B calculated using four IRT parameter-linking methods: mean-mean, mean-sigma, Haebara, and Stocking-Lord methods. The first two methods are based on means and standard deviations of the parameter estimates for the common items, and the last two are based on the item characteristic curves defined by the IRT model used. For details about the different linking methods, refer, for example, to Sect. 6.2 in Kolen and Brennan (2014).

IRT parameter linking can be performed in `equateIRT` using the mean-mean method by writing the following code:

```
> library(equateIRT)
> est3pl.x <- import.ltm(mod3pl.x, display=FALSE)
> est3pl.y <- import.ltm(mod3pl.y, display=FALSE)

> dimnames(est3pl.x$coef) [[1]] [1:40] <- paste
+   ("it", 1:40, sep="")
> dimnames(est3pl.y$coef) [[1]] [1:40] <- paste
+   ("it", 1:40, sep="")

> p12 <- modIRT(coef=list(est3pl.x$coef,
+ est3pl.y$coef), var=list(est3pl.x$var, est3pl.y$var),
+ ltparam=TRUE, display=FALSE)

> p12mm <- direc(mods=p12, which=c(1,2),
+   method="mean-mean", D=1.7)
> summary(p12mm)
```

Link: T1.T1

Method: mean-mean

Equating coefficients:

	Estimate	StdErr
A	0.896092	0.037336
B	0.093922	0.095740

The output shows that the mean-mean method was used, and it gives the estimates and standard errors of the equating constants. Other options for the method argument are mean-gmean, mean-sigma, Haebara, and Stocking-Lord.

1.2.2 IRT True-Score and Observed-Score Equating

To perform IRT equating in `SNSequate`, we can use the following code, where the first two lines are used to perform IRT true-score equating and the second two lines are used to perform IRT observed-score equating:

```
> res.3pl.tse<-irt.eq(120, param_x, param_y,
+ method="TS",method_link="mean/mean", common=1:40)

> res.3pl.ose<-irt.eq(120, param_x, param_y,
+ method="OS",method_link="mean/mean", common=1:40)

> outirt.3pl <- cbind(
+ Theta=res.3pl.tse$theta_equivalent, Scale=0:120,
+ IRTTSE=res.3pl.tse$tau_y, IRTOSE=res.3pl.ose$e_Y_x)

> outirt.3pl[26:28,]
      Theta Scale  IRTTSE  IRTOSE
[1,] -1.895891   25 27.13547 27.64139
[2,] -1.655602   26 28.57855 28.74090
[3,] -1.469780   27 29.90523 29.84072
```

We summarize the output for both methods in the object `outirt.3pl` and show the results only for the three score values 25, 26, and 27. The first column of the output, labeled `Theta`, shows the value of the ability for a given score on the scale (column `Scale`), together with the IRT true-score (column `IRTTSE`) and IRT observed-score (column `IRTOSE`) equated values.

To perform IRT true-score and IRT observed-score equating in `equateIRT`, we can use the following code which gives the subsequent outputs:

```
> score(p12mm, method = "TSE",se=FALSE,scores = 25:27)
      theta T2  T1.as.T2
1 -3.636381 25 23.94389
2 -3.211220 26 24.66996
3 -2.882307 27 25.42281

> score(p12mm, method = "OSE",se=FALSE,scores = 25:27)
      T2  T1.as.T2
26 25 22.99109
27 26 23.91655
28 27 24.84573
```

In the first output, the column labeled `theta` displays the value of the ability for a given score on the scale for scores 25 to 27 (column T2) and the corresponding equated values using IRTTSE (column T1 . as . T2). Similarly, the second output shows the equated values using the IRTOSE (column T1 . as . T2) for the scores 25 to 27 (column T2).

1.3 Kernel Equating Methods

KE has traditionally been presented as a method comprising the following five steps: (i) presmooth the score distributions, (ii) estimate the score probabilities, (iii) continuize the discrete score distributions, (iv) perform the actual equating, and (v) evaluate the equating transformation. KE can be performed using either the package `kequate` (Andersson et al. 2013) or `SNSequate` (González 2014). Although there is some overlap between these packages, they also contain some unique features such as the choice of the kernel for continuization, the method to select the bandwidth parameters, and the choice of the presmoothing model.

1.3.1 Kernel Equating with Kequate

To perform KE with `kequate`, after loading the data, the function `kefreq()` is used to obtain score frequency distributions. Under the NEAT design, two samples of test takers from two different populations, denoted here as P and Q , are administered test forms X and Y , respectively, and, additionally, a set of common items in form A is administered to both samples. Thus, we first find the sum scores from test forms X and Y and store them in the objects `verb.x` and `verb.y`, respectively. The sum scores from the common item anchor test A are stored in the object `verb.xa` for those test takers that were administered test form X and in `verb.ya` for those taking test form Y . The following code can be used for loading the data, constructing sum score vectors, and sorting the data:

```
> library(kequate)
> verb.xa <- apply(ADMneatX[,1:40],1,sum)
> verb.x <- apply(ADMneatX[,41:120],1,sum)
> verb.ya <- apply(ADMneatX[,1:40],1,sum)
> verb.y <- apply(ADMneatY[,41:120],1,sum)
> neatk.x <- kefreq(verb.x, 0:80, verb.xa, 0:40)
> neatk.y <- kefreq(verb.y, 0:80, verb.ya, 0:40)
```

The regular `glm()` function in R can be used to presmooth the score data by fitting appropriate log-linear models. For example, assume that we fit log-linear models to the test scores X and A as follows:

$$\log(p_{jl}) = \beta_0 + \sum_{i=1}^{T_r=2} \beta_i^X (x_j)^i + \sum_{i=1}^{T_s=1} \beta_i^A (a_l)^i + \beta_{il}^{XA} x_j^i a_l^i, \tag{2}$$

where $p_{jl} = \Pr(X = x_j, A = a_l)$. The same or a similar model can be fitted for the Y scores. Once we have models for X and Y , equating can be performed. For the NEAT design, we can choose to either use chained equating, in which case the equating is performed from test form X to the anchor test A and then to the test form Y , or we can perform a poststratification equating, in which case we assume a target population T , where $T = wP + (1 - w)Q$ and $0 \leq w \leq 1$ is a weight. We can perform poststratification equating (`neatPSEadm`) or chained equating (`neatCEadm`) by simply writing one line of code. The following code can be used for fitting two log-linear models for test scores from test forms X and Y , respectively, and perform poststratification and chained KE in R with `kequate`:

```
> NEATvX <- glm(frequency~I(X)+I(X^2)+I(A)+I(X):I(A),
+ family = "poisson", data = neatk.x, x = TRUE)
> NEATvY <- glm(frequency~I(X)+I(X^2)+I(A)+I(X):I(A),
+ family = "poisson", data = neatk.y, x = TRUE)
> neatPSEadm <- kequate("NEAT_PSE",0:80,0:80,NEATvX,
+ NEATvY)
> neatCEadm <- kequate("NEAT_CE",0:80,0:80,0:40,
+ NEATvX, NEATvY)
```

Using the `summary()` function on the chained KE object `neatCEadm` in the last line of the code, we get the following output when excluding equated scores for test scores 2–78:

```
Design: NEAT CE equipercentile

Kernel: gaussian

Sample Sizes:
  Test X: 2000
  Test Y: 2000

Score Ranges:
  Test X:
    Min = 0 Max = 80
  Test Y:
    Min = 0 Max = 80
  Test A:
    Min = 0 Max = 40
```

```

Bandwidths Used:
      hxP      hyQ      haP      haQ      hxPlin
1 0.7076599 0.6901055 0.4791841 0.4791372 12896.49
      hyQlin      haPlin      haQlin
13333.52 12374.88 1 12403.57

Equating Function and Standard Errors:
      Score      eqYx      SEEYx
1      0 -0.07056532 0.1470844
2      1  0.86454785 0.2657626
-
80     79 79.53264834 0.1893297
81     80 80.29402720 0.1126502

```

```

Comparing the Moments:
      PREAx      PREYa
1  0.0024998851  0.07006644
2  0.0054265212 -0.12833451
3  0.0009268535 -0.37683680
4  0.0136679929 -0.75152590
5  0.0410000237 -1.32311537
6  0.0823055646 -2.14353840
7  0.1373475690 -3.25047851
8  0.2060684759 -4.66835536
9  0.2884883735 -6.40798466
10 0.3846608957 -8.46680452

```

The output shows that the default Gaussian kernel was used and both test forms had 2,000 test takers. Also the score ranges and the used bandwidths are provided. Finally the equated scores and the standard error of equating (SEE) are given for each score followed by the percent relative error (PRE) of the first ten moments.

The `kequate` package can handle all the commonly used equating designs; SG, CB, EG, NEAT, and in addition the NEC design as described in Wiberg and Bränberg (2015) and Wallin and Wiberg (2019). As the `kequate` package reads in log-linear models using `glm` objects, one can build a model as complicated as one prefers. Besides the default penalty method, one can select other methods to choose the bandwidth parameters such as cross-validation (CV) (Wallin et al. 2018) and double-smoothing (DS) (Hägström and Wiberg 2014). Also different kernels can be used besides the default Gaussian kernel (e.g., `uniform` and `logistic`). Get commands can be applied to the equating objects to obtain commonly used evaluation measures such as PRE (`getPre()`), SEE (`getSee()`), equated values (`getEq()`), comparison of SEE (`genSeed()`), etc. Another interesting feature is the possibility of, instead of using log-linear models in the presmoothing step, using IRT models to conduct IRT KE as described in Andersson and Wiberg (2017).

1.3.2 Kernel Equating with SNSequate

Different functions are available to carry out the five steps in the KE method including `loglin.smooth()` (presmoothing), `bandwidth()` (bandwidth selection), `ker.eq()` (KE), `PREP()` (assessment through the percent relative error), and `SEED()` (SEE and SEE differences) functions. The following code is used for running KE under the EG design when a polynomial log-linear model of degree 2 for the score variables X and Y is used for presmoothing (`degree = c(2, 2)`), when the bandwidth parameters are automatically calculated using the penalty method (`hx = NULL, hy = NULL`), and when the Gaussian kernel is used for continuization of the CDFs (`kert = "gauss"`):

```
> ker.ADM<-ker.eq(scores = cbind(egADM.x, egADM.y) ,
+ kert = "gauss", hx = NULL, hy = NULL,
+ degree = c(2, 2), design = "EG")
```

The output shows summary statistics for the observed scores, the estimated bandwidth parameters, and the equated scores together with the SEE in both directions (i.e., X to Y in `eqYx` and `SEYx`; and Y to X in `eqXy` and `SEEXy`):

```
> summary(ker.ADM)
```

Call:

```
ker.eq.default(scores = cbind(egADM.x, egADM.y) ,
  kert = "gauss",
  hx = NULL, hy = NULL, degree = c(2, 2),
  design = "EG")
```

Summary statistics

	X	Y
Total	10000.0000	10000.0000
Mean	38.3189	36.7006
SD	12.7181	12.2498
Skewness	0.3393	0.5680
Kurtosis	2.4713	2.8189

Bandwidth parameters used

	hx	hy
1	0.700719	0.7012886

Kernel type used

```
[1] "Gaussian"
```

Equated values and SEE under the EG design

Score		eqYx	eqXy	SEEXx	SEEXy
1	0	-0.05213171	0.05424558	0.06203064	0.06693562
2	1	0.89803256	1.10823013	0.11182611	0.12493300
3	2	1.84024486	2.16927630	0.16200559	0.17938774
.....					
33	32	30.61294861	33.44205773	0.18591825	0.18884808
34	33	31.57480377	34.48168516	0.18285541	0.18627619
35	34	32.53670047	35.52128440	0.18025731	0.18426909
.....					
78	77	74.91704424	78.55757083	0.31354605	0.17526158
79	78	76.21407883	79.21108155	0.28279095	0.13123453
80	79	77.66694896	79.82457171	0.22696717	0.09066394

Other supported options for the argument `kert` are "gauss," "logis," "uniform," "epan", and "adap" for the gaussian, logistic, uniform, Epanechnikov, and adaptive kernels, respectively. The equating designs implemented in `SNSequate` are the EG, SG, CB, and NEAT designs.

2 Discussion

In this paper we have listed current R packages to perform traditional equating, KE, IRT, and KE IRT methods. Short illustrations were given for the EG and NEAT designs when using the packages `equate`, `kequate`, `equateIRT`, and `SNSequate`. More extensive examples covering all equating methods and exploring more features of these packages are shown in González and Wiberg (2017).

In the future we believe that test equating will continue to play a crucial role in standardized achievement tests. There will be several challenges, especially if machine learning and artificial intelligence are being implemented as part of admission processes. This will require the development of new methods to ascertain that selected candidates' knowledge can be compared to each other in a valid and reliable way. This means that it will be essential to create new packages or extend the current ones to include equating methods that account for these new technologies.

Acknowledgments This research was partially funded by the Swedish Research Council grant number 2014-578.

References

- Albano, A. D. (2016). `Equate`: An R package for observed-score linking and equating. *Journal of Statistical Software*, 74(8), 1–36.
- Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika*, 82(1), 48–66.

- Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, 55(6), 1–25.
- Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software*, 68(7), 1–22.
- Battauz, M. (2017). equateMultiple: Equating of multiple forms using item response theory methods. <https://cran.r-project.org/web/packages/equateMultiple/index.html>, R Package Version 0.0.0.
- Braun, H., & Holland, P. (1982). Observed-score test equating: A mathematical analysis of some ets equating procedures. In: P. Holland & D. Rubin (Eds.), *Test equating* (Vol. 1, pp. 9–49). New York: Academic Press.
- Choi, S., Gibbons, L., & Crane, P. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of Statistical Software*, 39(8), 1–30.
- von Davier, A. A., Holland, P., & Thayer, D. (2004). *The kernel method of test equating*. New York: Springer.
- von Davier, M., & von Davier, A. (2011). A general model for IRT scale linking and scale transformations. In A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (Vol. 1, pp. 225–242). New York: Springer.
- González, J. (2014). SNSequate: Standard and nonstandard statistical models and methods for test equating. *Journal of Statistical Software*, 59(7), 1–30.
- González, J., & Wiberg, M. (2017). Applying test equating methods using R. New York: Springer.
- González, J., Wiberg, M., & von Davier, A. A. (2016). A note on the Poisson's binomial distribution in item response theory. *Applied Psychological Measurement*, 40(4), 302–310.
- Häggeström, J., & Wiberg, M. (2014). Optimal bandwidth selection in observed-score kernel equating. *Journal of Educational Measurement*, 51(2), 201–211.
- Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates.
- Lord, F., & Wingersky, M. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings”. *Applied Psychological Measurement*, 8(4), 453–461.
- Partchev, I. (2014). Irtoys: Simple interface to the estimation and plotting of IRT models. <http://CRAN.R-project.org/package=irtoys>, R Package Version 0.1.7.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25.
- Robitzsch, A. (2016). sirt: Supplementary item response theory models. <https://cran.r-project.org/web/packages/sirt/index.html>, R Package Version 1.12.2.
- Wallin, G., & Wiberg, M. (2019). Propensity scores in kernel equating for non-equivalent groups. *Journal of Educational and Behavioral Statistics*, 44(4), 390–414.
- Wallin, G., Häggeström, J., & Wiberg, M. (2018). How to select the bandwidth in kernel equating? – An evaluation of five different methods. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative Psychology. The 82nd Annual Meeting of the Psychometric Society*, Zurich, 2017 (pp 91–100). Springer.
- Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, 35(12), 1–33. <http://www.jstatsoft.org/v35/i12/>.
- Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement*, 39(5), 349–361.

Predictive Validity Under Partial Observability



Eduardo Alarcón-Bustamante, Ernesto San Martín, and Jorge González

Abstract To assess the predictive capacity of selection tests is a challenge because the response variable is observed only in selected individuals. In this paper we propose to evaluate the predictive capacity of selection tests through marginal effects under a partial identification approach. Identification bounds are defined for the marginal effects under monotonicity assumptions of the response variable. The performance of our method is assessed using a real data set from the university selection test applied in Chile and compared with the marginal effect of the traditional model used in Chile to evaluate the predictive capacity of the selection test.

Keywords Selection problem · Marginal effects · Partial identification · Identification bounds

E. Alarcón-Bustamante (✉)

Department of Statistics, Faculty of Mathematics, Pontificia Universidad Católica de Chile, Santiago, Chile

Interdisciplinary Laboratory of Social Statistics, Santiago, Chile

e-mail: esalarcon@mat.uc.cl

E. San Martín

Department of Statistics, Faculty of Mathematics, Pontificia Universidad Católica de Chile, Santiago, Chile

Interdisciplinary Laboratory of Social Statistics, Santiago, Chile

The Economics School of Louvain, Université Catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium

e-mail: esanmart@mat.uc.cl

J. González

Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile

e-mail: jorge.gonzalez@mat.uc.cl

1 Introduction

A test is used to learn about a behavior of interest. The relationship between test scores and any variable external to the test may be used to predict some (future) behavior of the individuals tested (Lord 1980) in the sense that we are interested in the conditional distribution of those external variables given test scores. In this paper, we focus our attention on tests that are used in a selection process, specifically on admission to the higher education. The purpose of the test is to select the “best applicants” in some specific sense which is typically operationalized through a cutoff score. It is supposed that the cutoff is defined in such a way that higher scores on the test would translate in better performance at higher education.

In this context, it is necessary to assess and measure the quality of the selection, which leads to analyze the *validity* and *reliability* of the admission test. Regarding the validity, it is defined as *the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing (U.S) 2014). In particular, the predictive validity of a test is defined as the evidence based on relations to other variables: in an admission test, these variables are supposed to be chosen according to the selection purposes of a higher educational system. Following this definition, the analysis of the relationship between test scores and any external variable to the test provide an important source of predictive validity evidence.

To assess the predictive validity of a selection test is a challenge because the outcome measured at higher education is observed only in the selected group, whereas the scores of the selection test are observed for the whole population of applicants. This problem is accordingly called *selection problem* and arises when the sampling process does not fully reveal the behavior of the outcome on the support of the predictors (Manski 1993).

Statistical procedures used for the evaluation of the predictive validity include regression models with truncated distributions (Nawata 1994; Heckman 1976, 1979; Marchenko and Genton 2012) and corrected Pearson correlation coefficient (Thorndike 1949; Pearson 1903; Mendoza and Mumford 1987; Lawley 1943; Guilliksen 1950). In the context of admission university selection tests, a common practice to evaluate the predictive validity of the selection tests is to measure the correlation between the obtained scores and the cumulative grade point average (GPA) at the first year of the students in the university.

Although those procedures constitute solutions to the problem of learning about the predictive validity, it is explicitly assumed a prior knowledge for the performance of the *whole population*,¹ that is, it is assumed that the conditional

¹The term whole population refers to the population that is integrated by two subpopulations: the population where the outcome is observed and the one where the outcome is not observed (from here on the observed group and the non-observed group, respectively).

distribution of the outcome given the scores is known up to some parameters. However, we argue that this assumption is not pertinent because the consequence of the partial observability is that the conditional distribution of the outcome given the scores is not identified and therefore assuming any structure for the non-observed group could not be assessed empirically (Manski 1993). For this reason, this approach does not solve satisfactory the problem of predictive validity. In the educational measurement literature, the predictive validity is typically analyzed through the *marginal effect* (for instance, see Leong 2007; Goldhaber et al. 2017; Geiser and Studley 2002), that is, the derivative of the conditional expectation of the outcome given scores, with respect to the scores. However, as the conditional expectation is not identified, the marginal effect is not identified either.

In this paper we propose a methodological approach that allows to learn about the predictive validity of selection tests through the marginal effects, under partial observability of the outcome. We use a partial identification approach in order to define a region that characterizes the set of all admissible values for the marginal effects. This region is delimited by identification bounds. This approach works if explicit assumptions on the unobservable distributions are made, the idea being that such assumptions be weaker than the standard ones abovementioned (Manski 2013). We propose to find identification bounds by assuming that the selection test is such that higher scores would translate to higher values of the outcome, i.e., it is considered that there is a positive relationship between test scores and the outcome. This assumption reflects an optimistic viewpoint on the selection test, and the idea is to get conclusions to be compared with other perspectives. Identification bounds are rigorously operationalized through the monotonicity of the conditional expectation of the outcome given the test score.

The paper is organized as follows. The general framework of the partial identification approach is introduced in Sect. 2. In Sect. 2.1 the partial identification framework of the conditional expectation is formalized. Identification bounds for marginal effects are formally described in Sect. 2.2. In Sect. 2.3 the identification bounds for the marginal effects in the selection problem context are formally characterized. In Sect. 3, the performance of the proposed methodology is illustrated on a real data set from the selection test used in the university admission Chilean system. Conclusions and further work are discussed in Sect. 4.

2 Partial Identification Framework

2.1 *Partial Identification of the Conditional Expectation*

Let Y denote the outcome variable, X a test score, and Z a binary random variable with $Z = 1$ if the outcome is observed and $Z = 0$ otherwise. Consequently, each member of the population is characterized by a triple (Y, Z, X) . In this paper the

attention is focused on the conditional expectation of the outcome Y given a test score X . By the law of total probability, it follows that

$$\mathbb{E}(Y|X) = \mathbb{E}(Y|X, Z = 1)\mathbb{P}(Z = 1|X) + \mathbb{E}(Y|X, Z = 0)\mathbb{P}(Z = 0|X). \quad (1)$$

In (1), $\mathbb{E}(Y|X, Z = 1)$, $\mathbb{P}(Z = 1|X)$, and $\mathbb{P}(Z = 0|X)$ are identified by the data generating process. However, $\mathbb{E}(Y|X, Z = 0)$ is not identified. Consequently $\mathbb{E}(Y|X)$ is not identified either.

One solution for this problem is to assume *weak ignorability*, namely, $Y \perp Z|X$,² which implies that $\mathbb{E}(Y|X) = \mathbb{E}(Y|X, Z = 1)$. The assumption of weak ignorability allows making inferences on $\mathbb{E}(Y|X)$ *ignoring* the non-observed values of Y , which can lead to underestimation of the predictive capacity of the selection test (Manzi et al. 2008).

Assuming that $Y \in [y_0, y_1]$ where y_0 and y_1 are the minimum and the maximum possible GPA, respectively, it follows that $y_0 \leq \mathbb{E}(Y|X, Z = 0) \leq y_1$. By applying this inequality to equation (1), we have

$$\begin{aligned} \mathbb{E}(Y|X, Z = 1)\mathbb{P}(Z = 1|X) + y_0\mathbb{P}(Z = 0|X) &\leq \mathbb{E}(Y|X) \\ &\leq \mathbb{E}(Y|X, Z = 1)\mathbb{P}(Z = 1|X) + y_1\mathbb{P}(Z = 0|X). \end{aligned}$$

The lower bound of the conditional expectation is interpreted as the value $\mathbb{E}(Y|X)$ takes if, in the non-observed group, Y is always equal to y_0 (i.e., if all students obtained the worst GPA). Regarding the upper bound, it is interpreted as the value $\mathbb{E}(Y|X)$ takes if, in the non-observed group, Y is always equal to y_1 (i.e., if all students obtained the best GPA) (Manski 1989).

2.2 Partial Identification of the Marginal Effects

As it is well-known, the marginal effect is defined as

$$M.E(X) = \frac{d\mathbb{E}(Y|X)}{dX}.$$

Figure 1 shows how variations in X reflect in variations on $\mathbb{E}(Y|X)$. These variations are quantified by the change of $\mathbb{E}(Y|X)$ with respect to changes in X , defined by $M.E(X)$. The blue dashed lines represent the marginal effect evaluated at the point $X = x$. From equation (1) it follows that

²In words, $Y \perp Z|X$ indicates that Y is conditionally orthogonal to Z (see Florens and Mouchart 1982).

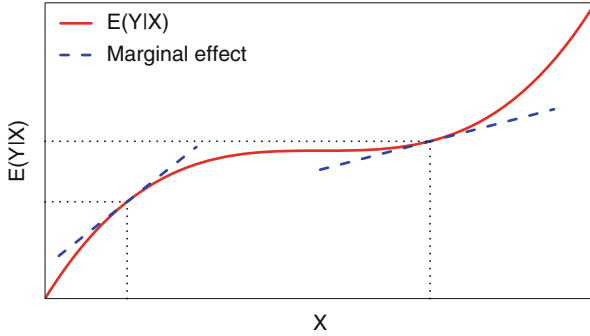


Fig. 1 The regression function and the marginal effect

$$\begin{aligned}
 M.E(X) &= \left(\mathbb{P}(Z = 0|X)M.E_{Z=0}^X + \mathbb{P}(Z = 1|X)M.E_{Z=1}^X \right) \\
 &+ \left([\mathbb{E}(Y|X, Z = 0) - \mathbb{E}(Y|X, Z = 1)] \frac{d\mathbb{P}(Z = 0|X)}{dX} \right), \tag{2}
 \end{aligned}$$

where $M.E_{Z=z}^X$ is the marginal effect of X over the group $Z = z$, with $z \in \{0, 1\}$.

As it was mentioned in Sect. 1, the conditional expectation is not identified in the context of the selection problem. Consequently, the marginal effect will not be identified either. In fact, in equation (2) both $\mathbb{E}(Y|X, Z = 0)$ and $M.E_{Z=0}^X$ are not identified by the sampling process.

In order to find the identification bounds for the marginal effects, suppose that $D_{0x} \leq M.E_{Z=0}^{X=x} \leq D_{1x}$, where $M.E_{Z=z}^{X=x}$ is the marginal effect of X over the group $Z = z$ evaluated at $X = x$. Thus, we assume that the marginal effect for the non-selected population exist, which means that if this population had been selected, the score of the selection test would have predicted the outcome with an associated marginal effect. This marginal effect is subject to uncertainty, as reflected by the bounds, which depends on specific values of X .

Considering the above assumption and that $y_0 \leq \mathbb{E}(Y|X, Z = 0) \leq y_1$, $M.E(X)$ evaluated at $X = x$, denoted by $M.E^{X=x}$, is bounded as follows

$$\begin{aligned}
 &D_{0x}\mathbb{P}(Z = 0|X = x) + \mathbb{P}(Z = 1|X = x)M.E_{Z=1}^{X=x} + \\
 &\quad [y_0 - \mathbb{E}(Y|X = x, Z = 1)] \frac{d\mathbb{P}(Z = 0|X)}{dX} \Big|_{X=x} \leq M.E^{X=x} \leq \\
 &D_{1x}\mathbb{P}(Z = 0|X = x) + \mathbb{P}(Z = 1|X = x)M.E_{Z=1}^{X=x} + \\
 &\quad [y_1 - \mathbb{E}(Y|X = x, Z = 1)] \frac{d\mathbb{P}(Z = 0|X)}{dX} \Big|_{X=x}
 \end{aligned}$$

Note that an assumption on $\mathbb{E}(Y|X, Z = 0)$ does not by itself restrict the marginal effect, but an assumption on both $\mathbb{E}(Y|X, Z = 0)$ and the marginal effect for the non-observed group does (Manski 1989).

2.3 Identification Bounds for Marginal Effects

According to Manski (2003, 2007, 2005), researchers sometimes take credible information about properties of the outcome. For example, there might be reasons to believe that the outcome increase/decrease monotonically when the predictor increase/decrease. Thus, in a University Admission System it can be assumed that, from an optimistic viewpoint, a higher score in the selection test implies a higher GPA at the first year of the university. What can be concluded for the marginal effect under this optimistic assumption? The partial identification analysis aims to answer this question.

Selection tests are used to select the best applicants such that higher scores, X , would imply higher values of the outcome, Y . This fact allows to think that the conditional expectation of Y given X is a non-decreasing function of X and, consequently, the marginal effect will be greater or equal than zero, i.e., $M.E(X) \geq 0$. In order to find an explicit expression for D_{0x} , note that equation (2) is the sum of two terms, namely

$$a = \mathbb{P}(Z = 0|X)M.E_{Z=0}^X + \mathbb{P}(Z = 1|X)M.E_{Z=1}^X, \text{ and}$$

$$b = [\mathbb{E}(Y|X, Z = 0) - \mathbb{E}(Y|X, Z = 1)] \frac{d\mathbb{P}(Z = 0|X)}{dX}.$$

In terms of a and b the marginal effect will be positive if $a \geq -b$. This fact implies that

$$M.E_{Z=0}^X \geq -\frac{\mathbb{P}(Z=1|X)}{\mathbb{P}(Z=0|X)} M.E_{Z=1}^X - \frac{[\mathbb{E}(Y|X, Z=0) - \mathbb{E}(Y|X, Z=1)] \frac{d\mathbb{P}(Z=0|X)}{dX}}{\mathbb{P}(Z=0|X)}. \quad (3)$$

It can be proved that (see Appendix A.1) an explicit expression for D_{0x} is given by

$$D_{0x} = \max_{x \in X} \left\{ \frac{2\mathbb{E}(Y|X=x, Z=1) - y_0 - y_1}{2\mathbb{P}(Z=0|X=x)} \frac{d\mathbb{P}(Z=0|X)}{dX} \Big|_{X=x} - \frac{\mathbb{P}(Z=1|X=x)}{\mathbb{P}(Z=0|X=x)} M.E_{Z=1}^X \right\}.$$

For D_{1x} , suppose that the predictability of X over Y in the observed group is at least equal to the one in the non-observed group. This assumption is realistic because one objective of selection tests is to choose those applicants that would obtain a better outcome than those who were not selected. In terms of marginal effects, this assumption translates to $M.E_{Z=1}^X \geq M.E_{Z=0}^X$ which implies that $D_{1x} = \max_{x \in X} \{M.E_{Z=1}^X\}$.

3 Illustration

The evolution of the university admission system in Chile includes the *baccalaureate* test, administrated during 1931 and 1966, and the *Prueba de Aptitud Académica* (PAA, for their initials in Spanish), administered during the period 1967–2002. These tests were criticized, among other reasons, because of their low predictive capacity (Grassau 1956; DEMRE 2016; Donoso 1998). Since 2003 the selection process is partially based³ on scores from the *Prueba de Selección Universitaria* (PSU, for their initials in Spanish). The PSU is elaborated based on the secondary school curriculum and includes two mandatory tests (Mathematics and Language and Communication) and two elective tests (Sciences and History, Geography and Social Sciences). According to Donoso (1998), one of the reasons for the evolution of the Chilean university admission system is the necessity to increase predictive capacity of the selection tests.

To illustrate the partial identification approach proposed in this paper, we analyze the predictive validity of the mandatory PSU tests over the GPA of students in the first year in a Chilean university. It is important to highlight that the analysis is based on the top one university,⁴ so that the assumption that the performance of non-enrolled students will be at most equal to that of enrolled students is tenable.

3.1 Estimation of the Identification Bounds

Let Y denote the GPA and X the score in the selection factor of interest. The conditional expectation $\mathbb{E}(Y|X, Z = 1)$ was estimated by an adaptive local linear regression model using a symmetric Kernel as implemented in the `loess.as` function from the `FANCOVA` R-package (Wang 2010). The probability of being observed was modeled assuming that $\mathbb{P}(Z = 1|X) = \Phi(\alpha X)$, where $\Phi(\nu)$ is the standard normal cumulative probability distribution evaluated at ν . We considered the standardized values for both X and Y . This means that, by taking into account that in Chile a score of 1 is the minimum GPA that could be obtained, and a score of 7 the maximum one, we evaluated our method using $y_0 = \frac{1 - \overline{GPA}}{sd(GPA)}$ and $y_1 = \frac{7 - \overline{GPA}}{sd(GPA)}$, where \overline{GPA} is the mean of the observed GPA (5.042) and $sd(GPA)$ is its standard deviation (0.568).

³Additional to these tests, there are other selection factors that are considered in the selection process, namely, Ranking and High school grade point average (NEM).

⁴According to the Quacquarelli Symonds University Rankings 2019.

3.2 Results

Figure 2 shows the identification bounds of the marginal effect for both the Mathematics test and the Language and Communication test. Our method is compared with the marginal effect of the multiple linear regression model, the traditional procedure that have been used in Chile in order to evaluate the predictive capacity of the selection factors. For this model, the marginal effect of X is given by its regression coefficient.

For the Mathematics test, it can be seen that the marginal effects are not necessarily a constant function of the test score. In this case, a higher score produces a higher marginal effect. In other words, a good performance in the Mathematics test stands for a better performance in the GPA. Note that these types of conclusions cannot be obtained using the traditional procedure. Regarding the marginal effects for the Language and Communication test, contrary to what was seen for the Mathematics test, the identification bounds are nearly a constant function of the test score. This means that a good performance in the Language and Communication test does not stand for a better performance in the GPA at the first year.

As it was mentioned before, the identification bounds were computed under an optimistic scenario of the selection process. In this context, when using the traditional procedure as implemented in Chile, the marginal effect of the Mathematics test is more optimistic than the optimism manifested by the bounds (see left panel in Fig. 2). In contrast, with the traditional procedure, the marginal effect of the Language and Communication test is less optimistic than the optimism manifested by the bounds (see right panel in Fig. 2).

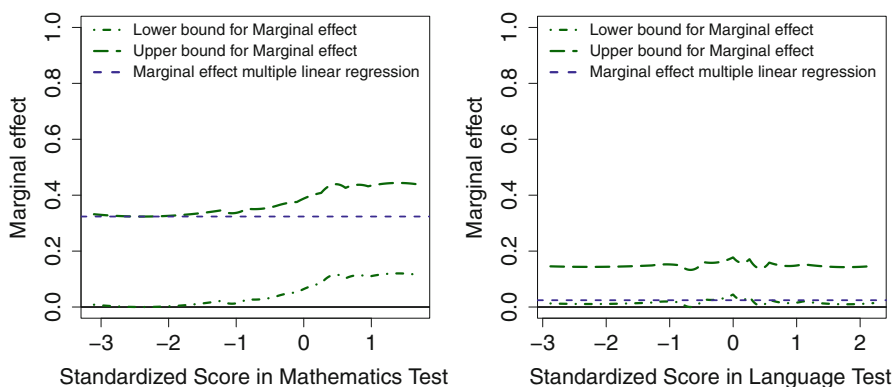


Fig. 2 Identification bounds for marginal effects

4 Conclusions and Discussion

We have presented a method that allows to learn about the predictive validity of selection tests through the marginal effect under partial observability.

Our partial identification-type solution characterizes the set of all admissible values of the marginal effect. i.e., if the proposed model for the evaluation of the predictive capacity captures the information “the performance of the non-observed group is at most equal to the performance of the observed group,” then the marginal effect of X must lie between the identification bounds.

Although other approaches have been proposed to tackle the selection problem by assuming that the regression line does not change between the observed group and the non-observed group, our proposal has the advantage of not assuming any (parametric) structure for the non-observed group, as we only use properties of the selection tests. More specifically, we used monotonicity assumptions in order to find the set of all the possible values of the marginal effect by considering that the selection process is correct. However, this scenario make sense only when information about the top one university is available and if it is assumed that the conditional expectation on the observed group is written only in terms of X . Extending the approach for the scenario where information of more universities is available is a topic for future research. Also extending the model by considering that the conditional expectation depends on more than one covariate is planned for future work.

Acknowledgments Eduardo Alarcón-Bustamante was funded by CONICYT Doctorado Nacional grant 2018-21181007. Ernesto San Martín was partially funded by the FONDECYT grant 1181261.

A Proofs

A.1 Explicit Expression for D_{0x}

In equation (3) $\mathbb{E}(Y|X, Z = 0)$ is not identified by the sampling process, but it is assumed that $y_0 \leq \mathbb{E}(Y|X, Z = 0) \leq y_1$. Then, restrictions for D_{0x} are given by

$$M.E_{Z=0}^{X=x} \geq \frac{[\mathbb{E}(Y|X = x, Z = 1) - y_0]}{\mathbb{P}(Z = 0|X = x)} \frac{d\mathbb{P}(Z = 0|X)}{dX} \Big|_{X=x} - \frac{\mathbb{P}(Z = 1|X = x)}{\mathbb{P}(Z = 0|X = x)} M.E_{Z=1}^{X=x}$$

$$M.E_{Z=0}^{X=x} \geq \frac{[\mathbb{E}(Y|X = x, Z = 1) - y_1]}{\mathbb{P}(Z = 0|X = x)} \frac{d\mathbb{P}(Z = 0|X)}{dX} \Big|_{X=x} - \frac{\mathbb{P}(Z = 1|X = x)}{\mathbb{P}(Z = 0|X = x)} M.E_{Z=1}^{X=x}$$

By combining these expressions it is obtained that

$$2 \cdot M \cdot E_{Z=0}^{X=x} \geq \frac{[\mathbb{E}(Y|X=x, Z=1) - y_0]}{\mathbb{P}(Z=0|X=x)} \frac{d\mathbb{P}(Z=0|X)}{dX} \Big|_{X=x} - \frac{\mathbb{P}(Z=1|X=x)}{\mathbb{P}(Z=0|X=x)} M \cdot E_{Z=1}^{X=x} \\ + \frac{[\mathbb{E}(Y|X=x, Z=1) - y_1]}{\mathbb{P}(Z=0|X=x)} \frac{d\mathbb{P}(Z=0|X)}{dX} \Big|_{X=x} - \frac{\mathbb{P}(Z=1|X=x)}{\mathbb{P}(Z=0|X=x)} M \cdot E_{Z=1}^{X=x}$$

this implies that,

$$M \cdot E_{Z=0}^{X=x} \geq \frac{2\mathbb{E}(Y|X=x, Z=1) - y_0 - y_1}{2\mathbb{P}(Z=0|X=x)} \frac{d\mathbb{P}(Z=0|X)}{dX} \Big|_{X=x} - \frac{\mathbb{P}(Z=1|X=x)}{\mathbb{P}(Z=0|X=x)} M \cdot E_{Z=1}^{X=x}$$

$$M \cdot E_{Z=0}^{X=x} \geq \max_{x \in X} \left\{ \frac{2\mathbb{E}(Y|X=x, Z=1) - y_0 - y_1}{2\mathbb{P}(Z=0|X=x)} \frac{d\mathbb{P}(Z=0|X)}{dX} \Big|_{X=x} - \frac{\mathbb{P}(Z=1|X=x)}{\mathbb{P}(Z=0|X=x)} M \cdot E_{Z=1}^{X=x} \right\}.$$

And therefore,

$$D_{0x} = \max_{x \in X} \left\{ \frac{2\mathbb{E}(Y|X=x, Z=1) - y_0 - y_1}{2\mathbb{P}(Z=0|X=x)} \frac{d\mathbb{P}(Z=0|X)}{dX} \Big|_{X=x} - \frac{\mathbb{P}(Z=1|X=x)}{\mathbb{P}(Z=0|X=x)} M \cdot E_{Z=1}^{X=x} \right\}.$$

References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- DEMRE. (2016, 9). *Prueba de selección universitaria* (Tech. Rep.). Universidad de Chile.
- Donoso, S. (1998). La reforma educacional y el sistema de selección de alumnos a las universidades: impactos y cambios demandados. *Estudios Pedagógicos*(24), 7–30.
- Florens, J., & Mouchart, M. (1982). A note on noncausality. *Econometrica*, 50(3), 583–592.
- Geiser, S., & Studley, R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*, 8(1), 1–26.
- Goldhaber, D., Cowan, J., & Theobald, R. (2017). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education*, 68(4), 377–393.
- Grassau, E. (1956). Análisis estadístico de las pruebas de bachillerato. *Anales de la Universidad de Chile*(102), 77–93.
- Guilliksen, H. (1950). *Theory of mental tests*. New York: Willey.

- Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 46, 931–961.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.
- Lawley, D. (1943). IV.-A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Science*, 62(1), 28–30.
- Leong, C.-H. (2007). Predictive validity of the multicultural personality questionnaire: A longitudinal study on the socio-psychological adaptation of Asian undergraduates who took part in a study-abroad program. *International Journal of Intercultural Relations*, 31, 545–559.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. New York: Routledge.
- Manski, C. (1989). Anatomy of the selection problem. *The Journal of Human Resources*, 24(3), 343–360.
- Manski, C. (1993). Identification problems in the social sciences. *Sociological Methodology*, 23, 1–56.
- Manski, C. (2003). *Partial identification of probability distributions*. New York: Springer.
- Manski, C. (2005). *Social choice with partial knowledge of treatment response* (1st ed.). New Jersey: Princeton University Press.
- Manski, C. (2007). *Identification for prediction and decision*. Cambridge: Harvard University Press.
- Manski, C. (2013). *Public policy in an uncertain world: Analysis and decisions*. Cambridge: Harvard University Press.
- Manzi, J., Bravo, D., Pino, G. del, Donoso, G., Martínez, M., & Pizarro, R. (2008, 7). *Estudio de la validez predictiva de los factores de selección a las universidades del consejo de rectores, admisiones 2003 al 2006* (Tech. Rep.). Comité Técnico Asesor, Honorable Consejo de Rectores de las Universidades Chilenas.
- Marchenko, Y. V., & Genton, M. G. (2012). A Heckman Selection-t model. *Journal of the American Statistical Association*, 107(497), 304–317.
- Mendoza, J., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, 12(3), 282–293.
- Nawata, K. (1994). Estimation of sample selection bias models by the maximum likelihood estimator and Heckman's two-step estimator. *Economics Letters*, 45(1), 33–40.
- Pearson, K. (1903). Mathematical contribution to the theory of evolution-XI on the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London*, 200(Ser. A), 1–66.
- Thorndike, R. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.
- Wang, X.-F. (2010). fANCOVA: Nonparametric Analysis of Covariance [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=fANCOVA> (R package version 0.5-1).

Multiple-Group Propensity Score Inverse Weight Trimming and Its Impact on Covariate Balance and Bias in Treatment Effect Estimation



Diego Luna-Bazaldua  and Laura O'Dwyer

Abstract Propensity scores have become a key technique to analyze causal effects of interventions in observational research. Contemporary developments in propensity score methods facilitate the estimation of treatment effects when more than two intervention groups are compared. Former research has documented the benefits and weaknesses of trimming inverse weights of propensity scores in the context of one treatment and one comparison group, but no research to date has explored the implications of trimming inverse weights in the context of multiple-group propensity scores. The present study adds to the current research on this topic by analyzing the impact of trimming multiple-group inverse weights on covariate balance and treatment effect estimation. Results from the simulation study showed that trimming inverse weights increased covariate bias and did not have a substantial impact on parameter recovery statistics of the treatment effect; moreover, data mining methods produced less extreme inverse weights compared to multinomial logistic regression models.

Keywords Propensity scores · Multiple-group treatments · Multinomial logistic regression · Generalized boosted models · Neural networks

1 Introduction

The use of propensity scores (PS) in observational research to estimate the causal effects of interventions or treatments on outcomes has been increasing over the last years in social and health sciences (Thoemmes and Kim 2011). Techniques based on PS allow to reduce the bias of observed baseline covariates between intervention and

D. Luna-Bazaldua (✉)
The World Bank Group, Washington, DC, USA
e-mail: dlunabazaldua@worldbank.org

L. O'Dwyer
Lynch School of Education, Boston College, Chestnut Hill, MA, USA
e-mail: laura.odwyer.1@bc.edu

comparison groups in order to improve the estimation of the effect of interventions on outcomes. The original framework on PS developed by Rosenbaum and Rubin (1983) considered instances with only one treatment and one comparison group, but more recent developments now allow for the estimation multiple-group (MG) PS in the context of multiple treatments or interventions (McCaffrey et al. 2013).

PS are defined as a participant's conditional probability of exposure to a treatment versus a comparison group given baseline covariates (Austin and Stuart 2015). Within the causal inference framework (Morgan and Winship 2007; Murnane and Willett 2010), PS were originally introduced as an approach to reduce the effect of sources of potential bias in the estimation of causal effects on an outcome variable in observational and quasi-experimental designs (Rosenbaum and Rubin 1983, 1984, Rubin 1997). Compared to other statistical adjustment techniques aimed to reduce bias in the estimation of treatment effects (e.g., analysis of covariance (ANCOVA)), PS have the advantage of minimizing baseline covariate differences between groups (West et al. 2000); in addition, conditional on the PS, baseline covariates are independent of the group assignment (Austin and Stuart 2015; Rosenbaum and Rubin 1983, 1984).

PS can be used in four different methods to reduce bias from baseline covariates on the treatment effect estimation: matching of cases between groups, stratification or subclassification of cases, PS included as covariate in ANCOVA models, and inverse probability weighting (Austin 2011; Thoemmes and Kim 2011). Former research has found that matching and inverse weighting methods decrease more bias on the baseline covariates between groups when compared to stratification methods (Austin 2011; Austin et al. 2007; Lunceford and Davidian 2004). In addition, approaches that simultaneously include the baseline covariates in addition to PS adjustments have shown to be doubly robust in the estimation of treatment effects (Hall et al. 2015).

Similar to the other three PS methods, inverse weights of propensity scores (IWPS) can be defined to estimate either the average treatment effects (ATE) on the outcome or the average treatment effect for the treated (ATT) on the outcome (Austin and Stuart 2015; Hirano et al. 2003). The ATE estimates the average treatment effect of every unit in the population that had been exposed to the treatment, whereas the ATT estimates the average treatment effect only for those units that were in the treatment group (Austin 2011; Morgan and Winship 2007). Given that PS close to the boundaries of zero or one produce extreme inverse weights, a drawback of using IWPS methods is that the treatment effect estimates tend to be influenced by those cases whose PS are close to values of zero or one (Austin and Stuart 2015; Thoemmes and Kim 2011). In those instances, former research done by Lee et al. (2011) has shown that trimming extreme IWPS improves balance in the baseline covariates and treatment effect estimates.

The original PS estimation approach proposed by Rosenbaum and Rubin (Rosenbaum and Rubin 1983, 1984) was focused on the estimation of PS in designs with only one treatment and one comparison group. However, recent extensions of the causal inference framework have developed methods for the estimation of multiple-group (MG) propensity scores. Specifically, McCaffrey et al. (2013) introduced an

MG PS framework to estimate treatment effects when the interest is focused on one specific treatment relative to three or more conditions. The MG PS approach uses inverse weights to balance the covariates among groups and improve the estimation of either the ATE or ATT treatment effect for the condition of interest.

Stuart et al. (2014) showed applications of the MG IWPS for the estimation of treatment effects in the context of difference-in-differences designs. In empirical research, MG IWPS have been employed in observational studies comparing the impact of three different surgery types for breast cancer on mortality rates in observational studies (Kurian et al. 2014). Similarly, this multigroup technique has been used to compare the effect of three different hospital-based care interventions on outcomes such as rehospitalization and death in a sample of cognitively impaired older adults (Naylor et al. 2014).

As mentioned before, prior research on IWPS for two groups has documented the benefits and limitations of IWPS trimming using different PS estimation techniques (Lee et al. 2011). However, no research to date has explored the implications of MG IWPS trimming on covariate balance improvement and treatment effect estimation. Thus, it is critical to analyze this topic to understand the role of trimming inverse weights on the estimation of treatment effects in the context of designs with more than two groups. This research focuses on the role of MG IWPS trimming under several simulated propensity score and treatment effect conditions.

Given the lack of previous research on truncating very large MG IWPS, the objective of this study was to analyze the impact of MG IWPS trimming on two key criteria in quasi-experimental and observational research: decrease of improvement in the mean maximum absolute (MMA) covariate balance between groups and decrease of bias in treatment effect estimation. To reach this objective, the current research reports the methodology, results, and discussion from a simulation study and a study with empirical data.

2 Methods

2.1 Simulation Study

The simulation study began with the random generation of four independent continuous covariates (X_1, \dots, X_4) and two dichotomous covariates (X_5 and X_6) for $i = 1, \dots, 2000$ individual cases. The continuous covariates were generated from independent standard normal distributions, while the dichotomous covariates from a binomial distribution $X_5 \sim \text{Bin}(N = 2000, p = 0.3)$ and $X_6 \sim \text{Bin}(N = 2000, p = 0.7)$. Treated as baseline covariates, these six variables had an impact on the formation of four treatment groups $T_i = t\{1, \dots, 4\}$ and, subsequently, on a continuous outcome Y_i .

Two factors, each with two levels, were considered in this simulation study. The first factor was determined by the multinomial logistic regression model used for

Table 1 Simulation conditions

Condition	Model for PS	Model for outcome
1	Main effects MLR	Main effects LR
2	MLR with interactions and squared terms	Main effects LR
3	Main effects MLR	LR with interactions and squared terms
4	MLR with interactions and squared terms	LR with interactions and squared terms

Note: MLR refers to multinomial logistic regression model and LR to linear regression model

the generation of the PS (i.e., a main effects model or a polynomial model with interactions among the covariates), and the second factor by the linear regression model used to generate the outcome (i.e., a main effects model or a polynomial model with interactions among the covariates). The combination of these factors resulted in four different conditions described in Table 1, with 250 replications of the experiment conducted in each condition.

For groups t_1 to t_3 , the multinomial logistic regression model with only main effects for the MG PS generation was defined as:

$$P(T_i = t) = \frac{\exp^{\beta_t X}}{1 + \sum_{t=1}^3 \exp^{\beta_t X}}$$

and group t_4

$$P(T_i = 4) = \frac{1}{1 + \sum_{t=1}^3 \exp^{\beta_t X}}$$

where $P(T_i)$ is the probability of being assigned to treatment t . The matrix of values for the covariate coefficients β was:

	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_4 X_4$	$\beta_5 X_5$	$\beta_6 X_6$
$T = t_1$	0.3	0.2	0	0	0.5	-1.0
$T = t_2$	0	0	0.5	0.5	0	-1.0
$T = t_3$	0.5	0	0	0.5	0	-1.0

The values for β were chosen to have a slightly lower proportion of units i in the treatment of interest (i.e., $t = 1$) compared to the other three groups. A second multinomial logistic regression model for the MG PS included the main effects above and additional squared terms and interactions among covariates.

	$\beta_7 X_1^2$	$\beta_8 X_2^2$	$\beta_9 X_3^2$	$\beta_{10} X_4^2$	$\beta_{11} X_5^2$	$\beta_{12} X_6^2$
$T = t_1$	0.1	0.1	0	0	0.1	-0.1
$T = t_2$	0	0	0.1	0.1	-0.1	-0.1
$T = t_3$	0.1	0	0	0.1	0	-0.1

	$\beta_{13}X_1X_2$	$\beta_{14}X_1X_3$	$\beta_{15}X_1X_4$	$\beta_{16}X_1X_5$	$\beta_{17}X_1X_6$	$\beta_{18}X_2X_3$	$\beta_{19}X_2X_4$	$\beta_{20}X_2X_5$
$T = t_1$	0.1	0.1	0	-0.1	-0.1	0.1	0	0
$T = t_2$	0	0	0.1	0	-0.1	0.1	0.1	0
$T = t_3$	0.1	0.1	0.1	-0.1	-0.1	0	0	-0.1

	$\beta_{21}X_2X_6$	$\beta_{22}X_3X_4$	$\beta_{23}X_3X_5$	$\beta_{24}X_3X_6$	$\beta_{25}X_4X_5$	$\beta_{26}X_4X_6$	$\beta_{27}X_5X_6$
$T = t_1$	-0.1	0	0	-0.1	0	-0.1	-0.1
$T = t_2$	-0.1	0.1	0	-0.1	0.1	-0.1	-0.1
$T = t_3$	-0.1	0	-0.1	-0.1	-0.1	-0.1	-0.1

The main effects linear regression model to estimate the treatment effect was:

$$Y_i = \beta_0 + \beta_1 T_{i[1]} + \beta_2 T_{i[2]} + \beta_3 T_{i[3]} + \beta_4 T_{i[4]} + \beta \mathbf{Covs} + E_i$$

where the vector of β coefficients corresponds to the intercept, treatment effects for the four observed conditions $T_{i[t]}$, and covariate effects on the outcome. The error term, E_i , is normally distributed with parameters $N(0, 2)$. The parameter values for the coefficients in this model are $\beta \in \{\beta_0=0, \beta_1=1, \beta_2=-.5, \beta_3=0, \beta_4=-1\}$, and $\beta_{covs} \in \{\beta_5=.4, \beta_6=.2, \beta_7=-.6, \beta_8=-.3, \beta_9=-.4, \beta_{10}=-.3\}$.

Finally, the linear regression model with polynomials and interactions includes additional coefficients for the two-way interactions among covariates and squared covariate effects on the outcome:

$$Y_i = \beta_0 + \beta_1 T_{i[1]} + \beta_2 T_{i[2]} + \beta_3 T_{i[3]} + \beta_4 T_{i[4]} + \beta_5 X_1 + \beta_6 X_2 + \beta_7 X_3 + \beta_8 X_4 + \beta_9 X_5 + \beta_{10} X_6 + \beta_{11} X_1 X_2 + \beta_{12} X_1 X_3 + \beta_{13} X_1 X_4 + \beta_{14} X_1 X_5 + \beta_{15} X_2 X_3 + \beta_{16} X_3 X_4 + \beta_{17} X_4 X_6 + \beta_{18} X_3 X_5 + \beta_{19} X_5 X_6 + \beta_{20} X_1^2 + \beta_{21} X_2^2 + \beta_{22} X_3^2 + \beta_{23} X_4^2 + \beta_{24} X_5^2 + \beta_{25} X_6^2 + E_i$$

with parameter values for the additional coefficients being $\beta_{poly} \in \{\beta_{11}=.3, \beta_{12}=-.2, \beta_{13}=.1, \beta_{14}=-.3, \beta_{15}=.1, \beta_{16}=.1, \beta_{17}=-.2, \beta_{18}=.3, \beta_{19}=-.3, \beta_{20}=-.2, \beta_{21}=.1, \beta_{22}=.2, \beta_{23}=-.1, \beta_{24}=.1, \beta_{25}=-.2\}$.

In this study, the treatment effect of interest is $T_{[1]}$. Thus, the PS estimation, MMA covariate balance improvement, and weighted treatment effect estimation were done taking $t = 1$ as the group of interest (McCaffrey et al. 2013).

2.2 PS and Treatment Effect Estimation

Three different models were used to estimate the MG PS: a main effects multinomial logistic regression (MLR) model, a generalized boosted model (GBM), and a neural networks (NN) model (Keller et al. 2015; McCaffrey et al. 2013; Stuart et al. 2014). The treatment effect was estimated using both unweighted and ATE weighted main effects linear regression models leaving out the dichotomous indicator for treatment 4 $T_{[4]}$ to prevent multicollinearity. Data generation and statistical analyses were

conducted in R (R Core Team 2016), using the “twang” (Ridgeway et al. 2017) and “nnet” (Venables and Ripley 2002) packages. The appendix A includes R code to estimate these three models.

To analyze the role of inverse weight trimming, the three estimated MG IWPS were trimmed based on their percentiles from the 99th to the 50th percentile value, which is the value range explored in prior research on IWPS trimming (Lee et al. 2011). Each one of these trimmed IWPS was used to calculate MMA covariate balance and treatment effect estimates within each replication. The criterion used to determine covariate balance improvement was the MMA covariate bias (McCaffrey et al. 2013). Absolute bias and mean squared error (MSE) were used as criteria for the treatment effect parameter recovery (Rizzo 2008).

3 Results

Table 2 shows that the GBM MG IWPS percentile distribution is consistent across conditions and it does not produce extreme propensity score weights. However, the MLR and NN approaches estimated some extreme propensity scores in conditions 2 and 4.

The effects of MG IWPS trimming are summarized in Tables 3, 4 and 5. Prior to trimming, GBM IWPS consistently produced the lowest MMA covariate bias across conditions (see Table 3). Conversely, MG IWPS estimated from MLR models yielded the highest MMA covariate bias in the four conditions explored. Furthermore, trimming those extreme MLR inverse weights down to the 99th

Table 2 Average IWPS percentiles by condition and PS estimation method

	1st quartile	Median	3rd quartile	Max
Condition 1				
MLR	1.000	1.000	1.000	1.388
GBM	1.028	1.130	1.427	17.776
NN	1.022	1.091	1.343	13.128
Condition 2				
MLR	1.000	1.000	1.021	4.64×10^{20}
GBM	1.051	1.195	1.548	17.833
NN	1.030	1.124	1.443	143.827
Condition 3				
MLR	1.000	1.000	1.000	1.362
GBM	1.029	1.132	1.429	16.758
NN	1.021	1.089	1.341	14.206
Condition 4				
MLR	1.000	1.000	1.022	8.33×10^{19}
GBM	1.050	1.191	1.542	17.984
NN	1.030	1.125	1.445	180.740

Note: MLR refers to multinomial logistic regression, GBM to generalized boosted models, and NN to neural networks

Table 3 Average absolute maximum covariate balance by inverse weight trimming percentile and PS estimation method

Trim percentile	Model	Condition 1	Condition 2	Condition 3	Condition 4
100	MLR	1.1952	2.32 e ¹⁷	1.1957	12,417,219
	GBM	1.0408	1.4703	1.0401	1.011
	NN	1.1264	1.4893	1.1249	1.0372
99	MLR	1.1954	1.0535	1.1959	1.0523
	GBM	1.0576	1.018	1.0561	1.0198
	NN	1.1271	1.0523	1.1266	1.05
98	MLR	1.1956	1.0771	1.1961	1.0768
	GBM	1.0672	1.023	1.0663	1.0252
	NN	1.1292	1.0575	1.1289	1.056
97	MLR	1.1957	1.0876	1.1961	1.0877
	GBM	1.0748	1.0276	1.0742	1.0297
	NN	1.1313	1.0613	1.1311	1.0602
96	MLR	1.1957	1.0945	1.1962	1.0946
	GBM	1.0814	1.0317	1.0809	1.0336
	NN	1.1333	1.0646	1.1333	1.0637
95	MLR	1.1957	1.0995	1.1962	1.0998
	GBM	1.087	1.0354	1.0868	1.0372
	NN	1.1353	1.0672	1.1353	1.0666
90	MLR	1.1957	1.1151	1.1962	1.1156
	GBM	1.109	1.0506	1.109	1.0524
	NN	1.1445	1.0778	1.1449	1.0778
80	MLR	1.1957	1.128	1.1962	1.1286
	GBM	1.1373	1.0729	1.1377	1.0744
	NN	1.16	1.093	1.1606	1.0934
70	MLR	1.1957	1.1315	1.1962	1.1322
	GBM	1.1571	1.0897	1.1574	1.091
	NN	1.1725	1.1053	1.1731	1.106
60	MLR	1.1957	1.1324	1.1962	1.1331
	GBM	1.1715	1.1031	1.1719	1.1043
	NN	1.1822	1.1155	1.1828	1.1163
50	MLR	1.1957	1.1325	1.1962	1.1332
	GBM	1.182	1.1137	1.1823	1.1149
	NN	1.1889	1.123	1.1894	1.1239

Note: MLR refers to multinomial logistic regression, GBM to generalized boosted models, and NN to neural networks

percentile improved covariate balance. However, trimming MG IWPS below the 98th percentile tended to increase covariate bias across conditions.

Table 4 presents the average parameter recovery statistics for the treatment effect parameter (i.e., $\beta_1 = 1$) for unweighted and MG IWPS weighted regression models in the four conditions. The unweighted results reported in Table 4 correspond to regression models that included only the treatment dummy indicators

Table 4 Average parameter recovery statistics of the treatment effect in the four conditions

β_1	Abs. bias	MSE	β_1	Abs. bias	MSE	
	Condition 1			Condition 2		
Unweighted	1.405	0.406	0.193	1.610	0.610	0.387
Unweightedw/ covariates	1.492	0.492	0.279	1.505	0.505	0.288
MLR	1.494	0.501	0.291	1.795	1.048	2.084
GBM	1.503	0.504	0.305	1.496	0.497	0.281
NN	1.542	0.543	0.344	1.517	0.537	0.380
	Condition 3			Condition 4		
Unweighted	1.674	0.674	0.490	2.097	1.097	1.222
Unweightedw/ covariates	1.299	0.317	0.135	1.464	0.465	0.360
MLR	1.293	0.305	0.127	1.800	1.271	2.875
GBM	1.334	0.354	0.172	1.408	0.411	0.207
NN	1.324	0.338	0.158	1.520	0.545	0.441

Note: MSE refers to the mean squared error of the estimator, and Abs. bias to absolute bias

as independent variables and models that also included the six additional covariates; the weighted results in Table 4 correspond to regression coefficients estimated with unrestricted (i.e., not trimmed) MG IWPS. The unweighted model without covariates produced the lowest absolute bias and MSE compared to the weighted estimators only in condition 1, but the unweighted models tended to produce a larger average bias and MSE in conditions 2 and 4. More research is needed to explain why the unweighted model with covariates produced a larger average bias than the equivalent model without covariates in condition 1. The weighted estimator using GBM inverse weights produced the lowest bias and MSE in conditions 2 and 4, whereas the weighted estimator using MLR inverse weights presented the lowest bias and MSE in condition 3. Since condition 4 is the closest to a real-life scenario with multiple interacting covariates implicated in the determination of groups and treatment effects, a practical implication of the results in Table 4 is that data mining models, in particular generalized boosted models, may be preferred over other approaches when comparing effects in observational studies.

Results for the average parameter recovery statistics of weighted models using trimmed inverse weights are reported in Table 5. Overall, the results in Table 5 show that parameter recovery of the treatment effect does not change in most conditions as the MG IWPS are trimmed down to the 50th percentile; only in condition 4 there is a consistent but minor improvement in the estimation of the parameter β_1 when the largest inverse weights are truncated to the 98th percentile. However, trimming the inverse weights do not produce a significant change in parameter recovery in the case of the MG IWPS estimated using MLR models.

Table 5 Average parameter recovery statistics of the treatment effect using trimmed inverse weights in the four conditions

Percentile	MLR			GBM			NN		
	β_1	Abs. bias	MSE	β_1	Abs. bias	MSE	β_1	Abs. bias	MSE
Condition 1									
99	1.492	0.502	0.292	1.492	0.492	0.283	1.532	0.543	0.326
98	1.492	0.503	0.295	1.490	0.490	0.280	1.528	0.543	0.320
97	1.492	0.503	0.295	1.490	0.490	0.279	1.526	0.542	0.317
96	1.492	0.503	0.295	1.490	0.490	0.278	1.523	0.542	0.313
95	1.492	0.503	0.296	1.490	0.490	0.278	1.521	0.542	0.310
90	1.492	0.502	0.295	1.490	0.490	0.277	1.512	0.543	0.300
80	1.492	0.502	0.295	1.490	0.490	0.277	1.501	0.543	0.289
70	1.492	0.502	0.295	1.490	0.491	0.277	1.497	0.543	0.284
60	1.492	0.501	0.294	1.491	0.491	0.278	1.494	0.544	0.281
50	1.492	0.500	0.293	1.491	0.491	0.278	1.493	0.544	0.280
Condition 2									
99	1.493	0.493	0.279	1.504	0.504	0.279	1.513	0.529	0.293
98	1.497	0.497	0.274	1.505	0.505	0.280	1.513	0.532	0.291
97	1.499	0.499	0.274	1.506	0.506	0.281	1.514	0.532	0.291
96	1.500	0.500	0.274	1.507	0.507	0.281	1.514	0.532	0.291
95	1.501	0.501	0.275	1.507	0.507	0.281	1.514	0.532	0.290
90	1.503	0.503	0.276	1.508	0.508	0.282	1.514	0.532	0.289
80	1.505	0.505	0.278	1.509	0.509	0.282	1.511	0.531	0.285
70	1.505	0.505	0.278	1.508	0.508	0.281	1.509	0.530	0.283
60	1.505	0.505	0.278	1.507	0.507	0.281	1.508	0.529	0.281
50	1.505	0.505	0.278	1.507	0.507	0.280	1.506	0.529	0.280
Condition 3									
99	1.293	0.305	0.127	1.344	0.356	0.169	1.336	0.340	0.159
98	1.293	0.305	0.126	1.340	0.352	0.163	1.335	0.339	0.157
97	1.293	0.305	0.126	1.337	0.348	0.159	1.333	0.338	0.155
96	1.293	0.305	0.126	1.334	0.345	0.156	1.331	0.338	0.154
95	1.293	0.305	0.126	1.332	0.343	0.154	1.330	0.337	0.152
90	1.293	0.304	0.127	1.324	0.335	0.147	1.322	0.335	0.146
80	1.293	0.304	0.126	1.314	0.325	0.140	1.310	0.334	0.138
70	1.293	0.304	0.126	1.307	0.319	0.136	1.304	0.334	0.134
60	1.293	0.304	0.126	1.302	0.314	0.132	1.299	0.333	0.131
50	1.293	0.304	0.126	1.298	0.310	0.130	1.296	0.333	0.129
Condition 4									
99	1.395	0.406	0.202	1.403	0.405	0.192	1.387	0.327	0.184
98	1.383	0.386	0.178	1.401	0.403	0.190	1.385	0.325	0.180
97	1.379	0.381	0.173	1.401	0.402	0.189	1.385	0.325	0.179
96	1.377	0.379	0.171	1.400	0.401	0.188	1.385	0.325	0.178
95	1.376	0.377	0.169	1.399	0.400	0.187	1.385	0.324	0.178
90	1.371	0.372	0.165	1.396	0.397	0.184	1.384	0.324	0.176
80	1.366	0.367	0.161	1.388	0.389	0.178	1.380	0.323	0.172
70	1.364	0.366	0.160	1.382	0.383	0.173	1.376	0.322	0.169
60	1.364	0.365	0.160	1.376	0.378	0.169	1.372	0.321	0.166
50	1.364	0.365	0.160	1.372	0.373	0.165	1.369	0.321	0.163

Note: MSE refers to the mean squared error of the estimator, and Abs. bias to absolute bias

4 Discussion

This study adds to the current research on the use of MG IWPS estimation and trimming for estimating causal effects in non-randomized designs. In summary, results from the simulation study showed that trimming inverse weights had a detrimental effect by increasing covariate bias and by not having a substantial impact on parameter recovery statistics of the treatment effect.

The results from the simulation study are similar to those found by Lee et al. (2011), who also reported that trimming tended to increase covariate bias in their simulation studies. These authors noted that trimming had a positive impact on reducing covariate bias and treatment effect bias for extreme weights produced by logistic regression models, which is parallel to our results for inverse weights produced by MLR models.

In addition, inverse weights from the machine learning models here analyzed – GBM and NN – did not yield extreme weights and tended to reduce covariate bias contrasted to MLR models. These results are consistent with the former research that has pointed out the advantages of machine learning models with respect to logistic and multinomial regression models in the estimation of propensity scores (Keller et al. 2015; Lee et al. 2011; McCaffrey et al. 2013). Results of the simulation study also showed that MG PS estimated from an MLR model produced larger inverse weights and higher covariate bias compared to machine learning models.

In empirical research dealing with the comparison of two or more treatment groups, researchers may have limited information about the baseline covariates related to specific treatments and to the outcome. Thus, based on the results from this study, here are some suggestions for the use of MG IWPS and trimming in empirical research:

First, use more than one method to estimate the MG IWPS; this includes MLR models with only main effects and models with interactions and polynomials (Keller et al. 2015; McCaffrey et al. 2013).

Second, compare the distribution of the unrestricted MG IWPS; if a PS estimation method yields considerable larger inverse weights compared to others, there could be some misspecification in that model (Lee et al. 2011).

Third, contrast the unrestricted MG IWPS with respect to baseline covariate balance criteria; models that reduce the most covariate bias tend to be preferred for further analyses of the treatment effect (McCaffrey et al. 2013).

Fourth, when trimming is considered necessary due to the presence of extreme weights, then compare the impact of trimming on covariate balance and on the standard errors of the treatment effect estimate with respect to results using unrestricted inverse weights. If the trimmed inverse weights yield more biased results, the researcher should be cautious about this approach.

Finally, more research must be done, particularly regarding omitted baseline covariates impacting the MG PS and treatment effect; prior research on this topic has found that omitted confounders may bias the estimated PS and treatment effects, but

a small amount of research has addressed this issue in the context of MG PS (Austin 2011; Weitzen et al. 2005).

Acknowledgments This project was done while the first author worked as an associate professor at Universidad Nacional Autónoma de Mexico (UNAM). He is now working at the Education Global Practice of the World Bank Group. The authors acknowledge UNAM’s Dirección General de Asuntos de Personal Académico for the research grant IA303018.

Appendix A – Code to Estimate Multiple-Group Propensity Scores

The code in R presented here exemplifies the estimation of multiple-group propensity scores on a categorical variable that identifies more than two groups (“**Group**” variable in the data set). Propensity scores are estimated using six covariates identified as X_1, \dots, X_6 . Code for data generation can be requested to the authors.

```
# Call libraries that will be needed for this exercise.
library(twang) # For generalized boosted regression
library(nnet)  # For NN Estimation

### 1. After calling the data to R, models are estimated.

# Multinomial Regression model.
MR <- multinom(as.factor(Data$Group) ~ X1 + X2 + X3 + X4 +
               X5 + X6, data = Data)

# Generalized Boosting model.
GBM4 <- mnps(Group ~ X1 + X2 + X3 + X4 + X5 + X6, data = Data,
             n.trees = 3000, interaction.depth = 2,
             shrinkage = 0.01, distribution = "multinomial",
             stop.method = "es.mean")

# Neural Network model. Using 4 group membership.
NN4 <- nnet(as.factor(Data$Group) ~ X1 + X2 + X3 + X4 +
            X5 + X6 , data = Data, size = 3, decay = 0.09)

### 2. Multiple-group propensity scores are extracted from
# the output of each function. Inverse weights can be generated
# by calculation ratios of PSs with respect to the group
# of interest.

# Multinomial Regression MG PS.
MR_PS <- MR$fitted.values

# Generalized Boosting MG PS.
GBM4_PS <- cbind(GBM4$psList[[1]]$ps[,1], GBM4$psList[[2]]$ps[,1],
                GBM4$psList[[3]]$ps[,1], GBM4$psList[[4]]$ps[,1])

# Neural Network MG PS.
NN4_PS <- NN4$fitted.values
```

References

- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424.
- Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679.
- Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*(4), 734–753.
- Hall, C. E., Steiner, P. M., & Kim, J. S. (2015). Doubly robust estimation of treatment effects from observational multilevel data. In L. A. van der Ark, D. Bolt, W. C. Wang, J. Douglas, & S. M. Chow (Eds.), *Quantitative psychology research* (pp. 321–340). New York: Springer.
- Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.
- Keller, B., Kim, J.-S., & Steiner, P. M. (2015). Neural networks for propensity score estimation: Simulation results and recommendations. In L. A. van der Ark, D. M. Bolt, S. M. Chow, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology research*. New York: Springer.
- Kurian, A. W., Lichtensztajn, D. Y., Keegan, T. H., Nelson, D. O., Clarke, C. A., & Gomez, S. L. (2014). Use of and mortality after bilateral mastectomy compared with other surgical treatments for breast cancer in California, 1998–2011. *JAMA*, *312*(9), 902–914.
- Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS One*, *6*(3), e18174. <https://doi.org/10.1371/journal.pone.0018174>.
- Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*(19), 2937–2960.
- McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, *32*(19), 3388–3414.
- Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press.
- Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Boston: Oxford University Press.
- Naylor, M. D., Hirschman, K. B., Hanlon, A. L., Bowles, K. H., Bradway, C., McCauley, K. M., & Pauly, M. V. (2014). Comparison of evidence-based interventions on outcomes of hospitalized, cognitively impaired older adults. *Journal of Comparative Effectiveness Research*, *3*(3), 245–257.
- R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2017). *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups* (Version 1.4–9.5). <https://CRAN.R-project.org/package=twang>
- Rizzo, M. L. (2008). *Statistical computing with R*. Boca Raton: Chapman & Hall/CRC.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.
- Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*, 757–763.
- Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Cherner, M. E., & Barry, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Services and Outcomes Research Methodology*, *14*(4), 166–182.

- Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*(1), 90–118.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York: Springer.
- Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2005). Weaknesses of goodness-of-fit tests for evaluating propensity score models: The case of the omitted confounder. *Pharmacoepidemiology and Drug Safety*, *14*, 227–238.
- West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–88). New York: Cambridge University Press.

Procrustes Penalty Function for Matching Matrices to Targets with Its Applications



Naoto Yamashita

Abstract Penalized estimation is widely used for obtaining sparse solutions, which facilitates easier interpretation compared with ordinal estimation procedures. In this research, as a generalized form of penalized estimation, a new penalty function is proposed. The proposed function shrinks solutions to a prespecified target matrix which possesses a certain simple structure. The resulting solution is therefore simple and easy to interpret, and its simplicity is controlled by some tuning parameters. Two applications of the proposed method are presented: sparse principal component analysis and three-way component analysis. The applications show that the proposed method surely produces sparse and interpretable solutions.

Keywords Penalized estimation · Regularization · Regression · Principal component analysis · PARAFAC

1 Introduction

Interpretability of solutions is one of the most important issues in modern multivariate analysis. Specifically, a growing interest in machine learning technology makes the issue more important, in that it often involves less interpretable and black box models (Rudin 2019). Penalized estimation (Hastie et al. 2015) is widely used for obtaining sparse and interpretable solutions in various procedures of multivariate analysis, which uses penalized functions in order to shrink some elements toward zero. For example, sparse principal component analysis (SPCA) (Jolliffe et al. 2003; Zou et al. 2006) with an L_1 penalty aims to minimize

$$\|\mathbf{Z} - \mathbf{FA}'\|^2 + \lambda\|\mathbf{A}\|_1 \quad (1)$$

N. Yamashita (✉)
Graduate School of Human Sciences, Osaka University, Suita, Japan
e-mail: nyamashita@hus.osaka-u.ac.jp

over an n (objects) $\times r$ (components) score matrix \mathbf{F} and a p (variables) $\times r$ loading matrix \mathbf{A} given an $n \times p$ data matrix \mathbf{Z} , where $\|\mathbf{A}\|_1 = \sum |a_{jk}| (j = 1, \dots, p; k = 1, \dots, r)$ and λ denote the L_1 norm of \mathbf{A} and a tuning parameter, respectively. The second term in (1) is called as a penalty function and serves to shrink the elements in \mathbf{A} toward zero, and therefore the estimated \mathbf{A} is of reduced cardinality.

The paper proposes a new penalty function which includes various penalty functions. The function is called as a Procrustes penalty function, which is formally expressed as

$$P_{Pro}(\mathbf{A}|\mathbf{\Lambda}, \mathbf{T}) = \|(\mathbf{A} - \mathbf{T})\mathbf{\Lambda}\|_1 \quad (2)$$

where $\mathbf{\Lambda}$ denotes the diagonal matrix of tuning parameters $\lambda_1, \dots, \lambda_r (> 0)$. λ_s controls the penalty strength on the s th column of \mathbf{A} . \mathbf{T} denotes a prespecified target matrix with the same dimension as \mathbf{A} . Here, consider to set \mathbf{T} as a matrix with a certain simple structure. The elements in \mathbf{A} shrink toward simple \mathbf{T} so as to approximate its structure, and the resulting \mathbf{A} is considered to be simple and easy to interpret. The novel point of the proposed method is that a solution matrix directly approximates a simple structure, while the existing one shrinks all elements toward zeros without considering whether the resulting \mathbf{A} has a simple structure or not. Procrustes penalty function therefore expected to yield a more simple and interpretable solution matrix compared with the existing methods. It should be noted that the article only considers the case with L_1 norm penalty, since the main goal of the research is to obtain sparse and interpretable solution matrices.

The remaining parts of the article are organized as follows. In the next section, an optimization algorithm for the proposed methods is presented, where a general case of regression model with Procrustes penalty is considered. There, a theorem of the minimizer for the regression problem is also presented. Further, the proposed procedure is applied to some machine learning problems, followed by their examples in Sect. 3. The final section is devoted to conclusions and future remarks.

2 Proposed Method

2.1 General Case: Multivariate Regression

Here, we consider a multiple regression problem with a Procrustes penalty function as a general case. It is formally expressed as the minimization of

$$f(\mathbf{W}) = \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|^2 + P_{Pro}(\mathbf{W}|\mathbf{\Lambda}, \mathbf{T}) \quad (3)$$

where $\mathbf{Y} = [y_1, \dots, y_q]$ denotes the $n \times q$ matrix of the dependent variables and it is regressed on the p independent variables in an $n \times p$ data matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$. \mathbf{W} is an unknown $p \times q$ matrix of regression coefficients. Above, the Procrustes penalty function applied to the coefficient matrix \mathbf{W} serves to obtain an interpretable

estimate of the parameter matrix \mathbf{W} . Here, as $P_{Pro}(\mathbf{W}|\mathbf{A}, \mathbf{T})$, the L_1 norm is considered, $P_{Pro}(\mathbf{W}|\mathbf{A}, \mathbf{T}) = \|\mathbf{A}(\mathbf{A} - \mathbf{T})\|_1$, because it is able to produce a \mathbf{W} including some exact zero elements.

For the minimization of equation (3) over \mathbf{W} , we have the following theorem. Note that the theorem is valid only in the case with L_1 norm penalty.

Theorem 1 *The minimizer of the penalized loss function $f_1(\mathbf{W}) = \|\mathbf{Y} - \mathbf{XW}\|^2 + \|(\mathbf{W} - \mathbf{T})\mathbf{A}\|_1$ is given by*

$$w_{jl} = \begin{cases} \hat{w}_{jl} - \frac{\lambda_j}{2\|\mathbf{x}_j\|^2} & \left(\hat{w}_{jl} > t_{jl} + \frac{\lambda_j}{2\|\mathbf{x}_j\|^2} \right) \\ \hat{w}_{jl} + \frac{\lambda_j}{2\|\mathbf{x}_j\|^2} & \left(\hat{w}_{jl} < t_{jl} - \frac{\lambda_j}{2\|\mathbf{x}_j\|^2} \right) \\ t_{jl} & \left(|w_{jl}| \leq t_{jl} + \frac{\lambda_j}{2\|\mathbf{x}_j\|^2} \right) \end{cases} \quad (4)$$

using $\hat{\mathbf{W}} = \{\hat{w}_{kl}\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, where t_{jl} is the (j, l) th element of \mathbf{T} corresponding to w_{jl} , the (j, l) th element of \mathbf{W} .

Proof. $f_1(\mathbf{W})$ can be rewritten as

$$f_1(\mathbf{W}) = \sum_{l=1}^q \sum_{j=1}^p \left(\|\mathbf{y}_l - w_{jl}\mathbf{x}_j\|^2 + \lambda_k |w_{jl} - t_{jl}| \right). \quad (5)$$

For minimizing it, consider the following three cases: [1] $w_{jl} > t_{jl}$, [2] $w_{jl} < t_{jl}$, and [3]otherwise. In case [1], $f_1(\mathbf{W})$ is reduced to

$$f_1(\mathbf{W}) = \sum_{l=1}^q \sum_{j=1}^p \left(\|\mathbf{y}_l - w_{jl}\mathbf{x}_j\|^2 + \lambda_j (w_{jl} - t_{jl}) \right) \quad (6)$$

and therefore its minimizer in the case is obtained as follows:

$$\frac{\partial f_1(\mathbf{W})}{\partial w_{jl}} = -2\mathbf{y}'_l \mathbf{x}_k + 2w_{jl}\|\mathbf{x}_j\|^2 + \lambda_k = 0 \Leftrightarrow w_{jl} = \hat{w}_{jl} - \frac{\lambda_j}{2\|\mathbf{x}_j\|^2}. \quad (7)$$

Further, the condition can be rewritten as

$$\hat{w}_{jl} - \frac{\lambda_j}{2\|\mathbf{x}_j\|^2} > t_{jl} \Leftrightarrow \hat{w}_{jl} > t_{jl} + \frac{\lambda_j}{2\|\mathbf{x}_j\|^2}, \quad (8)$$

which leads the first case of equation (4). In the similar manner, the minimizer in case [2] can be given by $w_{jl} = \hat{w}_{jl} - \frac{\lambda_j}{2\|\mathbf{x}_j\|^2}$, and the condition indicates that $\hat{w}_{jl} < t_{jl} - \frac{\lambda_j}{2\|\mathbf{x}_j\|^2}$. Case [3] indicates $w_{jl} = t_{kl}$, and its condition is $|w_{jl}| \leq t_{jl} + \frac{\lambda_j}{2\|\mathbf{x}_j\|^2}$, which is not covered by the first and second conditions. These results immediately lead to equation (4). \square

Equation (4) can be simply expressed as below, with $(x)_+$ being x if $x \geq 0$ and $-x$ otherwise:

$$w_{jl} = \text{sign}(\hat{w}_{jl} - t_{jl}) \left(|\hat{w}_{jl} - t_{jl}| - \frac{\lambda_j}{2\|\mathbf{x}_j\|^2} \right)_+ + t_{jl} \quad (9)$$

It should be noted that equation (4) is a generalization of LASSO solution in that both are equivalent when $\mathbf{T} = \mathbf{O}_{p \times q}$, where $\mathbf{O}_{p \times q}$ denotes the $p \times q$ matrix filled with 0s. This implies that a LASSO penalty function is aimed to shrink a solution matrix toward the zero matrix with the same dimension.

2.2 Applications in Machine Learning Problems

The above minimization procedure can be applied to SPCA, in which a component loading matrix \mathbf{A} is sparsely estimated. It is formulated as the minimization of

$$f_{SPCA}(\mathbf{F}, \mathbf{A}) = \|\mathbf{Z} - \mathbf{F}\mathbf{A}'\|^2 + \|(\mathbf{A} - \mathbf{T})\mathbf{A}\|_1 \quad (10)$$

subject to $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_r$. Note that it is equivalent to SCoTLASS (Jolliffe et al. 2003) if we set $\mathbf{T} = \mathbf{O}$. The minimization of $f_{SPCA}(\mathbf{F}, \mathbf{A})$ is attained by repeating the following two steps until the decrement of the function value converges, starting from suitable initial values for \mathbf{F}, \mathbf{A} :

1. Update \mathbf{F} by $\mathbf{F} = n^{1/2}\mathbf{K}_r$, where \mathbf{K}_r ($n \times r$) denotes the matrix of r right eigenvectors corresponding the r -largest eigenvalues of \mathbf{X} .
2. Update \mathbf{A} by equation (9) with setting $\mathbf{W} \rightarrow \mathbf{A}', \mathbf{X} \rightarrow \mathbf{F}$, and $\mathbf{Y} \rightarrow \mathbf{Z}$.

The example of the SPCA procedure is shown in the next section.

Also, the proposed procedure is applicable for three-mode component analysis (Kroonenberg 2008). Tucker3 (Tucker 1966; Kroonenberg and De Leeuw 1980) and CANDECOMP/PARAFAC (Carroll and Chang 1970) models are common choices for the problem. It is known that the former is less restrictive but difficult to interpret its result, while the latter is easier to interpret but too restrictive. It is formally expressed as follows. Let $\bar{\mathbf{X}}$ be an $n \times p \times k$ data array. Tucker3 and PARAFAC models either compress the first, second, and third modes into $s, t,$ and u components, respectively, and an $s \times t \times u$ array $\bar{\mathbf{C}}$ termed as a core array is obtained. Note that $s = t = u$ in PARAFAC model. In PARAFAC, the core array is restricted to be super-diagonal array and easier to be interpreted in that it contains several zero elements; the number of linkages between the components is limited. On the other hand, no constraint is imposed on the array in Tucker3 and thus often difficult to interpret (Frølich et al. 2018). In terms of fitness to the data array, Tucker3 is better because it has more unknown parameters than the other.

Here, Procrustes penalty function is used to an intermediate solution between two models in order to balance interpretability and fitness to the data array. It is accomplished by minimizing

$$f_{TP}(\mathbf{G}, \mathbf{C}, \mathbf{H}, \mathbf{E}) = \|\mathbf{Z} - \mathbf{GC}(\mathbf{H} \otimes \mathbf{E})'\|^2 + \|(\mathbf{C} - \mathbf{T})\mathbf{\Lambda}\|_1 \quad (11)$$

where the former is Tucker3's loss function. Above, $\mathbf{G}(n \times s)$, $\mathbf{H}(p \times t)$, and $\mathbf{E}(k \times u)$ are unknown component loading matrices of the first, second, third mode of $\tilde{\mathbf{X}}$, respectively. $\mathbf{Z} = \{\mathbf{X}_1, \dots, \mathbf{X}_k\}$ is the $n \times pk$ matrix of the horizontal slices of $\tilde{\mathbf{X}}$. Also, $\mathbf{C} = \{\mathbf{C}_1, \dots, \mathbf{C}_k\}$ is the $s \times tu$ matrix composed of the horizontal slices of $\tilde{\mathbf{C}}$, and it is matched to \mathbf{T} having a suitable dimension. Here, consider to set \mathbf{T} as the matrix of horizontal slices of the core array estimated by PARAFAC. The estimated core array by minimizing equation (11) therefore shrinks toward the one by PARAFAC, and thus the array is considered to be in the middle of PARAFAC and Tucker3. The relative strength of Procrustes penalty to the Tucker3's loss function is controlled by $\mathbf{\Lambda}$; $\mathbf{\Lambda}$ having high values for its diagonal elements leads to a more PARAFAC-like core array and vice versa.

The minimization of equation (11) is accomplished as follows. Nelder-Mead's numerical optimization is used for minimizing $f_{TP}(\mathbf{G}, \mathbf{C}, \mathbf{H}, \mathbf{E})$ over \mathbf{G} , \mathbf{H} , and \mathbf{E} . The core array minimizing the loss function is obtained by extending Theorem 1 to Penrose regression in which

$$f_{Pen}(\mathbf{W}) = \|\mathbf{Y} - \mathbf{X}_1 \mathbf{W} \mathbf{X}_2\|^2 + \|\mathbf{\Lambda}(\mathbf{W} - \mathbf{T})\|^2 \quad (12)$$

is minimized over \mathbf{W} .

Corollary 1 *The minimizer of the loss function in Equation (12) over \mathbf{W} is given by*

$$w_{jk} = \text{sign}(\tilde{w}_{jk} - t_{jk}) \left(|\tilde{w}_{jk} - t_{jk}| - \frac{\lambda_j}{2\|\mathbf{x}_j^{(1)}\|^2\|\mathbf{x}_k^{(2)}\|^2} \right)_+ + t_{jk} \quad (13)$$

where $\mathbf{x}_j^{(1)}$ and $\mathbf{x}_k^{(2)}$ denote the j th and k th columns of \mathbf{X}_1 and \mathbf{X}_2 , respectively.

Proof. It can be easily verified in the same way as Theorem 1. \square

Using the corollary, the optimal \mathbf{C} is obtained by setting $\mathbf{X}_1 \rightarrow \mathbf{G}$ and $\mathbf{X}_2 \rightarrow \mathbf{H} \otimes \mathbf{E}$. The solution for the minimization of equation (11) is obtained by repeating the following steps, updating \mathbf{C} by (13), and updating the other parameter matrices by numerical optimization of the loss function.

3 Illustrations

The proposed procedures are illustrated in order to show how well they work in dealing with real datasets.

3.1 SPCA to Wine Data

The SPCA procedure combined with a Procrustes penalty function is to be demonstrated. The dataset used here is wine data (Dua and Karra Taniskidou 2017) that consists of 178 samples (wines) and 13 variables (chemical ingredients), and it is available at UCI Machine Learning Repository.

First, the target matrix \mathbf{T} was specified by the following manner. An original principal component analysis with three components was applied to the dataset, and the score and loading matrices were obtained denoted as $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{A}}$, respectively. Let $\tilde{\mathbf{A}}^\sharp = \{\tilde{a}_{jk}^\sharp\}$ be a varimax-rotated $\tilde{\mathbf{A}}$, and \mathbf{T} was set by

$$t_{jk} = \begin{cases} \text{sign}(a_{jk}^\sharp) & (|\tilde{a}_{jk}^\sharp| > \tau) \\ 0 & (\text{otherwise}) \end{cases} \quad (14)$$

where τ is a suitable threshold. The value of τ was set at 0.4 in this example. \mathbf{T} is considered as a possible simple structure because it is based on varimax-rotated loading matrix and its simplicity was further emphasized by substituting all elements with ± 1 or zeros.

We restricted $\lambda_k = \lambda$ for $k = 1, \dots, r$ for the simplicity of the example and estimated \mathbf{A} s for λ s within the range of $\lambda \in [20, 250]$. Figure 1 shows solution paths of \mathbf{A} where each of the estimated elements of \mathbf{A} is plotted against λ s. It can be seen that some of the elements shrink toward zero as λ increases, while the others shrink toward 1 or -1 . The figure indicates that most of the elements in \mathbf{A} take 0, 1, or -1 when λ is large, and therefore λ controls the resulting simplicity of the

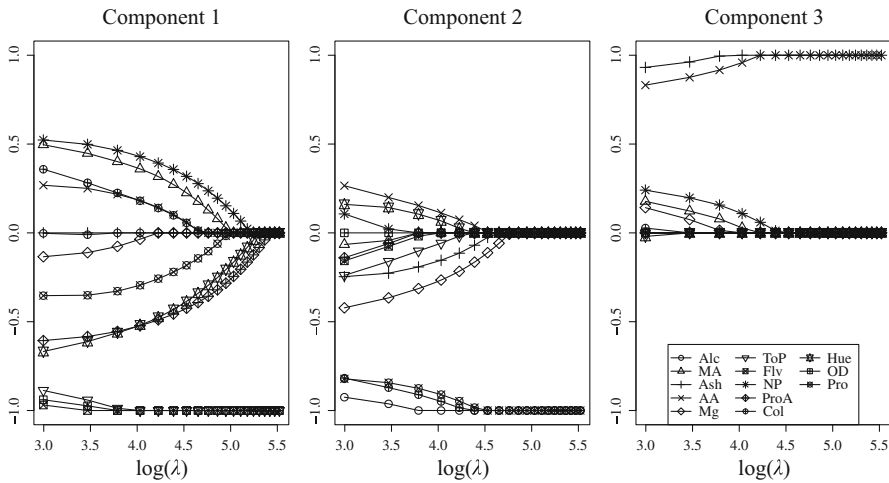


Fig. 1 Solution paths for three components; names of the 13 variables are shortened. The y-axes stand for the value of elements. The legend is common for all the plots

solution. To illustrate this, we picked up some λ s within the range of λ s and showed the estimated \mathbf{A} s with the target matrix \mathbf{T} as Table 1. The correspondences between the variables and the components are clearly captured, because the estimated loading matrix contains many exact zero elements when λ is sufficiently large and also has simple structure as well as \mathbf{T} . For example, referring \mathbf{A} with $\lambda = 50$, the second component is characterized by *alcohol* (Alc) and *color intensity* (Col) and *proline* (Pro), but it is not clearly captured in the one with $\lambda = 20$.

3.1.1 Three-Mode Component Analysis to Multiple Personality Data

The proposed method is also used for estimating an intermediate core array between Tucker3 (unconstrained) and PARAFAC (constrained to be super-diagonal). The three-mode component analysis with Procrustes penalty function was applied to multiple personality data (Kroonenberg 2008) in which 15 concepts were evaluated by 10 scales by 6 personalities. The data array $\bar{\mathbf{X}}$ thus has a dimension of $13 \times 10 \times 6$, and see Osgood and Luria (1954) for details of the data. To set a target matrix \mathbf{T} , PARAFAC with $s = t = u = 2$ was firstly applied to $\bar{\mathbf{X}}$, and the resulting $2 \times 2 \times 2$ core array $\bar{\mathbf{C}}$ is transformed into 2×4 a target matrix \mathbf{T} . For simplicity, the isotropic penalty parameters were used: $\mathbf{\Lambda} = \lambda \mathbf{I}_2$. With the target, the proposed method was applied to the data, using the penalty parameters $\lambda = 1, 10, \text{ and } 100$.

Table 2 shows frontal slices of the estimated core arrays together with Tucker3 ($\lambda = 0$) and PARAFAC ($\lambda = \infty$) solutions. There are two extreme cases, Tucker3 and PARAFAC, and components are fully connected in the former solution, while they are sparsely connected in the latter. The proposed method allows to obtain intermediate solutions between them. As λ gets increased, the solution gradually approximates PARAFAC solution that is the most restrictive but interpretable core array. Users freely choose appropriate value of λ , by checking resulting simplicity of the core array. If one considers $\lambda = 10$ as appropriate, connections between several components can be ignored, while all connections have to be interpreted in Tucker3.

4 Conclusions

The research proposes a new penalty function for penalized optimization in multivariate analysis procedures. It is termed as Procrustes penalty function, and it shrinks a solution matrix to a prespecified target matrix. The target matrix possesses a certain desired structure which the resulting solution should possess, such as a simple structure. If a target with a simple structure is employed, the resulting solution matrix is thus simple and interpreted, while the existing penalty functions hardly consider the matrix-wise simplicity of a solution matrix. The proposed method is applicable to various multivariate analysis procedures, in order to obtain an interpretable solution matrix or other purposes. We used the Procrustes penalty in

Table 1 Estimated A_s for $\lambda = 20, 50,$ and 100 and the target matrix. Blanc cells indicate exact zero elements

	$\lambda = 20$			$\lambda = 50$			$\lambda = 100$			Target matrix T		
	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3
	Alc	-0.003	-0.925			-1.000			-1.000			-1.000
MA	0.497	-0.066	0.177	0.381		0.055	0.196					
Ash		-0.247	0.931		-0.174	1.000			1.000			1.000
AA	0.268	0.265	0.832	0.201	0.132	0.937	0.031		1.000			1.000
Mg	-0.134	-0.422	0.142	-0.058	-0.290			-0.079				
ToP	-0.887	-0.239		-1.000	-0.078					-1.000		
Flv	-0.969	-0.157		-1.000						-1.000		
NP	0.523	0.107	0.241	0.448		0.135	0.293					
ProA	-0.606	-0.140		-0.540			-0.405					
Col	0.358	-0.818	0.027	0.201	-0.930		0.030		-1.000		-1.000	
Hur	-0.668	0.161	-0.019	-0.540	0.090		-0.358					
OD	-0.939			-1.000						-1.000		
Pro	-0.354	-0.822		-0.314	-0.891		-0.159		-1.000		-1.000	

Table 2 Estimated core arrays by Tucker3, PARAFAC, and the proposed method with three λ s. Blanc cells stand for exact zero elements

	Mode2 scales	Mode3 personalities			
		Comp.1		Comp.2	
		Comp.1	Comp.2	Comp.1	Comp.2
Tucker3	Mode1 concepts				
	Comp.1	-15.803	-5.723	3.437	-3.363
	Comp.2	4.423	-12.243	-1.229	-2.460
$\lambda = 1$	Comp.1	-5.185		-0.189	-2.404
	Comp.2	1.295	-0.934		-3.008
$\lambda = 10$	Comp.1	-0.277		0.088	0.075
	Comp.2	0.022	-0.305		0.266
$\lambda = 100$	Comp.1	0.582			
	Comp.2				0.493
PARAFAC	Comp.1	0.582			
	Comp.2				0.493

SPCA and three-mode component analysis, aiming to simplify the loading matrix in the former and to obtain an intermediate solution between Tucker3 and PARAFAC.

So far, the author treated the target matrix \mathbf{T} as a prespecified matrix need to be fixed. \mathbf{T} is, however, possible to be estimated jointly as well as other parameter matrices. There exist some cases where it is difficult to specify \mathbf{T} in advance, and such extension of the proposed method serves to relax the hurdle for using the Procrustes penalty function.

References

- 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.
- Dua, D., & Karra Taniskidou, E. (2017). UCI machine Learning Repository. <http://archive.ics.uci.edu/ml>.
- Frølich, L., Andersen, T. S., & Mørup, M. (2018). Rigorous optimisation of multilinear discriminant analysis with Tucker and PARAFAC structures. *BMC Bioinformatics*, 19(1), 197.
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Ratón: CRC Press.
- Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3), 531–547. <https://doi.org/10.1198/1061860032148>.
- Kroonenberg, P. M. (2008). *Applied multiway data analysis* (Vol. 702). Hoboken: Wiley.
- Kroonenberg, P. M., & De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45(1), 69–97.
- Osgood, C. E., & Luria, Z. (1954). A blind analysis of a case of multiple personality using the semantic differential. *The Journal of Abnormal and Social Psychology*, 49(4p1), 579.

- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206.
- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*(3), 279–311.
- Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.

Factor Score Estimation from the Perspective of Item Response Theory



David Thissen and Anne Thissen-Roe

Abstract The factor scores of confirmatory factor analysis (CFA) models and the latent variables of item response theory (IRT) models are similar statistical entities, so one would expect that their estimation or characterization would follow parallel tracks in CFA and IRT. However, historically they have not. Different procedures have been used to derive factor score estimates and latent variable estimates in IRT, and different computational procedures have been the result. In this chapter we approach factor score estimation for some simple CFA models from the perspective of IRT, with the kinds of graphics that are used to explain IRT estimates of proficiency, and the computational procedures that are used in test theory. We compare traditional “regression” and “Bartlett” factor score estimates with alternative computational approaches to likelihood-based factor score estimates, referring to the expected a posteriori and maximum likelihood estimates of IRT latent variables to clarify relations among the scores. This provides insights into the ways in which the data are combined into factor score estimates. The results provide an alternative method to compute factor scores in some simple models in the presence of observations that may be missing at random for some variables.

Keywords Factor scores · Item response theory

1 Introduction

It can be useful in explaining item response theory (IRT) to say that IRT scores are the same as factor score estimates: point estimates summarizing the location of a

D. Thissen (✉)

Department of Psychology and Neuroscience, The University of North Carolina
at Chapel Hill, Chapel Hill, NC, USA
e-mail: dthissen@email.unc.edu

A. Thissen-Roe

pymetrics, New York, NY, USA
e-mail: anne@pymetrics.com

likelihood or posterior for a latent variable. While this statement can be helpful if the students of IRT already know something about factor analysis (or vice versa), its usefulness is limited by the fact that neither the IRT nor the factor analysis literature makes the analogy clear in any detail. Procedures to compute factor score estimates originated with ideas based on regression (Thomson 1935, 1936, 1938; Thurstone 1935; Bartlett 1937), and the usual textbook presentation of factor analysis uses those lines of argument. The literature on factor score estimates (that is not preoccupied with factor score indeterminacy Grice 2001) is about the *properties* of the estimates: whether they are conditionally unbiased, whether they have the right variances and correlations with each other or other variables, etc.; for examples see Skrondal and Laake (2001); Devlieger et al. (2015); or Hoshino and Bentler (2013). This presentation is not about those topics, but rather “What if you want factor scores estimates to use like test scores, assuming you know the structural parameters from previous large scale calibration and you probably want to have a method tolerant of observations missing at random?” Mardia et al. (1979), Hoshino and Bentler (2013), Estabrook and Neale (2013), and Loncke et al. (2018) touch on this topic, and Skrondal and Rabe-Hesketh (2004) have a chapter on it, but it has not been salient in the literature on factor analysis.

Our intention is to provide an explanation that is useful for pedagogy, for both IRT and factor analysis, and that supports the use of factor score estimates for the usual purposes of test scores: reporting or subsequent analysis. There are contexts in which a large number of (effectively) continuous variables (say, a few dozen) serve as indicators for a smaller number (say, less than a dozen) latent variables or factors. If it is desirable in such a context to report something like scores to the respondents, or to have summaries for subsequent analyses, the use of factor score estimates by direct analogy with IRT score computations presents itself.

In this chapter we trace the development of factor score estimates from the same likelihood principles as are used for IRT scores, to make the direct analogy clear; then we show which regression methods are the same as IRT scores. In the process, we describe ways the factor score estimates can be computed with some of the observations missing at random, again in parallel to standard IRT methods to score around missing item responses. And we illustrate the computations with graphics that are rarely used in the factor analytic literature.

2 Unidimensional Likelihood-Based Score Estimates

2.1 IRT Scoring

Thissen and Orlando (2001) and Thissen et al. (2001) summarize decades of work that has culminated in contemporary IRT test scoring. A common (if misguided) IRT convention is to refer to the latent variable measured by a test as θ (or $\boldsymbol{\theta}$ if multidimensional; we will discuss only unidimensional models in this section). The basic element of the model is the item response function $T(u_{ij}|\theta)$, the line tracing

the probability of categorical response u (that may be dichotomous or polytomous) of person i to item j as a function of θ . In this section (only) we use the most common IRT notation, even though the use of θ conflicts with its reuse in the factor analytic context in subsequent sections; we assume the context will make meaning clear to the reader.

Under IRT's defining assumption of conditional or local independence, the likelihood of the set of responses from person i is

$$L(\mathbf{u}_i|\theta) = \prod_j T(u_{ij}|\theta). \quad (1)$$

Historically the mode of equation 1, the *maximum likelihood* (ML) estimate, has been used as an IRT test score, because it can be described as the most likely value of θ given the response pattern \mathbf{u} .

However, the ML estimate has disadvantages, chief among which is that it is not finite for some response patterns for commonly used IRT models. Equation 1 is also an incomplete representation of the model. There must also be some density $\phi(\theta)$ for the latent variable itself. Including that, a more complete likelihood is

$$L_p(\theta|\mathbf{u}_i) \propto \prod_j T(u_{ij}|\theta)\phi(\theta). \quad (2)$$

Equation 2 is often referred to as the *posterior* density for θ because of the formal analogy of the equation with likelihood-times-prior in Bayesian analysis. That makes the population distribution for the latent variable, $\phi(\theta)$, a prior distribution, which it really is not. $\phi(\theta)$ is part of the model for the categorical responses. However, again for historical reasons, the nomenclature has become so solidified that IRT estimates based on equation 2 are referred to as a posteriori so the modal estimate is the maximum a posteriori (MAP) and the mean of equation 2 is the expected a posteriori (EAP) estimate. The EAP estimate has the advantage of minimizing squared error, but can be challenging to compute for models with higher-dimensional θ , for which the modal estimate remains valuable.

Unidimensional IRT score computation is illustrated on the left side of Fig. 1. The mode of the blue likelihood in the lower panel is the ML estimate. The mode of the magenta "posterior" is the MAP estimate, and the mean of that curve is the EAP estimate. The two modal estimates are computed with an optimization method, usually Newton-Raphson; the EAP is computed using numerical integration.

2.2 Unidimensional Factor Analysis

2.2.1 The One-Factor Model

To be parallel with unidimensional IRT, we begin with the one-factor model for continuous response y_{ij} for person i and observed variable j ,

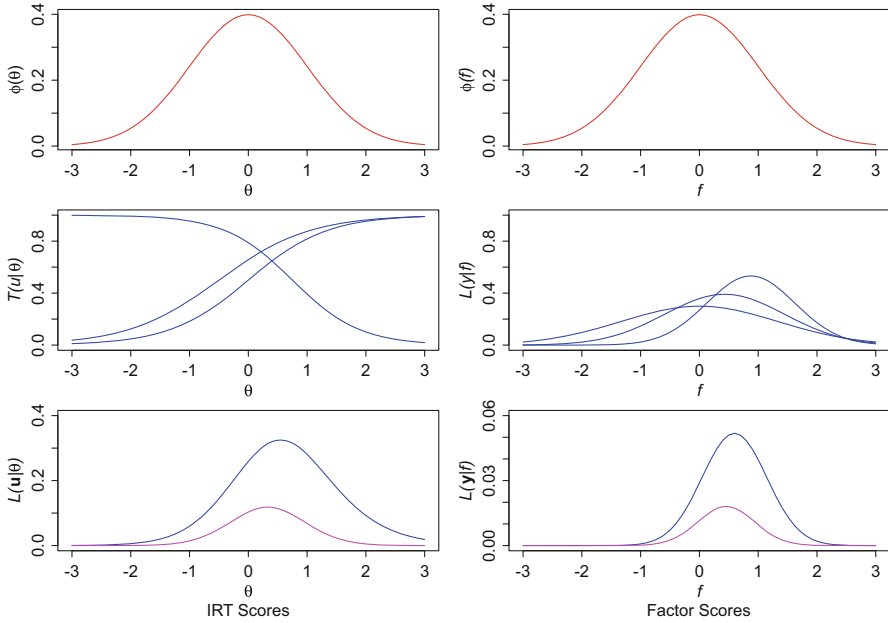


Fig. 1 Left side: IRT scoring. Upper panel: The population distribution $\phi(\theta)$, usually $N(0, 1)$ for IRT models. Center panel: Three trace lines $T(u|\theta)$, two for correct or positive responses to dichotomous items and one incorrect or negative. Lower panel: The blue curve is the likelihood $L(u_i|\theta)$ from equation 1, the product of the three trace lines in the center panel; the magenta curve is the posterior, $L_p(\theta|u_i)$ from equation 2, which is the blue likelihood times the red population distribution in the upper panel. **Right side: Factor scoring.** Upper panel: The population distribution $\phi(f)$, often $N(0, 1)$ for CFA models. Center panel: Three Gaussian likelihoods $L(y_{ij}|f_i)$ for three values of “ $y = [0.0, 0.3, 0.7]$ ” and $\lambda = [0.6, 0.7, 0.8]$ ”. Lower panel: The blue curve is the likelihood $L(y_j|f)$ from equation 5, the product of the three likelihoods in the center panel; the magenta curve is the posterior, $L_p(f|y_i)$ from equation 6, which is the blue likelihood times the red population distribution in the upper panel

$$y_{ij} = \lambda_j f_i + \epsilon_{ij}, \tag{3}$$

in which the observations y_{ij} and the factor scores f_i are both assumed to be standardized (hence the absence of an intercept in equation 3). λ_j is the regression parameter (or factor loading) for y_j on f (and also the correlation for standardized y and f), and ϵ_{ij} is $N(0, \theta_j)$ in which θ_j is the unique, or error, variance for observed variable j . Note that with everything standardized, $\theta_j = 1 - \lambda_j^2$. [Also note the reuse of the notation θ here following widely conventional notation used with structural equation modeling; the meaning differs from θ in IRT in the previous section.]

The context for computing factor score estimates is one in which the (usually previously estimated) values of λ_j and θ_j are taken to be fixed and known, just as it is for IRT scoring using item parameters from calibration.

The (Gaussian) likelihood for response y_{ij} , analogous to the IRT trace line, is

$$L(y_{ij}|f_i) = \phi(y_{ij}|f_i) = \frac{1}{\sqrt{2\pi\theta_j}} e^{-\frac{(y_{ij}-\lambda_j f_i)^2}{2\theta_j}}. \quad (4)$$

Assuming local independence, the (also Gaussian) likelihood of the vector of responses \mathbf{y}_i for person i is

$$L(\mathbf{y}_i|f_i) = \prod_j L(y_{ij}|f_i). \quad (5)$$

Equation 5 is directly analogous to equation 1 from IRT. It ignores the population distribution for simplicity. This is harmless; unlike in IRT, for all patterns of observed responses, estimates can still be computed. However, factor analysis often makes use of the assumption that the factor scores f are normally distributed. The likelihood that includes the population distribution is

$$L_p(f_i|\mathbf{y}_i) \propto \prod_j L(y_{ij}|f_i)\phi(f), \quad (6)$$

analogous to equation 2.

Because both likelihoods are Gaussian, the modes and means are the same, and derivational and computational approaches to compute either provide the same factor score estimates.

2.2.2 Modal (or Maximum Likelihood or ML, or MAP) Estimation

Mardia et al. (1979, p. 274) point out that likelihood-based factor score estimates are the same as Bartlett's and Thomson's regression-based factor score estimates, and Skrondal and Rabe-Hesketh (2004, p. 239) make the same observation about Bartlett's estimates. Hoshino and Bentler (2013, p. 47) observe about "(1) Bartlett's method and (2) the regression method. The former can be considered the ML estimator of the factor score vector in the fixed effect factor analysis, while the latter can be regarded as the Bayes posterior mean estimator." Estabrook and Neale (2013) discuss the performance of likelihood-based estimates of factor scores, as compared with Bartlett's method. None of those sources provide much detail about their derivation or computation.

The maximum likelihood, or ML, factor score estimate can be computed by locating the modal value of the likelihood $L(\mathbf{y}_i|f)$ in equation 5. Iterative computation of the mode of a likelihood is usually done with a Newton-Raphson algorithm applied to the log likelihood. In this case, the log likelihood is

$$\ell = \log L(\mathbf{y}_i | f_i) = \sum_j \log L(y_{ij} | f_i) \propto \sum_j \frac{-(y_{ij} - \lambda_j f_i)^2}{2\theta_j}; \quad (7)$$

the maximum is the value of f where the first derivative of ℓ equals zero:

$$\frac{\partial \ell}{\partial f} = \sum_j \frac{\lambda_j (y_{ij} - \lambda_j f_i)}{\theta_j} = 0. \quad (8)$$

The Newton-Raphson algorithm locates the maximum as the convergence of a sequence of values defined by

$$f_{\text{next}} = f_{\text{current}} - \frac{\partial \ell}{\partial f} / \frac{\partial^2 \ell}{\partial f^2}, \quad (9)$$

in which the second derivative is

$$\frac{\partial^2 \ell}{\partial f^2} = \sum_j \frac{-\lambda_j^2}{\theta_j}. \quad (10)$$

Because the likelihood is Gaussian, the log likelihood is quadratic, and Newton-Raphson always converges in one step to the mode, which is also the mean. Following standard likelihood theory, the error variance of the estimate is the negative inverse of the second derivative (equation 10).

The log likelihood and its derivatives for L_p , including a normal population distribution, add terms to equations 7, 8, and 10. Then procedures to compute analogs to IRT's MAP and EAP estimation can make use of either Newton-Raphson iteration to the mode (which requires only one evaluation of the derivatives) or numerical integration (which is more computationally intensive).

For the special case of a single observed variable, equation 8 can be solved by visual inspection to yield the mean value of f for an item given the response, $\mu_{f|y_{ij}} = y_{ij}/\lambda_j$; the associated variance comes from equation 10: $\sigma_{f|y_{ij}}^2 = \theta_j/\lambda_j^2$. These can be used to plot graphics for factor score estimation parallel to the left side of Fig. 1, as shown on the right side.

2.2.3 Quick Closed Form Computation for Models in which Each Variable Is Associated with Only One Factor

It is sometimes desirable to compute factor score estimates in the presence of individual observed variables that may be missing at random. That can be done with any of the procedures described in the preceding or subsequent sections. For the IRT-like likelihood-based computations described in Sect. 2.2.2, one simply omits the terms in the log likelihood and derivatives associated with the missing observa-

tion(s). Regression-based solutions (to follow in Sect. 3.4) require recomputation of the regression coefficients for the subset of variables observed for a given person. There is a shortcut for some simple models: those in which each observed variable is associated with only one factor. The simplest special case is the unidimensional model.

The likelihood in equation 5 is a product of Gaussian likelihoods. The mean of a product of Gaussian likelihoods is the average of the means of the component normal distributions, each weighted by the inverse of the associated variance. We determined at the end of the preceding section that the means and variances of the likelihoods for each response are $\mu_{f|y_{ij}} = y_{ij}/\lambda_j$ and $\sigma_{f|y_{ij}}^2 = \theta_j/\lambda_j^2$.

Then a closed form computation of the ML estimate \hat{f}_i is

$$\hat{f}_i = \frac{\sum_j \frac{\mu_{f|y_{ij}}}{\sigma_{f|y_{ij}}^2}}{\sum_j \frac{1}{\sigma_{f|y_{ij}}^2}} = \frac{\sum_j \frac{y_{ij}/\lambda_j}{\theta_j/\lambda_j^2}}{\sum_j \frac{1}{\theta_j/\lambda_j^2}} \quad (11)$$

and the associated error variance is

$$\sigma_{f_i}^2 = \frac{1}{\sum_j \frac{1}{\sigma_{f|y_{ij}}^2}} = \frac{1}{\sum_j \frac{1}{\theta_j/\lambda_j^2}}. \quad (12)$$

The summations run over all non-missing responses.

3 Multidimensional Likelihood-Based and Regression-Based Score Estimates

3.1 The Multiple Factor Model

Express the multiple factor model for the vector of p observed responses \mathbf{y}_i for person i as

$$\mathbf{y}_i = \mathbf{A}\mathbf{f}_i + \boldsymbol{\epsilon}_i, \quad (13)$$

in which the observations \mathbf{y}_i and the vector of k factor scores \mathbf{f}_i are standardized, \mathbf{A} is $p \times k$ matrix of regression coefficients (or factor loadings) for \mathbf{y}_i on \mathbf{f}_i , and $\boldsymbol{\epsilon}$ is multivariate $N(\mathbf{0}, \boldsymbol{\Theta})$ in which $\boldsymbol{\Theta}$ is the variance-covariance matrix of the residuals (or errors or “unique factors”).¹

¹Often this would quickly be changed into a model for the covariance matrix among the observations, $\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Theta}$, in which $\boldsymbol{\Phi}$ is the covariance (here, correlation) matrix among the factors, for estimation of the parameters in \mathbf{A} and $\boldsymbol{\Theta}$ by Wishart maximum likelihood. However,

Then the likelihood for the data as a function of the factor scores (assuming \mathbf{A} and Θ are known) is

$$L(\mathbf{y}_i | \mathbf{f}_i) = \frac{|\Theta|^{-\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}(\mathbf{y}_i - \mathbf{A}\mathbf{f}_i)' \Theta^{-1} (\mathbf{y}_i - \mathbf{A}\mathbf{f}_i)}. \quad (14)$$

The log likelihood is then proportional to

$$\ell = \log L(\mathbf{y}_i | \mathbf{f}_i) \propto -\frac{1}{2}(\mathbf{y}_i - \mathbf{A}\mathbf{f}_i)' \Theta^{-1} (\mathbf{y}_i - \mathbf{A}\mathbf{f}_i). \quad (15)$$

If a standard normal population distribution for \mathbf{f} is included, with correlation matrix Φ among the factors, the likelihood becomes

$$L_p(\mathbf{f}_i | \mathbf{y}_i) \propto \left[\frac{|\Theta|^{-\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}(\mathbf{y}_i - \mathbf{A}\mathbf{f}_i)' \Theta^{-1} (\mathbf{y}_i - \mathbf{A}\mathbf{f}_i)} \right] \left[\frac{|\Phi|^{-\frac{1}{2}}}{(2\pi)^{\frac{k}{2}}} e^{-\frac{1}{2}\mathbf{f}_i' \Phi \mathbf{f}_i} \right]; \quad (16)$$

the log likelihood is then proportional to

$$\ell_p = \log L_p(\mathbf{f}_i | \mathbf{y}_i) \propto -\frac{1}{2}(\mathbf{y}_i - \mathbf{A}\mathbf{f}_i)' \Theta^{-1} (\mathbf{y}_i - \mathbf{A}\mathbf{f}_i) - \frac{1}{2}\mathbf{f}_i' \Phi \mathbf{f}_i. \quad (17)$$

3.2 On the Normal Likelihood of the Multiple Factor Model

An IRT-style graphical representation of the combination of the likelihoods for individual variables that constitute the likelihood in equation 14 for a person's vector response is shown in Fig. 2. The illustration is for three hypothetical variables in a two factor solution with loadings

$$\mathbf{A} = \begin{bmatrix} 0.60 & 0.00 \\ 0.60 & 0.60 \\ 0.75 & -0.50 \end{bmatrix}. \quad (18)$$

In the left panel of Fig. 2, color density represents the likelihood of responses $\mathbf{y}' = [0.1 \ 0.2 \ 0.3]$ to these three items. The variables in the order of rows in equation 18 are shown as magenta, yellow, and cyan. Each colored band is a unidimensional normal curve in cross section, with a ridge along a line that is the mean for one factor conditional on the value of the other. The magenta ridge is vertical, because the response to the first variable is related only to f_1 , not f_2 . The ridge for the

here we are concerned with the factor scores \mathbf{f} , treating \mathbf{A} , Φ , and Θ as fixed and known, so we mention this only to clarify notation.

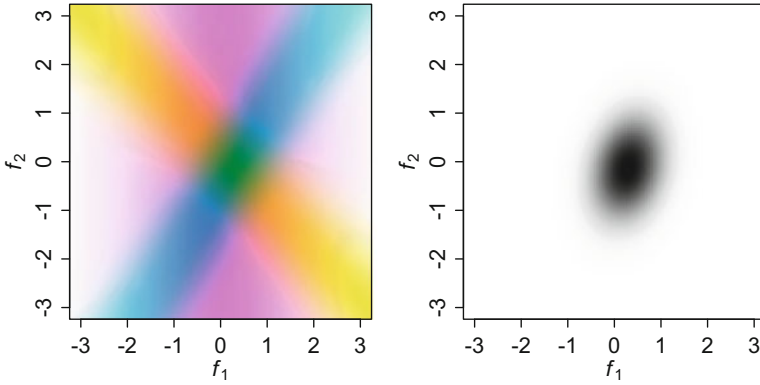


Fig. 2 Left panel: Color density represents the likelihood of responses $y' = [0.1 \ 0.2 \ 0.3]$ to three items with factor loadings in equation 18. The variables in the order of rows in equation 18 are shown as magenta, yellow, and cyan. Right panel: Grayscale density indicates the joint likelihood of the three responses plotted separately in the left panel; this is the bivariate normal likelihood of equation 14

yellow band, representing variable 2, descends at a 45° angle because the two factor loadings are equal, so for a given response the value of f_2 must become equally lower as the value of f_1 increases. The third (cyan) band shows the likelihood for a variable positively related to f_1 and negatively related to f_2 .

As is the explicit case in IRT, the joint likelihood for the three responses is the product of the separate likelihoods, but in writing the likelihood for the factor model that product is buried in the multivariate normal density, equation 14. That product, or multivariate normal, likelihood is shown with grayscale density in the right panel of Fig. 2.

3.3 Modal Estimation

The matrix form of the log likelihood is equation 15; the vector derivative with respect to f is

$$\frac{\partial \ell}{\partial f} = \Lambda' \Theta^{-1} (y_i - \Lambda \hat{f}_i), \tag{19}$$

and the second derivative is

$$\frac{\partial^2 \ell}{\partial f^2} = -\Lambda' \Theta^{-1} \Lambda. \tag{20}$$

So the multivariate Newton update to obtain the vector ML estimate \hat{f}_i is

$$\hat{\mathbf{f}}_{i,\text{next}} = \hat{\mathbf{f}}_{i,\text{current}} - \left[\frac{\partial^2 \ell}{\partial \mathbf{f}} \right]^{-1} \frac{\partial \ell}{\partial \mathbf{f}}. \quad (21)$$

MAP estimates are obtained by adding the derivative components with respect to \mathbf{f} from the standard normal population distribution to equations 19 and 20, as follows:

$$\frac{\partial \ell_p}{\partial \mathbf{f}} = \mathbf{\Lambda}' \mathbf{\Theta}^{-1} (\mathbf{y}_i - \mathbf{\Lambda} \hat{\mathbf{f}}_i) + \mathbf{\Phi}^{-1} \hat{\mathbf{f}}_i, \quad (22)$$

and the second derivative is

$$\frac{\partial^2 \ell_p}{\partial \mathbf{f}} = -\mathbf{\Lambda}' \mathbf{\Theta}^{-1} \mathbf{\Lambda} - \mathbf{\Phi}^{-1}. \quad (23)$$

3.4 The Regression Solutions Derived from the Likelihood

With the log-likelihood in equation 15, we can also find the maximum likelihood estimates of \mathbf{f} by locating the minimum of Q

$$Q = \text{tr}((\mathbf{y}_i - \mathbf{\Lambda} \mathbf{f}_i)' \mathbf{\Theta}^{-1} (\mathbf{y}_i - \mathbf{\Lambda} \mathbf{f}_i)) \quad . \quad (24)$$

Following the standard derivation of multivariate regression coefficients (see, e.g., Bock (1975, pp. 168–170)), inverting the roles of the explanatory variables and the regression coefficients, because here we know the regression coefficients $\mathbf{\Lambda}$ and we want to estimate the explanatory variables \mathbf{f} , we expand equation 24 to become

$$Q = \text{tr}(\mathbf{y}_i \mathbf{\Theta}^{-1} \mathbf{y}_i - 2\mathbf{y}_i \mathbf{\Theta}^{-1} \mathbf{\Lambda} \mathbf{f}_i + \mathbf{f}_i' \mathbf{\Lambda}' \mathbf{\Theta}^{-1} \mathbf{\Lambda} \mathbf{f}_i). \quad (25)$$

Using matrix derivatives, the partial of Q with respect to \mathbf{f} is equal to zero at the minimum of Q

$$\frac{\partial Q}{\partial \mathbf{f}} = -2\mathbf{\Lambda}' \mathbf{\Theta}^{-1} \mathbf{y}_i + 2\mathbf{\Lambda}' \mathbf{\Theta}^{-1} \mathbf{\Lambda} \mathbf{f}_i = 0, \quad (26)$$

so

$$\mathbf{\Lambda}' \mathbf{\Theta}^{-1} \mathbf{\Lambda} \mathbf{f}_i = \mathbf{\Lambda}' \mathbf{\Theta}^{-1} \mathbf{y}_i, \quad (27)$$

and

$$\hat{\mathbf{f}}_i = (\mathbf{\Lambda}' \mathbf{\Theta}^{-1} \mathbf{\Lambda})^{-1} \mathbf{\Lambda}' \mathbf{\Theta}^{-1} \mathbf{y}_i, \quad (28)$$

in which $(\Lambda'\Theta^{-1}\Lambda)^{-1}\Lambda\Theta^{-1}$ is the formula for the (Bartlett 1937) regression solution. Standard regression results lead to the conclusion that the error covariance matrix for \hat{f}_i is

$$\Sigma_{f_i} = (\Lambda'\Theta^{-1}\Lambda)^{-1}. \quad (29)$$

If we repeat this exercise with the log-likelihood including the population distribution in equation 17, we have

$$Q_p = \text{tr}(\mathbf{y}_i - \Lambda\mathbf{f}_i)'\Theta^{-1}(\mathbf{y}_i - \Lambda\mathbf{f}_i) + \text{tr}(\mathbf{f}_i)'\Phi^{-1}\mathbf{f}_i. \quad (30)$$

Then the derivative of Q_p with respect to \mathbf{f} is equal to zero at the minimum of Q

$$\frac{\partial Q_p}{\partial \mathbf{f}} = -2\Lambda'\Theta^{-1}\mathbf{y}_i + 2\Lambda'\Theta^{-1}\Lambda\mathbf{f}_i + 2\Phi^{-1}\mathbf{f}_i = 0, \quad (31)$$

so

$$\Lambda'\Theta^{-1}\Lambda\mathbf{f}_i + \Phi^{-1}\mathbf{f}_i = \Lambda'\Theta^{-1}\mathbf{y}_i, \quad (32)$$

and

$$\hat{\mathbf{f}}_i = (\Lambda'\Theta^{-1}\Lambda + \Phi^{-1})^{-1}\Lambda'\Theta^{-1}\mathbf{y}_i, \quad (33)$$

in which $(\Lambda'\Theta^{-1}\Lambda + \Phi^{-1})^{-1}\Lambda\Theta^{-1}$ is the formula for the Thomson-Thurstone regression solution (Thomson 1935, 1936; Thurstone 1935). Standard regression results lead to the conclusion that the error covariance matrix for \hat{f}_i is

$$\Sigma_{f_i} = (\Lambda'\Theta^{-1}\Lambda + \Phi^{-1})^{-1}. \quad (34)$$

Mardia et al. (1979, p. 274) provide both the Bartlett and Thomson-Thurstone results from the log likelihood but without any intermediate steps. Estabrook and Neale (2013, p. 3) also provide the Bartlett result as well as the version of the Thomson-Thurstone coefficients with $\Phi=\mathbf{I}$ for orthogonal factors; Bartholomew et al. (2011, p. 48) also provide the latter, but without any reference to Thomson or Thurstone.

Thomson (1938, p. 246) understood that his estimator included the population distribution and Bartlett's did not; he wrote, explaining the difference between his estimates and Bartlett's, "My formulae were arrived at by the ordinary regression method. Bartlett's estimates and the regression estimates attain different ends, and it is agreed that each method is correct in the right place. The regression estimates minimize the squares of the discrepancies between the estimates and the true values, summed over the population of persons. Bartlett's estimates minimize the squares of a man's specific factors, summed over tests." That is to say, the Thomson-Thurstone estimates are MAPs or EAPs, which we know minimize squared error

across the population, while Bartlett's estimates minimize the squared residuals within a person. In that same note to *Nature*, Thomson (1938, p. 246) provided the equations for the linear relationship between his estimates and Bartlett's.

3.5 *Observations Missing at Random*

There are a number of ways to compute factor score estimates in the general multiple-factor model with some observations potentially missing at random. One method that corresponds with common practice in IRT is to compute the modal estimates (either ML or MAP) of Sect. 3.3, omitting values corresponding to missing data from the derivatives; general software for full information maximum likelihood (FIML) estimation of factor models may be able to do this as a special case. Another more computationally intensive method, inspired by EAP estimation in IRT, would be to use numerical integration to compute the mean of the likelihoods in equations 14 or 16, similarly omitting unobserved vector elements.

However, a method that is probably less computationally intensive than either of the above strategies was mentioned by Loncke et al. (2018, p. 9): It is reasonably easy to recompute either the Bartlett or Thomson-Thurstone regression coefficients (in equation 28 or 33) for any observation with missing data, omitting matrix elements corresponding to missing observations. With modern computational equipment and software, recomputing the regression coefficients even for every observation is less computation than is routinely done for IRT scoring. And it doesn't have to be redone for each observation, only for those with missing data. For larger numbers of factors or observed variables, it may be more efficient to solve the equations in the form $A\mathbf{f} = \mathbf{b}$ rather than collecting all terms on the right-hand side; on modern computing systems, linear solvers scale better than inversion.

3.6 *The Mean and Variance of the One-Factor Score for One Variable, Redux*

In Sect. 2.2.3, we discussed a one-factor model and showed that the means and variances of the likelihoods for each response are $\mu_{f|y_{ij}} = y_{ij}/\lambda_j$ and $\sigma_{f|y_{ij}}^2 = \theta_j/\lambda_j^2$. Using Bartlett's regression to obtain the expected value and its variance, we arrive at the same result. If there are only one observation y and one factor f , everything in equation 28 is scalar. After cancellation we have $\mu_{f|y_{ij}} = y_{ij}/\lambda_j$; $\sigma_{f|y_{ij}}^2 = \theta_j/\lambda_j^2$ is the scalar form of equation 29.

The weighted-means method of Sect. 2.2.3 works for multiple factor models in which each observed variable is related to only one factor (no cross loadings). A generalization of the weighted means algorithm is possible for models with

cross loadings, modeled after the system described by Thissen et al. (2001) for approximately combining IRT trace lines.

4 Conclusion

We have elaborated on the parallelism between the IRT-based test scoring and the factor score estimates for linear-normal CFA models. Because “Bartlett’s” and “regression” (Thomson-Thurstone) factor score estimates correspond exactly with ML and MAP/EAP estimation in IRT, we have concentrated on those methods. This omits consideration of alternative methods for factor score estimation like those proposed by Skrondal and Laake (2001) and Hoshino and Bentler (2013), but some of the reasoning described in this presentation can be applied to those as well.

Hopefully this integration of IRT and the factor analytic traditions serves to make both easier to understand.

Acknowledgments We thank Li Cai and Alexis Georgeson for bringing the material by Mardia et al. (1979) and Bartholomew et al. (2011) to our attention.

References

- Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. West Sussex: Wiley.
- Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology*, 28, 97–104.
- Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Devlieger, I., Mayer, A., & Rosseel, Y. (2015). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, 76, 741–770.
- Estabrook, R. & Neale, M. (2013). A comparison of factor score estimation methods in the presence of missing data: Reliability and an application to nicotine dependence. *Multivariate Behavioral Research*, 48, 1–27.
- Grice, J. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6, 430–450.
- Hoshino, T. & Bentler, P. M. (2013). Bias in factor score regression and a simple solution. In de Leon, A. R., Chough, & K. C. (Eds.), *Analysis of mixed data: Methods and applications* (pp. 43-61). Boca Raton, FL: Chapman and Hall.
- Loncke, J., Eichelsheim, V., Branje, S., Buysse, A., Meeus, W., & Loeys, T. (2018). Factor score regression with social relations model components: A case study exploring antecedents and consequences of perceived support in families. *Frontiers in Psychology*, 9(1699), 1–19.
- Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate analysis*. London: Academic.
- Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, 66, 563–576.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman and Hall–CRC.
- Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring*. Mahwah: Lawrence Erlbaum Associates.

- Thissen, D., Nelson, L., & Swygert, K. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—approximation methods for scale scores. In D. Thissen & H. Wainer (Eds.), *Test scoring*. Mahwah: Lawrence Erlbaum Associates.
- Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring*. Mahwah: Lawrence Erlbaum Associates.
- Thomson, G. H. (1935). The definition and measurement of “g” (general intelligence). *The Journal of Educational Psychology*, 26, 241–262.
- Thomson, G. H. (1936). Some points of mathematical technique in the factorial analysis of ability. *Journal of Educational Psychology*, 27, 36–54.
- Thomson, G. H. (1938). Methods of estimating factor scores. *Nature*, 141, 246.
- Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University of Chicago Press.

On the Precision Matrix in Semi-High-Dimensional Settings



Kentaro Hayashi, Ke-Hai Yuan, and Ge Jiang

Abstract Many aspects of multivariate analysis involve obtaining the precision matrix, i.e., the inverse of the covariance matrix. When the dimension is larger than the sample size, the sample covariance matrix is no longer positive definite, and the inverse does not exist. Under the sparsity assumption on the elements of the precision matrix, the problem can be solved by fitting a Gaussian graphical model with lasso penalty. However, in high-dimensional settings in behavioral sciences, the sparsity assumption does not necessarily hold. The dimensions are often greater than the sample sizes, while they are likely to be comparable in size. Under such circumstances, introducing some covariance structures might solve the issue of estimating the precision matrix. Factor analysis is employed for modeling the covariance structure and the Woodbury identity to find the precision matrix. Different methods are compared such as unweighted least squares and factor analysis with equal unique variances (i.e., the probabilistic principal component analysis), as well as ridge factor analysis with small ridge parameters. Results indicate that they all give relatively small mean squared errors even when the dimensions are larger than the sample size.

Keywords Factor analysis · Graphical lasso · Inverse covariance matrix · Probabilistic principal component analysis · Woodbury identity

K. Hayashi (✉)

Department of Psychology, University of Hawaii at Manoa, Honolulu, HI, USA
e-mail: hayashik@hawaii.edu

K.-H. Yuan

Department of Psychology, University of Notre Dame, Notre Dame, IN, USA
Department of Statistics, Nanjing University of Posts and Telecommunications, Nanjing, China
e-mail: kyuan@nd.edu

Ge Jiang

Department of Educational Psychology, University of Illinois at Urbana-Champaign, Campaign, IL, USA
e-mail: gejiang2@illinois.edu

1 Introduction

In a variety of applications in multivariate analysis, the inverse of the covariance matrix (i.e., the precision matrix) is required. The necessity of computing the inverse of covariance matrix may be even stronger than that for the covariance matrix itself (Pourahmadi 2013). For example, the precision matrix is involved in the quadratic form of the log-likelihood function of the multivariate normal distribution. In the classical multivariate analysis, we assume that the covariance matrix is positive definite so that all the eigenvalues are positive. In such a case, the inverse always exists.

However, under high-dimensional settings, the situations are different. It is well known that when the number of variables exceeds the sample size, the sample covariance matrix based on the observed data is singular (i.e., some eigenvalues are zero), and the inverse of the covariance matrix no longer exists.

Under high dimensionality, a common approach is to assume sparsity in the covariance matrix or the precision matrix (i.e., assuming that many off-diagonal elements are either zero or near zero) and apply some regularized methods. Well-known regularized methods for estimating the covariance and/or the precision matrices include the thresholding (e.g., Bickel and Lavina 2008) and the graphical lasso (Friedman et al. 2008). See, e.g., Pourahmadi (2013) and Engel et al. (2017), for an overview of estimation of high-dimensional covariance and precision matrices.

However, in most applications in the behavioral sciences, the covariance matrix and the precision matrix are not necessarily sparse. Consequently, the assumption of the sparse covariance matrix and/or the sparse precision matrix may not hold. Also, in the behavioral sciences, even when the number of variables p is greater than the sample size n , they are still comparable in sizes. We rarely encounter situations with the dimension far exceeding the sample size (i.e., “ $p \gg n$ ”). We call such situations (i.e., (i) the covariance matrix or the precision matrix is not sparse, and (ii) p and n are comparable in size though p can be larger than n) “semi-high”-dimensional settings to distinguish them from the high-dimensional settings encountered in, e.g., statistical learning, in recent years.

If we cannot assume sparsity of the covariance matrix or the precision matrix, the most promising existing approach that can be used to find the precision matrix under semi-high-dimensional settings seems to be the ridge-type estimation (e.g., Yuan and Chan 2008, 2016). Yuan and Chan (2008) employed a ridge maximum likelihood (ML) approach to estimate the parameters of structural equation models (SEM). SEM is a structured version of factor analysis (FA; see, e.g., Lawley and Maxwell 1971) and is probably the most frequently used method to model covariance structures in the behavioral sciences. Also employing the FA model, Hayashi et al. (2019) showed that under high dimensions, the precision matrix can be approximated by the inverse of the unique (i.e., error) variance matrix. Because the ML and the generalized least squares (GLS) methods cannot be used without regularization, Hayashi et al. (2019) suggested estimating the parameters of FA by

the unweighted least squares (ULS) method or performing the FA with equal unique variances (e.g., Hayashi and Bentler 2000), also called the probabilistic principal component analysis (the probabilistic PCA; Tipping and Bishop 1999).

Thus, there exist different methods to construct the precision matrix when $p > n$ under semi-high-dimensional settings without assuming sparsity of the covariance matrix. Then, a natural question arises as to which method would perform better among the ULS, the FA with equal variances, and the ridge ML. The purpose of this article is to address this question. Because answering this question seems to be beyond the reach of mathematical analysis, we do so through a simulation study.

2 Factor Analysis and Its Estimation Methods

1. Factor analysis

In the FA model, the p -dimensional mean-centered vector of the observed variables y_i , $i = 1, \dots, n$, is linearly related to an m -dimensional vector of latent factors f_i via $y_i = \Lambda f_i + \epsilon_i$, where Λ is a $p \times m$ matrix of factor loadings (with $p > m$) and ϵ_i is a p -dimensional vector of errors. For the orthogonal factor model with uncorrelated factors, the three assumptions are typically imposed: (i) $f_i \sim N_m(\mathbf{0}, I_m)$; (ii) $\epsilon_i \sim N_p(\mathbf{0}, \Psi)$, where Ψ is a diagonal matrix with positive elements on the diagonals; and (iii) $Cov(f_i, \epsilon_i) = \mathbf{0}$. In words, factors and errors are normally distributed, errors corresponding to different observed variables are uncorrelated, and there are no correlations between factors and errors. Under these assumptions, the covariance matrix of y_i is given by $\Sigma = \Lambda \Lambda' + \Psi$. If y_i is standardized, Σ is a correlation matrix.

2. Woodbury identity

Even if Σ is positive definite so that Σ^{-1} exists in the population, the inverse S^{-1} of the sample covariance matrix S does not exist when $p > n$. When S^{-1} does not exist, we cannot estimate the unique variances Ψ (neither the inverse Ψ^{-1}) under the FA model using the generalized least squares (GLS) or the maximum likelihood (ML) method, which minimizes the fit function $F_{GLS}(S, \Sigma) = tr\{[(S - \Sigma)S^{-1}]^2\}$ and $F_{ML}(S, \Sigma) = tr(\Sigma^{-1}S) - \log |\Sigma^{-1}S| - p$, respectively, without resorting to certain regularization methods.

In this article, the key idea in computing the precision matrix when $p > n$ is to utilize the following Woodbury identity (see, e.g., Chapter 16 of Harville, 1997):

$$\Sigma^{-1} = \Psi^{-1} - \Psi^{-1} \Lambda \left(I_m + \Lambda' \Psi^{-1} \Lambda \right)^{-1} \Lambda' \Psi^{-1}. \quad (1)$$

The identity shows that if the covariance matrix Σ can be expressed as the covariance structure of the FA model $\Sigma = \Lambda \Lambda' + \Psi$, then the precision matrix can also be defined by the FA parameters. It implies that the estimated precision matrix can be estimated as long as we can estimate the FA parameters. Once the

estimated factor loadings $\hat{\Lambda}$ and the estimated unique variances $\hat{\Psi}$ are obtained, we can compute the estimate of the right-hand side (RHS) of the Woodbury identity, as long as all the elements of $\hat{\Psi}$ are strictly positive. $\hat{\Lambda}$ does not need to be of full column rank to compute the RHS of (1). But the parameterization is not proper if it is not of full rank, or a model with fewer factors is better.

Using the argument given by Bentler (1976), Hayashi et al. (2019) pointed out that if the sum of the squared loadings on each factor goes to infinity as the number of observed variables p increases (i.e., $\lambda_k' \lambda_k \rightarrow \infty$, $k = 1, \dots, m$, as $p \rightarrow \infty$, where λ_k is the k -th column of the factor loading matrix), the second term in the RHS of the Woodbury identity vanishes. They proposed to approximate the precision matrix Σ^{-1} by the first term in the RHS of the Woodbury identity, that is, the inverse of unique variances Ψ^{-1} . When Σ has a compound symmetry structure, the order of the second term in the RHS of the Woodbury identity is $O(-\rho(1 - \rho)^{-1}\{(1 - \rho) + \rho \cdot p\}^{-1}) = O(1/p)$, where ρ ($0 < \rho < 1$) is the common correlation in the population. More generally, the approximation depends on the speed with which the smallest eigenvalue of $\Lambda' \Psi^{-1} \Lambda$ goes to infinity (i.e., $\Psi^{-1} - \Sigma^{-1} \rightarrow \mathbf{0}$ as $ev_m(\Lambda' \Psi^{-1} \Lambda) \rightarrow \infty$).

However, we should note that the Woodbury formula itself holds regardless of the sizes of eigenvalues of $\Lambda' \Psi^{-1} \Lambda$. The quadratic form $\Lambda' \Psi^{-1} \Lambda$ is at least positive semi-definite, and $I_m + \Lambda' \Psi^{-1} \Lambda$ is positive definite. Thus, the inverse of $I_m + \Lambda' \Psi^{-1} \Lambda$ always exists. Therefore, we do not actually have to resort to the approximation to computing the inverse of Σ , and instead, we can use the entire RHS of the Woodbury formula in constructing the precision matrix.

3. Unweighted least squares (ULS) and factor analysis with equal unique variances

When S^{-1} does not exist, we cannot estimate the unique variances Ψ nor the inverse Ψ^{-1} under the FA model with either the GLS or the ML method. Thus, Hayashi et al. (2019) suggested either employing the ULS method or performing FA with equal unique variances, which do not require either S^{-1} nor the estimated model-reproduced precision matrix $\hat{\Sigma}^{-1}$. The ULS method minimizes the fit function $F_{ULS}(S, \Sigma) = tr\{(S - \Sigma)^2\}$ which does not involve either S^{-1} or (the estimate of) Σ^{-1} . Thus, estimation with ULS is simpler than that with the GLS or the ML method. The FA model with equal unique variances (with standardized variables) approximates the correlation matrix as:

$$\Sigma \approx \Lambda^* \Lambda^{*'} + kI_p, \quad (2)$$

with a positive constant k , whose ML estimate is given by the average of the $p-m$ smallest eigenvalues of S (i.e., $\hat{k} = \{1/(p-m)\} \sum_{j=m+1}^p ev_j(S)$, where $ev_j(S)$ is the j -th largest eigenvalue of S). Here, note that the eigenvectors of $\Sigma - kI_p$ are the same as the eigenvectors of Σ , and the eigenvalues of $\Sigma - kI_p$ are simply those of Σ minus k . Thus, the FA model with equal unique variances can be considered as a variant of PCA. In fact, this model is also called the probabilistic PCA in statistics (Tipping and Bishop 1999). Also, it has been known that as p increases

(with $m/p \rightarrow \infty$), the loading matrices from the FA and the PCA have the same limiting values, up to a rotational indeterminacy (see, e.g., Guttman 1956; Krijnen 2006; Schneeweiss 1997). Thus, the FA model with equal unique variances is a viable approach to computing the precision matrix under the semi-high-dimensional settings with a large p .

4. Ridge maximum likelihood method

Alternatively, of course, we can employ other estimation methods with a regularization. For example, we can use either ridge GLS or the ridge ML (Yuan and Chan 2008, 2016), neither of which require sparsity of the covariance or precision matrix. In this article, we employ the ridge ML method for SEM proposed by Yuan and Chan (2008). They added a ridge constant $k > 0$ to the diagonals of \mathbf{S} and used $\mathbf{S} + k\mathbf{I}_p$ in place of \mathbf{S} in order to stably estimate the parameters in SEM, a structured FA-type model. Note that now, $\mathbf{S} + k\mathbf{I}_p$ is positive definite and all the eigenvalues of $\mathbf{S} + k\mathbf{I}_p$ are strictly positive, so that the inverse $(\mathbf{S} + k\mathbf{I}_p)^{-1}$ always exists. There is no need for any modifications in applying ridge ML to estimate the FA parameters. Furthermore, Yuan and Chan (2008) proved that adding a ridge term to the sample covariance matrix does not change the ML estimates $\hat{\mathbf{\Lambda}}$ of factor loadings $\mathbf{\Lambda}$ and only changes the ML estimates $\hat{\mathbf{\Psi}}$ of unique variances $\mathbf{\Psi}$ by the ridge constant k .

Unfortunately, the performance of the ridge ML method may depend on the choice of the ridge constant. In Yuan and Chan (2008), the ratio p/n was chosen as the value of the ridge constant. However, in the scenario of p exceeding n , use of p/n as the ridge constant may be too large, in view of the original purpose of the ridge method which is to add a small constant to stabilize the solution in a regression (as in $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$). Therefore, we choose some small values as ridge constants. In this article, no efforts are made to optimize the value of a ridge constant with, e.g., cross-validation, to simplify the process of our simulation.

3 Graphical Lasso

In the simulation study, we also include graphical lasso (Friedman et al. 2008; see also Mazumder and Hastie 2012). Graphical lasso is a popular method used to graphically represent the relations among observed variables, and it obtains a precision matrix as a by-product. Besides sparsity of the precision matrix, it does not result in a structured covariance matrix such as in the FA model. It is concerned with minimizing a l_1 -regularized negative log-likelihood of the form

$$f(\boldsymbol{\Theta}) = -\log\{\|\boldsymbol{\Theta}\|\} + \text{tr}(\mathbf{S}\boldsymbol{\Theta}) + k\|\boldsymbol{\Theta}\|_1, \quad (3)$$

where $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix, $\|\boldsymbol{\Theta}\|_1$ is the sum of the absolute values of $\boldsymbol{\Theta}$, and k is a tuning parameter, which leads to the estimating equation

$$-\Theta^{-1} + S + k\Gamma = \mathbf{0}, \quad (4)$$

where the (j, k) element of Γ is $\gamma_{jk} = \text{sign}(\theta_{jk})$ if $\theta_{jk} \neq 0$; $\gamma_{jk} \in [-1, 1]$ if $\theta_{jk} = 0$.

Now, partition Σ , Σ^{-1} , S , and Γ as

$$\begin{aligned} \Sigma &= \begin{pmatrix} \Sigma_{11} & \sigma_{12} \\ \sigma_{12}' & \sigma_{22} \end{pmatrix}, \quad \Sigma^{-1} = \Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}' & \theta_{22} \end{pmatrix}, \\ S &= \begin{pmatrix} S_{11} & s_{12} \\ s_{12}' & s_{22} \end{pmatrix} \quad \text{and} \quad \Gamma = \begin{pmatrix} \Gamma_{11} & \gamma_{12} \\ \gamma_{12}' & \gamma_{22} \end{pmatrix}, \end{aligned} \quad (5)$$

where the dimensions of Σ_{11} , Θ_{11} , S_{11} , and Γ_{11} are $(p-1) \times (p-1)$; σ_{12} , θ_{12} , s_{12} , and γ_{12} are $(p-1) \times 1$; and σ_{22} , θ_{22} , s_{22} , and γ_{22} are scalars, respectively. Using the relation $\Sigma\Theta = I_p$ for the partitioned matrices, the estimating equation leads to

$$\Sigma_{11}\beta + s_{12} + k\gamma_{12} = \mathbf{0}, \quad (6)$$

where $\beta = \theta_{12}/\theta_{22}$. Because $\theta_{22} > 0$, it is the stationarity equation for $\min_{\beta \in R^{p-1}} \{(1/2)\beta' \Sigma_{11}\beta + \beta' s_{12} + k\|\beta\|_1\}$, where $\Sigma_{11} (> 0)$ is assumed to be fixed. We can regard this as a lasso regression of the last variable on the rest, replacing S_{11} by its current estimate for Σ_{11} . Note that the algorithm also uses the formula for the inverse of a partitioned matrix (see, e.g., Chapter 16 of Harville, 1997):

$$\begin{pmatrix} \Sigma_{11} & \sigma_{12} \\ \sigma_{12}' & \sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} (\Theta_{11} - \theta_{12}\theta_{12}'/\theta_{22})^{-1} & -\Sigma_{11}\theta_{12}/\theta_{22} \\ -\theta_{12}'\Sigma_{11}/\theta_{22} & 1/\theta_{22} - \theta_{12}'\Sigma_{11}\theta_{12}/\theta_{22}^2 \end{pmatrix}. \quad (7)$$

Thus, the algorithm of graphical lasso (Friedman et al. 2008, 2018) solves for a row or column of the estimating equation at a time, holding the rest fixed. Concretely, the algorithm is given as follows:

1. Initialize $\Sigma = S + kI_p$ as in the ridge ML method.
2. Cycle around the columns repeatedly, performing the following steps till convergence:
 - (i) Rearrange the rows and columns such that the target is the p -th column.
 - (ii) Solve the l_1 -regularized quadratic problem: $\hat{\beta} = \arg \min_{\beta \in R^{p-1}} \{(1/2)\beta' \Sigma_{11}\beta + \beta' s_{12} + k\|\beta\|_1\}$.
 - (iii) Update the row and column of the correlation matrix using $\sigma_{12} = -\Sigma_{11}\theta_{12}/\theta_{22} = -\Sigma_{11}\beta$.
 - (iv) Save $\hat{\beta}$ for this column in the matrix B .
3. Finally, for every row and column, compute the diagonal entries $\hat{\theta}_{jj}^{-1}$ using $\hat{\theta}_{22}^{-1} = \hat{\sigma}_{22} = -\hat{\beta}'\hat{\sigma}_{12}$, and convert the B matrix to Θ .

4 Simulation

To empirically examine the performance of FA with equal unique variances, ULS, ridge ML, and graphical lasso, we conducted a simulation study. We employed three structures of loading matrices which were created by vertically concatenating the loading structure with 12 observed variables (see Table 1). (For example, the loading matrix for the condition $p = 240$ was obtained by vertically concatenating the loading matrix with 12 variables $240/12 = 20$ times.) (i) The first structure is a simple structure with no cross loadings. Note that under the first structure, population unique variances are all equal. Thus, it is expected to be an ideal condition for the FA with equal unique variances. (ii) The second structure is with cross loadings. Under the second loading structure, population unique variances are no longer equal. (iii) The third structure is the same as the second one, except that one loading in each loading structure with 12 variables is very high (0.95), which leads to a small unique variance of 0.0475 in the population.

Throughout the simulation, the sample size was set equal to $n = 200$, which we consider to be a sample size comparable to those chosen in research in the behavioral sciences. The number of observed variables were chosen to be $p = 240, 360, 480, \text{ and } 600$, all greater than the sample size of $n = 200$ but not far greater than n , which are consistent with what we call the semi-high-dimensional settings. Because of $p > n$, it is not possible to directly estimate the precision matrix from the sample covariance (or correlation) matrix by inverting the latter. For each simulation condition, we replicated 300 times.

For the estimation methods, we employed (i) the FA with equal unique variances, (ii) ULS, and (iii) ridge ML, as well as (iv) graphical lasso as a reference. For the

Table 1 Three population unit structures of loading matrices

Equal unique variances			Unequal unique variances			Small unique variance		
No cross loadings			With cross loadings			With cross loadings		
0.8	0	0	0.8	0.2	0.1	0.8	0.2	0.1
0.8	0	0	0.8	0.2	0.1	0.8	0.2	0.1
0.8	0	0	0.7	0.1	0.2	0.7	0.1	0.2
0.8	0	0	0.7	0.1	0.2	0.7	0.1	0.2
0	0.8	0	0.2	0.8	0.1	0.2	0.8	0.1
0	0.8	0	0.2	0.8	0.1	0.2	0.8	0.1
0	0.8	0	0.1	0.7	0.2	0.1	0.7	0.2
0	0.8	0	0.1	0.7	0.2	0.1	0.7	0.2
0	0	0.8	0.2	0.1	0.8	0.2	0.1	0.8
0	0	0.8	0.2	0.1	0.8	0.2	0.1	0.8
0	0	0.8	0.1	0.2	0.7	0.1	0.2	0.7
0	0	0.8	0.1	0.2	0.7	0.1	0.2	0.95

Note: Each unit loading matrix was concatenated vertically 20, 30, 40, and 50 times for $p = 240, 360, 480, \text{ and } 600$, respectively.

ridge ML, two small ridge constants (0.05 and 0.10) and the ratio of p/n used in Yuan and Chan (2008) were selected. Because the initial step in graphical lasso is the same as ridge ML, the same tuning parameters of 0.05 and 0.10 (but not p/n) were selected for graphical lasso. For the ULS, we assumed a convergence when the maximum change in unique variances became less than 10^{-7} within 100 iterations. For ridge ML, the convergence was checked via the `factanal.fit.mle` function inside the `factanal` package in R. The `factanal.fit.mle` function uses the generic `optim` function for optimization, and it passes the convergence information to the `factanal.fit.mle` function. Inside the `optim` function, the “L-BFGS-B” method (a modified quasi-Newton method by Byrd et al. 1995) is used. The maximum number of iterations is 100. The default convergence criteria for the `optim` function were used, in which the algorithm stops if it is unable to reduce the estimates of unique variances by a factor of 10^{-8} . For the graphical lasso, the `glasso` package in R (Friedman et al. 2018) was used for estimation. We set the convergence criteria such that the average absolute parameter change becomes less than 10^{-7} times the average of the absolute values of the off-diagonal elements of the sample correlation matrix \mathbf{R} within the maximum of 10,000 iterations. In all the cases, a convergence was reached. For the FA with equal unique variances, because it does not require numerical iterations, the issue of convergence was not a problem. The final estimates of the unique variances for the FA with equal unique variances were computed as $\Psi = \text{diag}(\mathbf{R} - \mathbf{\Lambda}\mathbf{\Lambda}')$ with $\mathbf{\Lambda} = \mathbf{\Omega}\mathbf{\Theta}^{1/2}$, where $\mathbf{\Theta}$ is the diagonal matrix whose diagonal elements are the first m largest eigenvalues of $\mathbf{R} - \hat{k}\mathbf{I}_p$, $\mathbf{\Omega}$ is the $(p \times m)$ matrix whose columns are the corresponding standardized eigenvectors, and \hat{k} was estimated as $\hat{k} = \{1/(p-m)\} \sum_{j=m+1}^p ev_j(\mathbf{R})$.

It is known that minimizing the principal factor method gives the same solution as the ULS method. So in the simulation, we used the following principal factor method: we computed the FA loading matrix as $\mathbf{\Lambda} = \mathbf{\Omega}\mathbf{\Theta}^{1/2}$, where $\mathbf{\Theta}$ is the diagonal matrix whose diagonal elements are the first m largest eigenvalues of $\mathbf{R} - \mathbf{\Psi}$ and $\mathbf{\Omega}$ is the $(p \times m)$ matrix whose columns are the corresponding standardized eigenvectors. (For the initial values for $\mathbf{\Lambda}$, we used the eigenvalues and eigenvectors of \mathbf{R} , not $\mathbf{R} - \mathbf{\Psi}$.) Then, for a given $\mathbf{\Lambda}$, we updated $\mathbf{\Psi}$ by $\mathbf{\Psi} = \text{diag}(\mathbf{R} - \mathbf{\Lambda}\mathbf{\Lambda}')$.

Note that two common initial values for the j -th communality are $1 - 1/r^{jj}$ and $1 - (2m/p)(1/r^{jj})$, where r^{jj} is the j -th diagonal element of the inverse of the sample correlation matrix but it cannot be used with our conditions because \mathbf{R}^{-1} does not exist. For the initial values for ridge ML, we chose $1 - (2m/p)(1/\tilde{r}^{jj})$, where \tilde{r}^{jj} is the j -th diagonal element of $(\mathbf{R} + k\mathbf{I}_p)^{-1}$ with ridge constant k .

5 Results

The differences between empirical and population values for the precision matrix and also the unique variances are summarized in terms of the mean squared errors (MSEs), which are shown in Table 2. Given $p > n$ in our study, the empirical values

are computed from the RHS of the Woodbury identity with the estimates of the FA models in equation (1). We list our findings in the following:

1. MSEs for the precision matrix:

- (i) *Poor performance of graphical lasso*: The most notable finding across different estimation methods and across different loading structures is that the magnitudes of MSEs for graphical lasso are substantially larger than those for the other estimation methods. It means that graphical lasso with the regularization parameters 0.05 and 0.10 performs the worst among the methods that we examine: FA with equal unique variances, ULS, ridge ML, and graphical lasso.
- (ii) *Poor performance of ridge ML with the ridge constant of p/n* : Another notable finding is that the MSEs for ridge ML when p/n are used for the ridge constant are higher than ridge ML when small values (0.05 and 0.10) are used for the ridge constant. It means that the small values of the ridge constant work better than p/n .
- (iii) *Good performance of other estimation methods*: MSEs are almost the same among different estimation methods across FA with equal unique variances, ULS, and ridge ML with small ridge constants (0.05 and 0.10), regardless of different values of p .
- (iv) *Poor performance with small unique variances*: For the third loading structure in Table 1 with small unique variances with cross loadings, the MSEs are substantially higher than those with the first two loading structures across the different estimation methods. It means that the existence of small unique variances, equivalently, the existence of large factor loadings, causes trouble for all the estimation methods considered, even with graphical lasso that does not estimate unique variances or factor loadings.
- (v) *Off-diagonals of precision matrix*: MSEs for the off-diagonals of the Woodbury identity are much smaller than those for the diagonals. Also, as the number of variables increases, the MSEs become smaller in the off-diagonals of the precision matrix.

2. MSEs for unique variances:

Regardless of different loading structures, the values of MSEs for unique variances are similar except for ridge ML with a ridge constant of p/n when $p = 360$ or larger. (Note that graphical lasso does not require estimation of FA parameters.)

3. Further examination of poor performance of graphical lasso:

To further examine why the performance of graphical lasso is poor, we compute proportions (as percentages) of the off-diagonal elements of the 12×12 upper left block of the precision matrix estimated as zero by graphical lasso for the case with $p = 240$ (see Table 3). In general, a large proportion of the off-diagonal elements of the precision matrix are estimated as zero when the corresponding elements of the population precision matrix are small. Also, even when the population values are not very small, some off-diagonal elements of the precision matrix are sometimes

Table 2 Mean squared errors of diagonal, off-diagonal elements of Woodbury formula, and unique variances

Loading structure	Estimation method	N	P	MSE diagonal	MSE off-diagonal	MSE Psi
1	Equal Psi	200	240	0.102305	1.81E-05	0.001613
1	ULS	200	240	0.106878	1.88E-05	0.001677
1	Ridge (0.05)	200	240	0.105336	1.88E-05	0.001643
1	Ridge (0.1)	200	240	0.111321	1.92E-05	0.001717
1	Ridge (p/n)	200	240	0.123496	2.02E-05	0.001743
1	glasso (0.05)	200	240	0.174678	0.001206	NA
1	glasso (0.1)	200	240	0.743764	0.000604	NA
1	Equal Psi	200	360	0.106081	8.23E-06	0.001645
1	ULS	200	360	0.108061	8.43E-06	0.001674
1	Ridge (0.05)	200	360	0.108407	8.42E-06	0.001649
1	Ridge (0.1)	200	360	0.108304	8.40E-06	0.001692
1	Ridge (p/n)	200	360	0.134967	9.43E-06	0.001791
1	glasso (0.05)	200	360	0.157987	0.000861	NA
1	glasso (0.1)	200	360	0.730270	0.000403	NA
1	Equal Psi	200	480	0.106565	4.67E-06	0.001632
1	ULS	200	480	0.109422	4.77E-06	0.001697
1	Ridge (0.05)	200	480	0.106612	4.70E-06	0.001669
1	Ridge (0.1)	200	480	0.106484	4.71E-06	0.001647
1	Ridge (p/n)	200	480	0.146488	5.53E-06	0.001900
1	glasso (0.05)	200	480	0.134898	0.000685	NA
1	glasso (0.1)	200	480	0.702913	0.000302	NA
1	Equal Psi	200	600	0.107517	2.99E-06	0.001618
1	ULS	200	600	0.111272	3.06E-06	0.001712
1	Ridge (0.05)	200	600	0.111583	3.07E-06	0.001652
1	Ridge (0.1)	200	600	0.111945	3.07E-06	0.001684
1	Ridge (p/n)	200	600	0.150304	3.59E-06	0.001900
1	glasso (0.05)	200	600	0.120045	0.000578	NA
1	glasso (0.1)	200	600	0.685683	0.000241	NA
2	Equal Psi	200	240	0.100350	2.09E-05	0.001806
2	ULS	200	240	0.104407	2.18E-05	0.001841
2	Ridge (0.05)	200	240	0.105240	2.20E-05	0.001858
2	Ridge (0.1)	200	240	0.108618	2.23E-05	0.001889
2	Ridge (p/n)	200	240	0.115976	2.29E-05	0.001889
2	glasso (0.05)	200	240	0.196374	0.001228	NA
2	glasso (0.1)	200	240	0.793978	0.000597	NA
2	Equal Psi	200	48	0.105526	9.57E-06	0.001820
2	ULS	200	360	0.107620	9.86E-06	0.001827
2	Ridge (0.05)	200	360	0.105916	9.82E-06	0.001827
2	Ridge (0.1)	200	360	0.107118	9.83E-06	0.001826
2	Ridge (p/n)	200	360	0.134016	1.10E-05	0.001942
2	glasso (0.05)	200	360	0.179615	0.000892	NA

(continued)

Table 2 (continued)

Loading structure	Estimation method	N	P	MSE diagonal	MSE off-diagonal	MSE Psi
2	glasso (0.1)	200	360	0.783000	0.000396	NA
2	Equal Psi	200	480	0.102823	5.37E-06	0.001764
2	ULS	200	480	0.104555	5.47E-06	0.001800
2	Ridge (0.05)	200	480	0.112516	5.61E-06	0.001852
2	Ridge (0.1)	200	480	0.107614	5.54E-06	0.001793
2	Ridge (p/n)	200	480	0.142027	6.41E-06	0.001996
2	glasso (0.05)	200	480	0.155783	0.000722	NA
2	glasso (0.1)	200	480	0.758072	0.000297	NA
2	Equal Psi	200	600	0.103230	3.46E-06	0.001826
2	ULS	200	600	0.107532	3.56E-06	0.001800
2	Ridge (0.05)	200	600	0.111029	3.58E-06	0.001887
2	Ridge (0.1)	200	600	0.106591	3.53E-06	0.001823
2	Ridge (p/n)	200	600	0.153773	4.25E-06	0.002067
2	glasso (0.05)	200	600	0.140010	0.000619	NA
2	glasso (0.1)	200	600	0.740493	0.000237	NA
3	Equal Psi	200	240	1.143269	0.000250	0.001584
3	ULS	200	240	0.822066	0.000193	0.001646
3	Ridge (0.05)	200	240	0.832872	0.000189	0.001649
3	Ridge (0.1)	200	240	0.819792	0.000191	0.001634
3	Ridge (p/n)	200	240	0.910596	0.000205	0.001667
3	glasso (0.05)	200	240	15.51614	0.003567	NA
3	glasso (0.1)	200	240	22.85946	0.003870	NA
3	Equal Psi	200	360	0.861037	8.45E-05	0.001623
3	ULS	200	360	0.841318	8.59E-05	0.001604
3	Ridge (0.05)	200	360	0.854895	8.58E-05	0.001644
3	Ridge (0.1)	200	360	0.864639	8.60E-05	0.001664
3	Ridge (p/n)	200	360	3.466345	0.000317	0.001784
3	glasso (0.05)	200	360	15.94882	0.001954	NA
3	glasso (0.1)	200	360	23.49369	0.001853	NA
3	Equal Psi	200	480	0.797377	4.41E-05	0.001617
3	ULS	200	480	0.903535	5.06E-05	0.001667
3	Ridge (0.05)	200	480	0.885663	4.96E-05	0.001623
3	Ridge (0.1)	200	480	0.829790	4.74E-05	0.001654
3	Ridge (p/n)	200	480	6.303237	0.000324	0.001869
3	glasso (0.05)	200	480	16.05256	0.001324	NA
3	glasso (0.1)	200	480	23.75300	0.001114	NA
3	Equal Psi	200	600	0.767052	2.71E-05	0.001622
3	ULS	200	600	0.888375	3.18E-05	0.001646
3	Ridge (0.05)	200	600	0.914167	3.21E-05	0.001614
3	Ridge (0.1)	200	600	0.862146	3.08E-05	0.001620
3	Ridge (p/n)	200	600	9.301400	0.000305	0.001874
3	glasso (0.05)	200	600	16.16034	0.001007	NA
3	glasso (0.1)	200	600	23.92592	0.000760	NA

estimated as zero by graphical lasso. Table 3 shows that the structure of the precision matrix is most sparse under the first loading structure and least sparse under the third loading structure. Accordingly, the performance of graphical lasso is the best under the first loading structure and the worst under the third loading structure.

6 Discussion

In our simulation, graphical lasso did not perform well compared to other estimation methods. Based on our further analysis, this seems to be due to the fact that the population precision matrices employed in the simulation are still not yet sparse enough. As we point out, Hayashi et al. (2019) showed that as the dimension increases, the precision matrix converges to a diagonal matrix (i.e., to the inverse of the unique variance matrix). This means that the off-diagonal elements of the precision matrix approach zero as the dimension increases. So, as the dimension increases, the precision matrix approaches a sparse matrix and the performance of graphical lasso should become better. Also, another feature of our simulation conditions is that the population correlation matrices all fit the FA models perfectly. Graphical lasso is the only method that did not take advantage of the correlation matrix having the FA models. Other methods are able to utilize the Woodbury formula because the FA parameter estimates are available.

Through the simulation, we were able to show that even when the sample size is smaller than the number of variables, the parameters of the FA model and the corresponding precision matrix can still be estimated relatively accurately using different methods such as the FA with equal unique variances, ULS, and ridge ML with small ridge constants. On average, the performances of these methods were nearly equal. The exceptions were when the ridge ML method was used with the ridge constant being p/n , which implies that the values of p/n are probably too large when p is greater than n . On the other hand, when the values of the ridge constants were small, the ridge ML method worked fine, even though the selected values of the ridge constant were not optimized.

The ridge method has been known to stabilize solutions by adding a small constant. Yuan and Chan (2008) showed that adding the ridge constant does not affect the values of factor loadings and only changes the values of unique variances by the size of the ridge constant. Thus, the ridge ML method still guarantees the consistency of the FA estimates if the ridge constant is subtracted from the estimates of the unique variances at the end. Thus, we expected that the ridge ML method would perform well, and it actually did, as long as small ridge constants were chosen.

It is noteworthy that the FA with equal unique variances performed excellently, even when the population unique variances were not equal, especially for our condition of the third loading structure where the existence of small unique variances made unique variances more unequal in the population. The results are

Table 3 Percentage of zeros in off-diagonal elements of reproduced precision matrix with graphical lasso for the first 12 variables (lower left) and corresponding off-diagonal elements of precision matrix in the population (upper right) (number of variables $p = 240$, tuning parameter $= 0.05$)

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12
Loading structure 1. Equal unique variances with no cross loadings												
1		-.0345				0	0	0	0	0	0	0
2	61.0		-.0345		0	0	0	0	0	0	0	0
3	66.0	61.7		-.0345	0	0	0	0	0	0	0	0
4	64.0	65.0	65.3		0	0	0	0	0	0	0	0
5	94.7	94.0	91.3	93.3		-.0345	-.0345	-.0345	0	0	0	0
6	94.0	92.0	90.3	94.3	64.0		-.0345	-.0345	0	0	0	0
7	93.0	93.7	91.0	92.7	65.0	63.0		-.0345	0	0	0	0
8	92.0	90.3	92.7	95.7	65.0	65.3	65.3		0	0	0	0
9	95.0	92.3	92.7	92.7	94.0	91.0	94.7	91.0		-.0345	-.0345	-.0345
10	91.0	93.7	93.3	94.3	94.3	92.3	93.3	92.7	62.3		-.0345	-.0345
11	95.0	93.7	93.0	93.7	94.7	93.7	92.7	94.3	61.7	59.0		-.0345
12	93.7	92.7	96.7	93.0	94.7	95.3	93.7	91.3	66.0	60.7	66.3	
Loading structure 2. Unequal unique variances with cross loadings												
1		-.0529	-.0303	-.0303	-.0057	-.0057	.0015	.0015	.0004	.0004	.0044	.0044
2	58.3		-.0303	-.0303	-.0057	-.0057	.0015	.0015	.0004	.0004	.0044	.0044
3	68.7	73.3		-.0183	.0012	.0012	.0029	.0029	-.0057	-.0057	-.0004	-.0004
4	69.3	70.0	71.0		.0012	.0012	.0029	.0029	-.0057	-.0057	-.0004	-.0004
5	93.0	91.3	97.0	92.7		-.0525	-.0303	-.0303	.0061	.0061	-.0021	-.0021
6	93.3	93.0	93.3	91.3	57.0		-.0303	-.0303	.0061	.0061	-.0021	-.0021
7	94.3	93.0	89.0	89.0	71.3	66.7		-.0185	-.0019	-.0019	-.0048	-.0048
8	92.7	91.7	85.3	84.0	68.3	70.3	71.0		-.0019	-.0019	-.0048	-.0048
9	95.7	97.7	90.3	90.0	91.3	91.7	93.7	90.0		-.0525	-.0301	-.0301

(continued)

Table 3 (continued)

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12
Loading structure 2. Unequal unique variances with cross loadings												
10	95.7	95.0	89.3	92.3	93.0	92.3	95.0	92.0	57.0		-.0301	-.0301
11	92.7	91.7	91.0	84.3	93.0	91.0	80.0	84.3	72.3	68.7		-.0183
12	90.7	88.3	87.3	85.7	93.0	92.7	82.7	88.3	69.7	68.0	69.7	
Loading structure 3. Small unique variances with cross loadings												
1		-.0517	-.0303	-.0303	-.0056	-.0056	.0007	.0007	-.0062	-.0062	.0004	.0175
2	53.0		-.0303	-.0303	-.0056	-.0056	.0007	.0007	-.0062	-.0062	.0004	.0175
3	70.7	68.3		-.0183	.0012	.0012	.0029	.0029	-.0056	-.0056	-.0004	-.0006
4	64.7	65.7	70.3		.0012	.0012	.0029	.0029	-.0056	-.0056	-.0004	-.0006
5	93.0	91.7	92.3	90.3		-.0525	-.0304	-.0304	.0057	.0057	-.0024	-.0005
6	92.7	91.3	93.0	90.3	56.0		-.0304	-.0304	.0057	.0057	-.0024	-.0005
7	91.7	94.0	86.3	85.3	65.7	69.7		-.0180	.0024	.0024	-.0022	-.0128
8	92.0	92.7	90.7	86.3	66.3	66.3	70.3		.0024	.0024	-.0022	-.0128
9	96.7	95.7	91.3	89.3	93.7	91.7	94.0	89.3		-.0151	-.0076	-.1031
10	98.0	95.3	87.3	88.3	92.3	89.7	89.0	92.0	60.0		-.0076	-.1031
11	89.7	92.7	85.3	88.3	92.7	90.0	87.7	86.0	74.7	74.3		-.0621
12	92.0	93.7	99.3	99.7	100.0	100.0	96.7	98.7	65.0	64.0	73.7	

likely due to the fact that the FA with equal unique variances is a variant of PCA, also called the probabilistic PCA (Tipping and Bishop 1999). It is known that the loading matrices from FA and PCA converge to the same limiting values (up to rotational indeterminacy) as the number of variables increases (Guttman 1956; Krijnen 2006; Schneeweiss 1997). Thus, it is not surprising that the FA with equal variances performed very well in the simulation, where the number of variables were large.

The performance of ULS in terms of MSEs was also excellent. This might be due to the fact that without using a covariance matrix as a weight matrix, ULS is a simple estimation method. Our experience indicates that simple estimation methods work well for a large number of variables. Though the context is different, the results in this article remind us of the strength of ULS in identifying small factors (MacCallum et al. 2007). Also, another example of a “simple” method that works well for a large number of variables is simple algorithms such as the coordinate descent (see, e.g., p. 118 of Hastie et al. 2015).

With the small unique variances in the third loading structure in Table 1, the values of MSEs were much higher than those corresponding to the first two loading structures. Related to this, we found that if the average diagonal element of the Woodbury formula is large, the corresponding MSE was also large. For example, the average values of the (12, 12) element of the Woodbury formula corresponding to a small unique variance in the third loading structure are between 20.55 and 25.27, as estimated by FA with equal unique variances, ULS, and ridge ML, whereas the average values of the same element in the first and the second loading structures are only between 2.18 and 2.95. The results indicate that small unique variances create challenges for estimating the precision matrix, primarily because equation (1) involves repeatedly inverting the unique variance matrix.

When the precision matrix was estimated by graphical lasso, the third loading structure in Table 1 resulted in very large values of MSEs. This can also be explained as follows: the small unique variances are the result of the existence of large factor loadings, which create a correlation structure with higher off-diagonal elements. It further results in larger off-diagonal elements of the precision matrix. Thus, the precision matrix becomes less sparse and graphical lasso is not a good fit for such conditions.

Note that MSEs for the off-diagonal elements of the Woodbury formula were much smaller than those for the diagonal elements. This is due to the fact that the diagonal elements of the Woodbury formula are much larger than the off-diagonal elements. Also, it is expected that the MSEs for the off-diagonal elements of the Woodbury formula became smaller as the number of variables gets larger. Again, this is due to the fact that as the number of variables increases, the precision matrix approaches a diagonal matrix. That is, the off-diagonal elements of the precision matrix become smaller as the dimension increases.

Our simulation design is relatively simple and limited. For example, in all the conditions, the FA model holds in the population. Also, the number of factors in the population was known a priori. Consequently, caution is needed when generalizing

the results to a broad scope, although all the studied methods possess certain robust properties. Additional studies are needed that include more conditions.

Acknowledgments The authors would like to thank Dr. Dylan Molenaar for his careful review of the manuscript and also thank Rachelle Podhorzer for clerical assistance. This work was supported by Grant 31971029 from the National Natural Science Foundation of China.

References

- Bentler, P. M. (1976). Multistructure statistical model applied to factor analysis. *Multivariate Behavioral Research, 11*, 3–15.
- Bickel, P., & Lavina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics, 36*, 2577–2604.
- Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing, 16*, 1190–1208.
- Engel, J., Buydens, L., & Blanchet, L. (2017). An overview of large-dimensional covariance and precision matrix estimator with applications in chemometrics. *Journal of Chemometrics, 31*, article e2880.
- Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*, 432–441.
- Friedman, J., Hastie, T., & Tibshirani, R. (2018). Graphical lasso: Estimation of Gaussian graphical models, Version 1.10. <https://cran.r-project.org/web/packages/glasso/glasso.pdf>
- Guttman, L. (1956). Best possible systematic estimates of communalities. *Psychometrika, 21*, 273–285.
- Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer.
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press.
- Hayashi, K., & Bentler, P. M. (2000). On the relations among regular, equal unique variances and image factor analysis. *Psychometrika, 65*, 59–72.
- Hayashi, K., Yuan, K.-H., & Jiang, G. (2019). On extended Guttman condition in high dimensional factor analysis. In M. Wilberg, S. Culpepper, R. Janssen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative psychology: The 83rd annual meeting of the psychometric Society, New York City, 2018* (pp. 221–228). New York: Springer.
- Krijnen, W. P. (2006). Convergence of estimates of unique variances in factor analysis, based on the inverse sample covariance matrix. *Psychometrika, 71*, 193–199.
- Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). New York: American Elsevier.
- MacCallum, R. C., Browne, M. W., & Cai, L. (2007). Factor analysis models as approximations. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 153–175). Mahwah, NJ: Erlbaum.
- Mazumder, R., & Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics, 6*, 2125–2149.
- Pourahmadi, M. (2013). *High-dimensional covariance estimation*. New York: Wiley.
- Schneeweiss, H. (1997). Factors and principal components in the near spherical case. *Multivariate Behavioral Research, 32*, 375–401.
- Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B, 61*, 611–622.
- Yuan, K.-H., & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics & Data Analysis, 52*, 4842–4858.
- Yuan, K.-H., & Chan, W. (2016). Structural equation modeling with unknown population distributions: Ridge generalized least squares. *Structural Equation Modeling, 23*, 163–179.

Performance of the Modified Continuous a -Stratification Indices in Computerized Adaptive Testing



Ya-Hui Su and Yan-Ling Lai

Abstract Computerized adaptive testing (CAT) has become increasingly popular for many purposes, including educational assessment in schools, personnel recruitment by institutions, and clinical diagnosis in hospitals. Practically, it is critical that items are not overexposed because exposure can translate into sharing with other examinees, which might threaten the validity of test scores. A popular method for monitoring item exposure is a -stratification. When removing old items from an item bank or adding new items to the item bank, the optimal strata for a -stratification would need to be obtained by having additional simulation studies. It is undesirable for practitioners in high-stakes testing when an item bank needs to be updated frequently. The continuous a -stratification index (CAI) not only avoids partitioning the item bank but also monitors item exposure; however, the CAI still yields a high percentage of overexposed items. Therefore, this study proposed three modified CAI methods for monitoring item exposure and compared their performance in item selection with some CAT item exposure control methods.

Keywords a -stratification · Computerized adaptive testing · Item exposure · Item selection

1 Introduction

Since the 1990s, computerized adaptive testing (CAT) has become increasingly popular for educational assessments in schools, personnel recruitment by institutions, and clinical diagnosis in hospitals. Because examinees are administered different sets of items in CAT, CAT can obtain efficient and precise estimations; it can also increase the security of testing materials. Additionally, CAT can provide cognitive diagnostic information or additional instruction to teachers, parents, and students

Y.-H. Su (✉) · Y.-L. Lai
Department of Psychology, National Chung Cheng University, Chiayi, Taiwan
e-mail: psyyhs@ccu.edu.tw

to improve students' learning process. Therefore, CAT may greatly improve the efficiency of assessments in many applications.

Several drawbacks have commonly been discussed regarding CAT. For example, constraints on content and answer keys are difficult to manage; additionally, a few aberrant responses at the early stage of testing might seriously affect the final estimate; besides, only a small portion of the items from the item bank are used during administration, which leads to a security risk. This is because the maximum Fisher information criterion tends to select items with high discrimination parameters in CAT, and these most-informative items may not be needed at the early stage of testing, at which time the ability estimator is not a considerable certainty. It is crucial, for practical reasons, to prevent items from being overexposed because the exposed items might be shared with current and future examinees, which can threaten the validity of test scores. Many item exposure control methods in relation to CAT have been summarized by Georgiadou et al. (2007).

The method of a -stratification is popular for monitoring item exposure (Chang and Ying 1999). Because the a -stratification method is easy to implement, it has been extended to one- and two-dimensional CAT (Chang et al. 2001; Chang and van der Linden 2003; Huebner et al. 2015; Lee et al. 2008; Leung et al. 2002). The a -stratification method can reduce the number of items with high information overexposure; however, measurement precision may be sacrificed to some degree, which many practitioners likely regard as negligible in their testing situations. In practice, when an item bank is constructed, each item is assigned to a stratum according to its discrimination parameter. Then, additional simulation studies are conducted to determine an appropriate stratum value with which to balance item exposure control as well as measurement precision. When removing old items from the item bank or adding new items to the item bank, the optimal strata for a -stratification must be obtained through additional simulation studies. It is undesirable for practitioners when an item bank needs to be updated frequently. It is even more challenging for high-stakes testing because item usage is closely monitored, in which overexposed items are frequently eliminated from the bank, and new items are regularly added to the bank.

Huebner et al. (2018) proposed a continuous a -stratification index (CAI), which incorporates item exposure control into the item selection process. In this situation, the CAI method does not need to partition an item bank into fixed and discrete strata. The researchers compared the CAI method with two existing a -stratification methods (i.e., a -stratification with matching- b [SMB] and a -stratification with the maximum Fisher information criterion [SMI]) through simulations under various test length and examinee's ability distribution conditions. It was found that the SMB method obtained the best item exposure control with the consistently smallest chi-square statistics, which was a measure of the evenness of item exposure rates; however, the SMB method showed the worst ability estimator with the consistently largest mean square error (MSE). It was also found that the CAI method performed similarly or better than the SMI method in terms of bias and MSE and also obtained smaller chi-square statistics. Huebner et al. (2018) recommended both the SMI and CAI methods to practitioners who would like to have a precise ability estimator.

In practice, the maximum item exposure rate is commonly set to 0.2. An item is considered overexposed if its exposure rate is larger than 0.2. According to Huebner et al. (2018), the percentage of overexposed items for the SMB, SMI, and CAI methods was 3.8%, 8.6%, and 6.9% under 20-item conditions, respectively. The percentage of overexposed items for the SMB, SMI, and CAI methods was 5.6%, 12.7%, and 12.3% under 30-item conditions, respectively. Based on the preceding discussion, the percentage of overexposed items was not low enough to be deemed negligible by many practitioners. These three α -stratification methods had a high percentage of items overexposed to examinees, entailing a substantial security risk because examinees might share testing information with current and future examinees. Huebner et al. (2018) suggested that the CAI method should be combined with item exposure control methods to limit item exposure. Thus, this study aimed to modify the CAI method for monitoring item exposure in CAT.

1.1 Continuous α -Stratification Index

Huebner et al. (2018) proposed the CAI method, which integrates item exposure control with the item selection index itself. Thus, the CAI method does not need to partition the item bank into fixed and discrete strata. The CAI method represents the similarity between the current testing stage and the percentile rank of the discrimination parameter for item i . Denote I as the number of items in the item bank, and I and L are the numbers of administered items and test length, respectively. After I items have been administered, the $(I+1)$ th item is selected by maximizing the quantity:

$$CAI_i \times Inf_i, \tag{1}$$

where the term Inf_i is the Fisher information criterion evaluated at the provisional ability estimator of item i . The term CAI_i for item i is defined as

$$CAI_i = \exp \left[-\beta \left(\frac{PR(a_i)}{I/L} - 1 \right)^2 \right], \tag{2}$$

where the term $PR(a_i)$ is the percentile rank of the discrimination parameter a for item i in the item bank that ranges between 0 and 1. Theoretically, the term $\frac{PR(a_i)}{I/L} - 1$ ranges between -1 and ∞ . Because the nature of the CAI method is to find $PR(a_i)$ as close as possible to I/L , the term $\frac{PR(a_i)}{I/L} - 1$ should be as close as possible to 0. The term β is a sensitivity parameter that determines the sensitivity of the discrepancy between $PR(a_i)$ and I/L during item selection. The term β was constrained as greater than 0. As Huebner et al. (2018) suggest, the sensitivity parameter β was set equal to 2 based on preliminary simulations.

As mentioned earlier, Huebner et al. (2018) compared three a -stratification methods (the SMB, SMI, and CAI methods) in their study. The SMB, SMI, and CAI methods were all found to identify a high percentage of overexposed items in CAT. In practice, the maximum item exposure rate is commonly set to 0.2, and an item is considered to be an overexposed item if its exposure rate is larger than 0.2. The CAI method was found to obtain 6.9% and 12.3% overexposed items for 20- and 30-item conditions, respectively. Such a high percentage of overexposed items would result in a security risk because the overexposed items may be shared with current and future examinees, which would hurt the validity of test scores. To learn more about how the CAI behaves, a preliminary replication was conducted here based on the study by Huebner et al. (2018). The CAI method has tended to administer an item where its $PR(a_i)$ was close to $1/L$, which matched the expression of the CAI method in Eq. (2). In addition to Eq. (2), the item selection process still needed to take Eq. (1) into consideration as well.

2 Method

Three modified CAI methods were derived in this study. Simulations were conducted to evaluate the performance of these modified CAI methods and three a -stratification methods (i.e., the SMB, SMI, and CAI methods) in CAT. The data generation, simulation design, and evaluation criteria were conducted under various conditions.

2.1 Modified Continuous a -Stratification Indices

Huebner et al. (2018) suggested that the CAI method combined with item exposure control methods would limit the maximum item exposure rate. To monitor item exposure, three modified CAI methods were proposed in the present study: the CAI + exposure, CAI + freeze, and CAI + SHOF methods.

One popular constraint-weighted item selection method is the maximum priority index (Cheng and Chang 2009), which has been proposed for simultaneously and efficiently monitoring many statistical and non-statistical constraints during item selection in unidimensional and multidimensional CAT (Cheng and Chang 2009; Cheng et al. 2009; Su 2015, 2016; Su and Huang 2015; Yao 2011, 2012, 2013). The constraint for monitoring item exposure control can be implemented as follows:

Assume that constraint k requires the item exposure rates of all items to be lower than or equal to a pre-specified maximum item exposure rate r_{\max} . After S examinees have taken the CAT, s examinees have seen item i . The term f_k can be defined as

$$f_k = \frac{1}{r_{\max}} \left(r_{\max} - \frac{s_i}{S} \right), \quad (3)$$

where $\frac{s_i}{S}$ is the provisional exposure rate of item i . To monitor item exposure, the CAI method can be integrated with the item exposure control in Eq. (3) as the CAI+exposure method, which is defined as

$$\text{CAI}_i \times \text{Inf}_i \times \frac{\left(r_{\max} - \frac{s_i}{S} \right)}{r_{\max}}. \quad (4)$$

When the selection method is the CAI+exposure method, item selection is performed by maximizing the quantity in Eq. (4).

Revuelta and Ponsoda (1998) proposed the restrictive maximum information method with freeze control to monitor item exposure. When the item exposure rate of an item reaches r_{\max} , the item is not included temporarily for item selection, a process also known as freeze control. Otherwise, item selection is performed by maximizing the quantity in Eq. (1). The CAI+freeze method should prevent items from being overexposed.

The Simpson and Hetter online procedure with freeze (SHOF; Chen et al. 2008) was proposed for handling item selection in CAT, and it can sufficiently prevent items from being overexposed well. In the present study, the CAI method is integrated with the SHOF as the third modified CAI method (i.e., the CAI+SHOF) to monitor item exposure. Let S_i and A_i denote the event of selecting item i and the event of administering item i , respectively. At the beginning of testing, all item exposure parameters $P(A|S)$ are set to 1. During an examinee's testing stage, an item with maximum item information is administered if a random number is less than $P(A|S)$; otherwise, the item is not administered, and the next item is selected. This procedure is repeated until testing is completed. After each examinee completes the administered items, the item exposure parameters must be adjusted accordingly:

$$\begin{aligned} &\text{If } P(A) > r_{\max}, \text{ then } P(A|S) = 0.0 \text{ (i.e., freeze control).} \\ &\text{If } P(A) \leq r_{\max}, \text{ and } P(A|S) \leq r_{\max}, \text{ then } P(A|S) = 1.0. \\ &\text{If } P(A) \leq r_{\max}, \text{ and } P(A|S) > r_{\max}, \text{ then } P(A|S) = \frac{r_{\max}}{P(S)}. \end{aligned} \quad (5)$$

The CAI+SHOF method should prevent items from being overexposed.

2.2 Simulation Design

The three-parameter logistic regression model describes the relationship between examinee and item parameters through mathematical models (Birnbaum 1968; Lord 1980; Wainer and Mislevy 2000). Conditional on the latent trait θ_n , the probability of getting a correct response for person n on item i is defined as

$$p_{ni1} = c_i + (1-c_i) \frac{\exp [a_i (\theta_n - b_i)]}{1 + \exp [a_i (\theta_n - b_i)]}, \quad (6)$$

where p_{ni1} is the probability of getting a correct response, and a_i , b_i , and c_i are the discrimination, difficulty, and guessing parameters of item i , respectively. In this study, item responses were generated according to the three-parameter logistic regression model in Eq. (6).

Through simulations, the efficiency of the three modified CAI methods was compared with that of the SMB, SMI, and CAI methods based on measurement precision and item exposure control. Specially, two factors were considered in the study: test length (2 levels; 20 and 30 items) and item selection methods (6 levels; SMB, SMI, CAI, CAI+exposure, CAI+freeze, and CAI+SHOF). The fixed-length stopping rule was used in this study. Following the design of Leung et al. (2002) and Huebner et al. (2018), true examinee abilities ($N = 5100$) were generated from a standard normal distribution. To compare the results with Huebner et al. (2018), the strata M was set to 4 for the SMB and SMI methods, and the sensitivity parameter β was set to 2 for the CAI and three modified CAI methods. For all conditions, 500 items were generated to form an item bank, for which the a , b , and c parameters were drawn from the Uniform(0, 1.3), Uniform(-1.3, 1.3), and Uniform(0.2, 0.3) distributions, respectively. The maximum likelihood estimator (MLE) was used to find examinees' ability to estimate.

2.3 Evaluation Criteria

The results of the simulation study were analyzed and discussed based on the following two aspects: (a) exposure control and (b) measurement precision. The measurement precision was evaluated by each latent ability recovery based on the bias, root mean squared error (RMSE), correlation between the estimated and true abilities ($r_{\theta, \hat{\theta}}$), and relative efficiency (RE). The formulas for bias, RMSE, r , and RE are defined as follows:

$$\text{bias} = \frac{1}{N} \sum_{n=1}^N (\hat{\theta}_n - \theta_n), \quad (7)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^N (\hat{\theta}_n - \theta_n)^2}, \quad (8)$$

$$r_{\theta, \hat{\theta}} = \frac{COV_{\theta, \hat{\theta}}}{S_{\theta} S_{\hat{\theta}}}, \text{ and} \quad (9)$$

$$RE = \frac{RMSE_{SMI}}{RMSE_{others}}, \tag{10}$$

where $\hat{\theta}_n$ and θ_n are the estimated and true abilities, respectively.

With respect to exposure control, the maximum item exposure rate, number of overexposed items (i.e., items with exposure rates larger than $r_{max}=0.2$), and number of unused items were reported. In addition, the χ^2 statistic was used to measure the skewness of item exposure rate distribution (Chang et al. 2001; Chang and van der Linden 2003; Chang and Ying 1999; Leung et al. 2002), which is defined as

$$\chi^2 = \frac{1}{L/I} \sum_{i=1}^I (r_i - L/I)^2, \tag{11}$$

where r_i is the exposure rate of item i and L is the test length. For each item selection method, qualifying the discrepancy between the observed and the expected item exposure rates was a good index of the efficiency of item bank usage. The smaller the χ^2 statistic attained, the better the item exposure control was.

3 Results

To evaluate the efficiency of the six item selection methods, measurement precision and exposure control under various conditions were reported. With respect to measurement precision, the bias, RMSE, RE, and $r_{\theta, \hat{\theta}}$ under various conditions are summarized in Table 1. The CAI, CAI+freeze, and CAI+SHOF methods yielded slightly larger RE than did the SMI method, meaning that the measurement precision

Table 1 Measurement precision of the six item selection methods under various conditions

Item length	Item selection methods	bias	RMSE	RE	$r_{\theta, \hat{\theta}}$
20	SMB	0.068	0.575	0.858	0.864
	SMI	0.024	0.494	1.000	0.901
	CAI	0.018	0.469	1.053	0.912
	CAI + exposure	0.026	0.496	0.996	0.901
	CAI + freeze	0.016	0.461	1.072	0.916
	CAI + SHOF	0.020	0.469	1.054	0.912
30	SMB	0.018	0.397	0.923	0.934
	SMI	-0.001	0.366	1.000	0.944
	CAI	0.007	0.363	1.009	0.943
	CAI + exposure	0.016	0.402	0.911	0.933
	CAI + freeze	0.012	0.361	1.013	0.944
	CAI + SHOF	0.011	0.372	0.983	0.940

was slightly better in the modified CAI methods than in the SMI method. The correlation between the estimated and true abilities for these three methods reached 0.91. The CAI+exposure method yielded slightly smaller RE and correlation than did the SMI method. Among the six methods, the SMB obtained the smallest RE and correlation. However, all item selection methods performed slightly better when the test length increased.

With respect to exposure control, the percentage of items under different item exposure rate, the maximum item exposure rate, and the chi-square statistics under various conditions are summarized in Table 2. The CAI+exposure, CAI+freeze, and CAI+SHOF methods had no items overexposed, and the maximum item exposure rates of these three methods were under 0.2, which means that the three modified CAI methods performed well regarding item exposure. Among the six modified CAI methods, the CAI+exposure method yielded the smallest chi-square statistics, which means that the CAI+exposure method outperformed the other five item selection methods. The SMB, SMI, and CAI methods obtained similar item exposure patterns, which matched the findings of Huebner et al. (2018).

4 Discussion

To monitor item exposure, three modified CAI methods (i.e., the CAI+exposure, CAI+freeze, and CAI+SHOF methods) were proposed in the present study. Among these three methods, the CAI+exposure method showed great potential for monitoring item exposure; however, it did not have the best measurement precision. Nevertheless, in general, the CAI+exposure method is recommended.

Some future research lines are addressed as follows. First, in practice, many statistical and non-statistical constraints are considered during test assembling. It is important to apply a constraint-weighted item selection method (e.g., the maximum priority index) to satisfy all the constraints simultaneously during item selection. However, there is no constraint that monitors discrimination parameters for the maximum priority index. Therefore, it would be worth integrating the four CAI methods (i.e., CAI, CAI+exposure, CAI+freeze, and CAI+SHOF) described herein with the maximum priority index for item selection in CAT when many constraints are considered for item selection in CAT. Second, this study investigated the three modified CAI methods under a unidimensional context. The idea of the modified CAI methods can easily be extended to multidimensional contexts for item selection. Third, only three item exposure components were integrated with the CAI method in this study. If other item exposure control methods or test overlap methods are considered, an investigation of the efficiency of item selection methods with different item exposure methods or test overlap methods should be conducted.

Table 2 Exposure control of the six item selection methods under various conditions

Test length	Item selection methods	Item exposure rates										Max.	Chi-square statistics
		0	0-0.05	0.05-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	>0.5				
20	SMB	0.036	0.874	0.012	0.006	0.028	0.026	0.012	0.006	0.079	116.67		
	SMI	0.724	0.078	0.044	0.082	0.046	0.012	0.004	0.010	0.81	125.65		
	CAI	0.680	0.106	0.070	0.076	0.040	0.016	0.008	0.004	0.78	104.51		
	CAI+exposure	0.058	0.610	0.306	0.026	-	-	-	-	0.13	9.64		
	CAI+freeze	0.570	0.164	0.090	0.176	-	-	-	-	0.20	56.22		
	CAI+SHOF	0.574	0.154	0.082	0.190	-	-	-	-	0.20	54.34		
30	SMB	0.028	0.768	0.084	0.006	0.024	0.072	0.012	0.006	0.80	109.59		
	SMI	0.658	0.066	0.062	0.084	0.068	0.040	0.012	0.010	0.81	117.84		
	CAI	0.600	0.086	0.084	0.116	0.060	0.038	0.006	0.010	0.60	93.97		
	CAI+exposure	0.028	0.408	0.432	0.132	-	-	-	-	0.14	7.53		
	CAI+freeze	0.436	0.186	0.084	0.294	-	-	-	-	0.20	51.30		
	CAI+SHOF	0.430	0.188	0.082	0.300	-	-	-	-	0.20	48.75		

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. 397–472). Reading: Addison-Wesley.
- Chang, H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement*, 25, 333–341.
- Chang, H., & van der Linden, W. J. (2003). Optimal stratification of item banks in *a*-stratified computerized adaptive testing. *Applied Psychological Measurement*, 27, 262–274.
- Chang, H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211–222.
- Chen, S.-Y., Lei, P.-W., & Liao, W.-H. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 61(2), 471–492. <https://doi.org/10.1348/000711007X227067>.
- Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62, 369–383.
- Cheng, Y., Chang, H.-H., Douglas, J., & Guo, F. (2009). Constraint-weighted *a*-stratification for computerized adaptive testing with nonstatistical constraints: Balancing measurement efficiency and exposure control. *Educational and Psychological Measurement*, 69, 35–49.
- Georgiadou, E., Triantafyllou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed for 1983 to 2005. *Journals of Technology, Learning, and Assessment*, 5(8). Available from <https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1647>
- Huebner, A., Wang, C., Daly, B., & Pinkelman, C. (2018). A continuous *a*-stratification index for item exposure control in computerized adaptive testing. *Applied Psychological Measurement*, 42(7), 523–537.
- Huebner, A., Wang, C., Quinlan, K., & Seubert, L. (2015). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation. *Behavior Research Methods*, 48, 1443–1453. <https://doi.org/10.3758/s13428-015-0659-z>.
- Lee, Y. H., Ip, E. H., & Fuh, C. D. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement*, 68, 215–232.
- Leung, C. K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the *a*-stratified design with the Symptom-Hetter algorithm. *Applied Psychological Measurement*, 26, 376–392.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum.
- Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 35(4), 311–327.
- Su, Y.-H. (2015). The performance of the modified multidimensional priority index for item selection in variable-length MCAT. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research* (Vol. 140, pp. 89–97). Cham: Springer.
- Su, Y.-H. (2016). A comparison of constrained item selection methods in multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 40(5), 346–360.
- Su, Y.-H., & Huang, Y.-L. (2015). Using a modified multidimensional priority index for item selection under within-item multidimensional computerized adaptive testing. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research* (Vol. 89, pp. 227–242). Cham: Springer.

- Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.
- Yao, L. (2011, October). *Multidimensional CAT item selection procedures with item exposure control and content constraints*. Paper presented at the (2011) International Association of Computer Adaptive Testing (IACAT) Conference, Pacific Grove, CA.
- Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika*, *77*, 495–523.
- Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement*, *37*, 3–23.

Constant CSEM Achieved Through Scale Transformation and Adaptive Testing



Dongmei Li

Abstract Conditional standard error of measurement (CSEM) indicates the level of measurement precision at a particular true score or ability level. Having a constant CSEM across all scores not only simplifies score interpretation and score reporting, but also contributes to the fairness of testing. This paper compares two fundamentally different approaches to achieving constant CSEMs: CSEM stabilizing scale transformations and computer adaptive tests (CATs) with fixed-precision stopping rules. Through conceptual comparison and empirical illustration, this study shows that the two approaches produce score scales that are nonlinearly related to each other, and each achieving the goal of equalizing the CSEMs on its own scale. Procedures for equalizing the CSEMs of a CAT that is not designed to have fixed precision are provided, and implications for transitioning from linear tests with equal CSEMs to CATs are also discussed.

Keywords Equal CSEM · Variance stabilizing transformation (GVS) · Fixed-precision CAT

1 Introduction

Under the framework of classical test theory (CTT), any observed test score is composed of a true score and some measurement error. Whereas the standard error of measurement (SEM), which is the standard deviation of measurement errors across all examinees, is an indication of the average measurement error across the entire score scale, the conditional standard error of measurement (CSEM) is an indication of measurement error at each true score level. If CSEMs vary across the score scale, knowing the CSEMs for specific true scores allows for the construction of more accurate confidence intervals around observed scores,

D. Li (✉)
ACT, Inc, Iowa City, IA, USA
e-mail: dongmei.li@act.org

compared to using the overall SEM, and therefore enhances the interpretation of scores. The Standards for Education and Psychological Testing (AERA, APA, & NCME 2014) recommends that CSEMs be reported in addition to the overall SEM when measurement errors vary substantially across the score scale.

Numerous methods have been proposed using CTT or item response theory (IRT) to estimate the CSEMs of raw scores (Feldt 1984; Felt and Qualls 1996; Lord 1955, 1957, 1980; Mollenkopf 1949) and reported scale scores (Brennan and Lee 1999; Felt and Qualls 1998; Kolen et al. 1992, 1996; Lee et al. 1998, 2000). In order to simplify score reporting and score interpretation while also increasing fairness, some testing programs, including the two major college admission tests in the United States—the ACT (ACT 2019) and the SAT (College Board 2017)—develop their tests to have approximately equal CSEMs across their score scales.

There are at least two different approaches to achieving constant CSEMs in operational testing programs. One is through CSEM stabilizing transformations using methodologies described by Kolen (1988), Li et al. (2014), or Moses and Kim (2017). The other is through computer adaptive tests (CATs) with fixed precision stopping rules (Wainer 2000). There are numerous applications of each approach in operational testing programs, but it seems that these two approaches are so fundamentally different that they are seldom compared directly, though some researchers briefly addressed the impact of both approaches (e.g., Yi et al. 2006).

The purposes of this study are to compare the two approaches, explore solutions for equalizing CSEMs from other CAT designs, and draw implications for transitioning from linear tests with constant CSEMs to CATs. Specifically, the study investigates the following research questions:

1. What are the differences and similarities for tests whose CSEM is made constant through CSEM stabilizing scale transformations versus fixed-precision CATs?
2. Is there a way for CAT administrations to report scores with constant CSEM if the CAT algorithm ends the test with rules other than fixed precision?
3. When transitioning from linear tests to CATs, if the linear test has been scaled to have constant CSEM, is it possible for the CAT to produce interchangeable scores with the linear forms and maintain the same constant CSEM property?

Investigations and findings for each research question are described below, followed by conclusions and discussion.

2 Comparison of the Two Approaches

The two approaches are first compared conceptually following a description of each. Then procedures and results from an empirical comparison based on simulated data are described.

2.1 *Conceptual Comparison*

CSEM stabilizing transformations apply non-linear transformations to the original score scale to alter the CSEM at different points along the score scale (Brennan and Lee 1999; Kolen et al. 1992; Lord 1980). To equalize the CSEMs, a transformation can stretch the part of the scale where the CSEMs are lower (with a high-slope transformation) and compress the part of the scale where the CSEMs are higher (with a low-slope transformation). Different CSEM stabilizing approaches have scale transformations based on different models and different methodologies, but they all work in a similar manner by stretching or compressing different parts of the score scale.

The most commonly used scale transformation for equalizing CSEMs is the arcsine transformation (Freeman and Tukey 1950). It has been applied to equalize the CSEM of scale scores of various testing programs (ACT 2019; Ban and Lee 2007; College Board 2017; Kolen 1988; Kolen et al. 1992). However, the arcsine transformation is only appropriate when the conditional errors follow a binomial distribution or at least when the shape of raw score CSEMs is similar to that of binomially distributed errors. The generalized variance stabilizing (GVS) method suggested by Li et al. (2014) is a more general method that can be used with any test score type or with any error model as long as raw score CSEMs can be obtained. Given the CSEMs of the raw scores, the CSEMs of any transformation of the raw scores can be estimated using the delta method. The GVS method obtains the transformation function through numerical integration by setting the CSEM of transformed scores to be a constant. A third method uses cubic polynomial transformations of raw scores with coefficients of the polynomial functions obtained by minimizing the CSEM differences across the transformed scale scores (Moses and Kim 2017). Results from these three approaches have been compared in both multiple-choice tests (Moses and Kim 2017) and mixed-format test scaling (Wang and Kolen 2016). As expected, these three approaches produce similar results when the raw score errors follow a binomial distribution or approximately so, but the GVS method and the cubic transformation method are more effective when error distributions deviate more from binomial distributions.

In fixed-precision CAT administrations, items are selected based on estimated student ability levels, and the test is stopped when a pre-specified measurement precision—usually defined as the inverse of the square root of the test information function—is achieved at estimated ability levels. Fixed-precision CAT can result in equal CSEM for all examinees in ideal situations with a sufficiently large item pool including items with a broad distribution of item difficulty. In practice, however, fixed-precision CAT is often combined with a maximum test length or a maximum test time, because the pre-specified precision may not be achieved for certain examinees within a reasonable amount of time given limitations of the item pool.

Perhaps the most obvious difference between the two approaches is that one changes the score scale and the other changes which test items are presented. CSEM stabilizing transformations change the original score scales of a linear test with non-

equal CSEMs to a score scale with equal CSEMs, and fixed-precision CAT changes the type and the number of test items selected depending on the ability level (denoted by θ) of the examinees until a pre-specified level of measurement precision (i.e., a certain CSEM) is achieved. Suppose that the linear test (with unequal CSEMs) before scale transformation and the fixed-precision CAT each comprises items drawn from the same IRT-calibrated item pool, the linear test and the fixed-precision CAT are on the same θ score scale. However, the linear test differs from the fixed-precision CAT in that the CSEM of the linear test varies along the θ score scale, but the CSEM of the fixed-precision CAT is stable along the θ score scale. After a CSEM stabilizing transformation is applied to the linear test, however, scores from the linear test and the fixed-precision CAT are no longer on the same score scale but are nonlinearly related because a nonlinear transformation has been applied to the θ score scale of the linear test. Since both linear and nonlinear transformations are permissible under the assumptions that are often made when fitting item response models, there are not any compelling psychometric reasons that one scale should be preferred over another because of the essentially non-interval nature of score scales in educational measurement (Feuerstahler 2019; Kolen and Brennan 2014; Lord 1980). Therefore, score scales obtained from these two fundamentally different approaches are equally defensible from a psychometric scaling perspective.

2.2 Empirical Comparison

To illustrate how these score scales are related and how they compare in terms of reliability, CSEM, and score distributions, an empirical comparison was conducted based on simulations. Provided below is a high-level description of the major steps taken in the data simulation process.

1. Items for a 60-item linear test and those for a 600-item CAT pool were generated using a 3-parameter logistic (3PL) IRT model, with the shapes of the item parameter distributions informed by those from a large-scale testing program: $a \sim \text{lognormal}(0, 0.35)$, $b \sim \text{normal}(0, 1)$, and $c \sim \text{beta}(4.8, 20)$.
2. Based on the CSEMs of the θ scores calculated using the linear form item parameters, a GVS transformation was applied to the θ scores so that the transformed scale scores have an approximately equal CSEM of 2 along the score scale.
3. In order to simplify future calculations of GVS transformed scores for any θ scores, a 7th-degree polynomial function was fit to the θ -to-GVS score transformations obtained in step 2.
4. Two samples of examinees were simulated: (1) 10,000 examinees from a standard normal distribution; (2) 500 examinees at each θ score point within the range of -4 to 4 at 0.2 intervals. The first sample was mainly used for the comparison of score distributions, and the second sample was mainly used for the calculation of CSEM and reliability.

5. Item responses for the two samples of examinees were generated for the linear test, and θ scores for all these examinees were estimated based on the item responses.
6. A fixed-precision CAT administration was simulated, and all examinee scores were estimated based on their item responses generated from the CAT administration. The CAT was stopped when a precision of at least 0.3 was reached. The R package *catR* (Magis and Barrada 2017; Magis et al. 2017) was used for all the CAT simulations. No content balance or exposure control was used for this illustration.
7. GVS transformed scale scores were calculated based on the estimated θ scores on the linear test using the polynomial function obtained in step 3.
8. The scatter plots of true and estimated scores were produced for the linear test on both the θ and the GVS transformed scales as well as for the fixed-precision CAT scores on the θ scale.
9. Reliability of the scores analyzed in step 8 was calculated as the squared correlation between true and estimated/observed scores.
10. CSEMs for the linear test GVS scores and those of the CAT θ scores were calculated as the standard deviation of the 500 examinee scores at each θ score (or at the corresponding GVS scale) score point.
11. To facilitate comparisons of score distributions between the linear test GVS scale scores and the CAT θ scores, the θ scores from the CAT were linearly transformed to scale scores so that the CAT scale scores would have the same mean and standard deviation as the linear test GVS scale scores.

Figure 1 shows the CSEMs of the linear test on the θ score scale calculated as the inverse of the square root of the test information function. As expected, the CSEM is lower in the middle and higher at the ends of the θ scale. To achieve constant CSEMs, a scale transformation is needed to stretch the middle and compress the ends of the θ scale (i.e., the transformation function should have a higher slope in the middle and lower slope at the ends). Figure 2 shows the raw-to-scale transformation

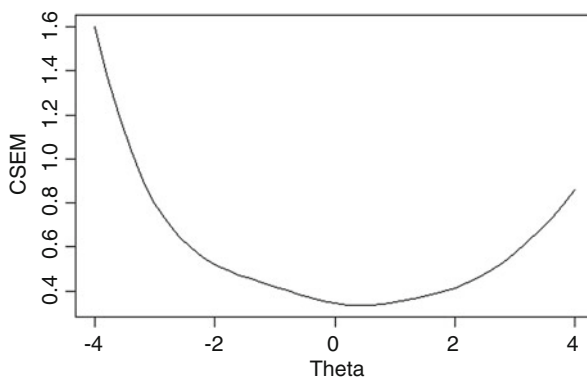


Fig. 1 CSEM of θ scores

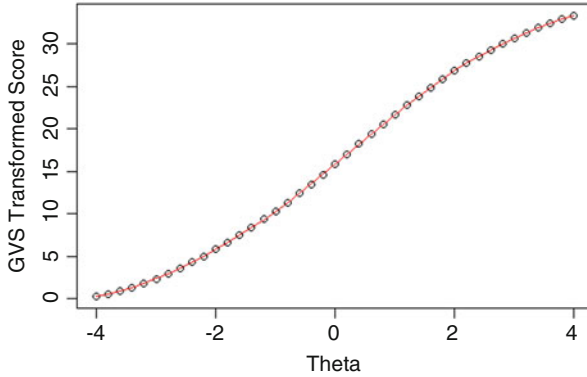
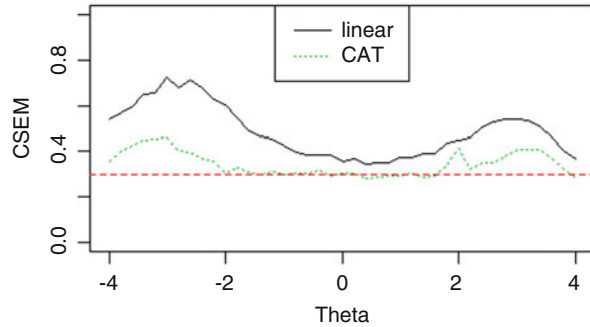


Fig. 2 GVS transformation of θ scores

Fig. 3 CSEM of the linear form and CAT

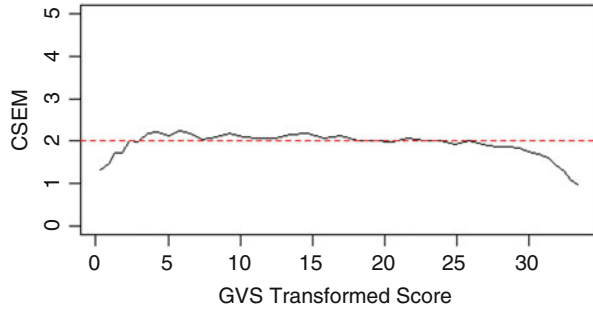


obtained using the GVS method through a numerical integration. The red curve represents the fitted 7th-degree polynomial function, which can be used to calculate the GVS transformed score for any θ value.

There is no easy formula to calculate the CSEMs of CAT scores, so the CAT CSEMs were obtained through simulations. As described in the steps above, the CAT CSEMs were obtained by simulating 500 examinees at each θ score point from -4 to 4 at intervals of 0.2 and calculated as the standard deviation of the estimated θ scores at each true θ value. This approach was also used to calculate the linear test CSEMs to be compared with that obtained from the test information function.

Figure 3 shows the CSEMs obtained through the simulation for the linear test and the fixed-precision CAT. Note that the shape of the linear test CSEMs obtained in this way is different from that obtained through the test information function (Fig. 1). Instead of continuously going higher toward the two ends of the θ scale, the CSEM obtained empirically shows a decline of the CSEM toward the ends. This was due to the fact that the IRT scale was truncated to a range of -4 to 4 for all the simulations. The fixed-precision CAT is stable with the pre-specified threshold of a CSEM of 0.3 within the range of -2 to 2 on the θ score scale, but CSEM values

Fig. 4 CSEM of the GVS transformed linear test



were higher outside that range. The reason was that within the limit of the item pool, the minimum CSEM that could be reached after exhausting all the items was greater than 0.3. It is also worth noting that the linear test has a higher CSEM than the CAT all along the score scale due to the specific precision level chosen for the CAT. A higher value for the stopping rule (e.g., 0.32) should have produced a CAT with more similar CSEMs to the linear test along parts of the scale, but a closer match of the CSEMs for these two tests was not required for the purposes of this illustration.

Figure 4 shows the CSEMs of the GVS transformed scores from the linear test based on the transformation function shown in Fig. 2. Though the GVS method was based on the theoretical CSEMs of θ scores but evaluated using an empirical approach that showed a different pattern at the two ends due to score truncation, the CSEM of the GVS transformed scores is fairly consistent along the score scale.

With the nonlinear relationship between θ scores and GVS scale scores, the shape of the distributions of examinee scores is slightly impacted. Figure 5 shows the CAT linearly transformed scale scores and the linear test GVS scale scores of the 10,000 examinees drawn from a standard normal distribution on the θ scale (with the CAT scores rescaled to have the same mean and standard deviation as the GVS scores). Whereas the CAT scale scores preserve the normality of the score generating distribution, the GVS score scale slightly reduced the score frequency in the middle and increased the score frequency at the ends. Yet, the GVS score distribution does not look dramatically different from that of the CAT scores.

Figure 6 shows the scatter plots of estimated scores vs. true scores from the linear test for both the GVS and the θ score scales, as compared with those of the θ scores from the fixed-precision CAT. Figure 6 also shows estimated reliability indices of the different test scores. These plots are based on the samples of 500 simulated examinees at each θ score point. A comparison of Fig. 6b, c shows that θ scores estimated from the fixed-precision CAT are more reliable with a more consistent conditional variance along the θ score scale. These results are expected given what is shown in Fig. 3. A comparison of Fig. 6a, b shows that the GVS transformation not only stabilized the conditional variance given true scores but also increased the reliability from 0.955 for the θ scores to 0.971 for the GVS scores.

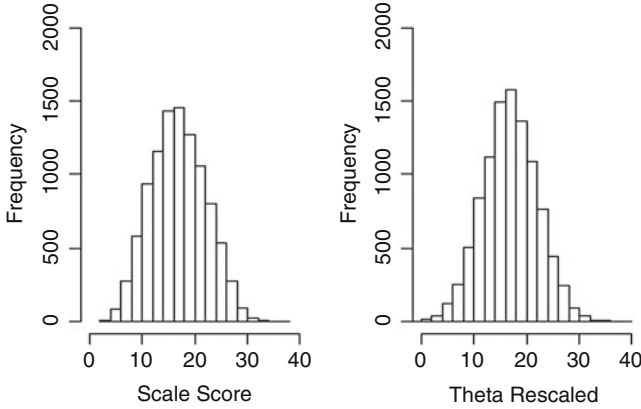


Fig. 5 Distributions of scores from the equal-CSEM linear test and the fixed-precision CAT

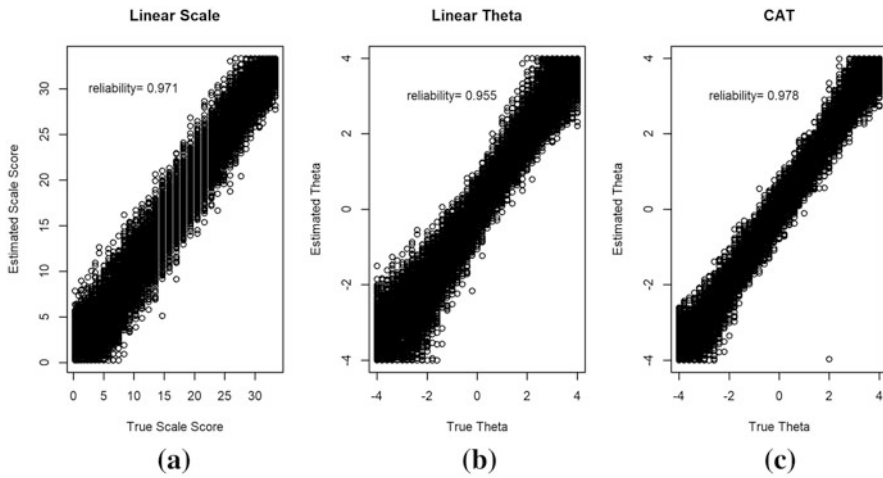
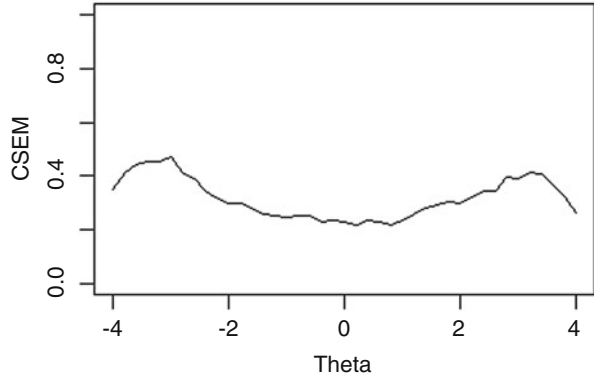


Fig. 6 Scatter plots of true and estimated scores and reliability (a) Linear scale score (b) Linear θ score (c) CAT θ score

3 CSEM Stabilizing Transformation of CAT Scores

Even though all CAT administrations select test items according to examinees' estimated ability level, not all CATs are designed to have fixed-precision stopping rules. Even for a fixed-precision CAT, a maximum test length or a maximum test time is often imposed to avoid very long testing times. When other stopping rules are used, CAT administrations may end up with CSEMs that vary across the score scale just like a linear test. The above section shows that scale transformation and fixed-precision CAT can be used to achieve a constant CSEM, but the use of scale transformation is not limited to linear tests. If desired, it can be used to equalize the CSEMs of CAT scores, too.

Fig. 7 CSEM of a fixed-length CAT



For example, Fig. 7 shows the CSEM of a 45-item fixed test length CAT using the 600-item pool, with CSEMs calculated as the standard deviation of the CAT scores of 500 examinees simulated at each θ value. Comparing Fig. 7 with Fig. 3 shows that, though the 45-item CAT test measures more precisely than the 60-item linear test across all levels of the θ score scale, the CSEMs of the 45-item CAT vary across the score scale in a manner similar to the 60-item linear test. If it is desirable to have a score scale for the CAT with a constant CSEM, the GVS transformation can be used to transform the θ scores. Provided below is a brief description of how this can be done.

1. Obtain CAT CSEMs on the θ score scale, usually through simulations of a CAT with a large number of examinees at each θ value within a range of θ values (e.g., -4 to 4) at small intervals (e.g., 0.1). Let $f(\theta)$ represent the resulting CSEMs as a function of ability.
2. Let $g(\theta)$ be the function used to transform θ to obtain a constant CSEM (i.e., the GVS transformation) and $g'(\theta)$ be the derivative of the transformation function.
3. According to the delta method (Dorfman 1938; Ver Hoef 2012), the CSEM of $g(\theta)$ at a given θ value is approximately $g'(\theta)$ times the CSEM of θ at that θ value. Therefore, if the CSEM of $g(\theta)$ is set to be a constant c , $g'(\theta)$ for each θ approximately equals c divided by the CSEM of θ .
4. Determine the desired scale score CSEM magnitude c on the transformed score scale (e.g., $c = 2$).
5. Calculate $cf(\theta)$ at each θ value, which approximates $g'(\theta)$ at each θ value.
6. Derive the GVS transformed scores $g(\theta)$ by integration or numerical integration.
7. Evaluate the CSEMs on the GVS transformed scores by transforming the CAT estimated θ scores of the 500 examinees simulated at each θ score point using $g(\theta)$ and calculate the standard deviation of the GVS scores at each θ point.

Figure 8 presents the results after step 7, which shows that the GVS transformation effectively stabilized the CSEM and provided a score scale that can be used to report the CAT scores with an approximately constant CSEM of 2 along the score scale.

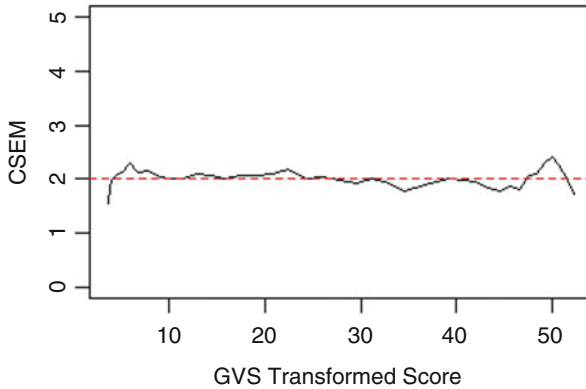


Fig. 8 CSEM of GVS-transformed CAT

4 Transitioning Linear Tests with Constant CSEM to CAT

When a testing program with linear tests with approximately equal CSEMs decides to transition to CAT, one of the major issues to be considered is whether the current score scale can be maintained. Certain features of the current scale can be maintained by using the same θ -to-scale score conversion for the CAT θ score estimates as the linear base form. However, whether the property of equal CSEM can be maintained depends on the specific features of the CAT.

To illustrate the CSEM of the score scales for different CAT designs when using the same GVS conversions as the linear test, CATs with different stopping rules were simulated using the same 600-item pool for which the b parameters were generated from a normal distribution (i.e., a peaked pool) and using another 600-item pool with b parameters generated from a uniform distribution (i.e., a flat pool). The stopping rules were: (1) fixed test length of 45 items (p_{45}), (2) fixed precision of 0.3 with a maximum test length of 45 items ($p_{se_{45}}$), (3) fixed test length of 45 items with maximum exposure rate of 0.2 ($p_{45_{20}}$), and (4) fixed test length of 45 items with maximum exposure rate of 0.2 and some content balance control ($p_{45_{20}_{cb}}$).

Figure 9 presents the resulting CSEMs of these various CAT administrations on the linear test score scale using the peaked pool (Fig. 9a) and the flat pool (Fig. 9b). Note that, in the legend of Fig. 9b, the letter “ f ” is used instead of “ p ” to represent the use of the flat pool. In only one condition—fixed test length with the peaked pool (p_{45} in Fig. 9a)—was the constant CSEM property maintained after the CAT scores were converted to the linear test scale scores. These results suggest that when transitioning from a linear test with a constant CSEM score scale to CAT, the equal CSEM property needs to be re-evaluated even when the scores are made comparable between the linear test and the CAT by using the same θ to scale score conversions. It may be difficult to maintain the constant CSEM property depending on the features of the item pool and the CAT design. A rescaling of the test may

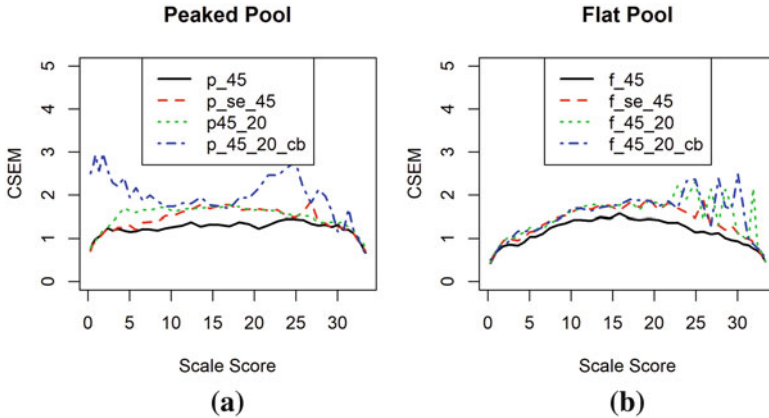


Fig. 9 CSEM of various CAT designs scaled to have comparable scores with an equal CSEM linear test (a) peaked pool (b) flat pool

be needed in order to create a score scale with constant CSEMs for the CAT test. Simulations like those done in this study can be used to inform decisions about whether a rescaling is needed for the CAT test if a constant CSEM is one of the primary requirements of the CAT score scale.

5 Conclusions and Discussion

This paper compared two different approaches to achieving constant CSEMs—CSEM stabilizing scale transformations and fixed-precision CATs. Using simulated data assuming a linear test and a CAT administration, the study demonstrated that these two approaches yielded different score scales that were nonlinearly related, but each possessing the equal CSEM property. Fixed-precision CAT achieved constant CSEM on the θ score scale, and non-linear transformations of the θ score scale were needed for the linear test to obtain constant CSEM. Other properties of the different score scales were also compared, including score distributions and reliability, with no strong evidence supporting the superiority of one approach over another, which was expected given that nonlinear transformations are permissible for IRT scales and that there are no compelling psychometric reasons to prefer one over another (Kolen and Brennan 2014). It might be worth pointing out that the GVS scale transformation used to stabilize the linear test CSEM in this study slightly increased the test score reliability. Though not presented, the reliability estimates based on the normally distributed 10,000 examinees showed a similar pattern of results. However, it is not known how the change in reliability can be generalized to other situations. Further studies are needed to investigate how equalizing CSEM through scale transformation impacts reliability of the test scores.

This study also demonstrated the feasibility of applying the GVS method to equalize the CSEMs of CAT administrations when stopping rules other than fixed precision are used. The GVS method was used because of its simplicity and flexibility. Other methods such as the cubic spline method may be applied also, though the arcsine transformation cannot be directly used because its application is limited to number correct score scales. The demonstration estimated CSEMs for CATs based on simulations. Further investigations can be done to evaluate the use of different scale transformation methods and the use of different CSEM estimations for the CAT θ scores.

This study also investigated potential outcomes when a linear test with constant CSEM transitions to CAT with the goal of maintaining the comparability of scores. Results indicated that unless the CAT administration has a CSEM function shaped similarly to the linear test, the equal CSEM property of the score scale may be compromised after transitioning to CAT, even when the CAT is designed to have a fixed-precision stopping rule.

One factor not accounted for in this study was the impact of equating on the CSEM stabilizing transformations applied to linear tests. Though a test can be scaled to have equal CSEM for a base form at the time of scaling, subsequent equating of new test forms may distort the CSEMs to various extents if the new forms are not strictly parallel to the base form. Further research should be conducted to investigate the impact of equating new test forms for linear tests and the impact of updating item pools for CAT on the stability of CSEMs over time.

References

- ACT. (2019). *The ACT technical manual*. Iowa City: Author.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: Author.
- Ban, J., & Lee, W. (2007). *Defining a score scale in relation to measurement error for mixed format tests* (CASMA research report, no. 24). Iowa City: Center for Advanced Studies in Measurement and Assessment.
- Brennan, R. L., & Lee, W.-C. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement, 59*(1), 5–24.
- College Board. (2017). *SAT[®] suite of assessments technical manual: Characteristics of the SAT*. New York: Author.
- Dorfman, R. (1938). A note on the δ -method for finding variance formulae. *The Biometric Bulletin, 1*(125, 126), 129–137.
- Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44*, 883–891.
- Feldt, L. S., & Qualls, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement, 33*, 141–156.
- 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.

- Feuerstahler, L. (2019). Metric transformations and the filtered monotonic polynomial item response model. *Psychometrika*, *84*(1), 105–123.
- Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and square root. *The Annals of Mathematical Statistics*, *21*, 607–611.
- Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement*, *25*(2), 97–110.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.
- 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.
- Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*(2), 129–140.
- Lee, W., Brennan, R. L., & Kolen, M. J. (1998). *A comparison of some procedures for estimating conditional scale-score standard errors of measurement* (Iowa Testing Programs Occasional Paper No. 43). Iowa City: University of Iowa.
- Lee, W., 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.
- Li, D., Woodruff, D., Thompson, D., & Wang, H. (2014, April 3–6). *An alternative way to achieve constant conditional standard error of measurement*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, Pennsylvania.
- Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*, 325–336.
- Lord, F. M. (1957). Do tests of the same length have the same standard errors of measurement? *Educational and Psychological Measurement*, *17*, 510–521.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates, Publishers.
- Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: recent updates of the package *catR*. *Journal of Statistical Software, Code Snippets*, *76*(1), 1–19. <https://doi.org/10.18637/jss.v076.c01>.
- Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. New York: Springer.
- Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika*, *14*, 189–229.
- Moses, T., & Kim, Y. (2017). Stabilizing conditional standard errors of measurement in scale score transformations. *Journal of Educational Measurement*, *54*(2), 184–199.
- Ver Hoef, J. M. (2012). Who invented the delta method? *The American Statistician*, *68*, 124–127.
- Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.
- Wang, S., & Kolen, M. J. (2016). Evaluation of scale transformation methods with stabilized conditional standard errors of measurement for mixed-format tests. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 4)* (CASMA Monograph Number 2.4, pp. 204–312). Iowa City: CASMA, The University of Iowa.
- Yi, Q., Wang, T., & Ban, J. (2006). Effects of scale transformation and test termination rule on the precision of ability estimates in CAT. *Journal of Educational Measurement*, *38*(3), 267–292.

Synergized Bootstrapping: The Whole is Faster than the Sum of Its Parts



Tim Loossens, Stijn Verdonck, and Francis Tuerlinckx

Abstract Re-sampling methods are popular for assessing uncertainty, for testing hypotheses, or for cross-validation because of their simplicity. They all rely on a similar scheme: generating replicated datasets by sampling data points from an original dataset, fitting a model or conducting a statistical test on each of these, and aggregating the results. However, when fitting the model or conducting the statistical test becomes time-consuming, re-sampling methods become impractical because of the many replications. Many methods have been proposed to alleviate the computational burden, but they generally do not incorporate two key features of re-sampled datasets. One, re-sampled datasets all stem from the same origin and therefore have similar characteristics. Two, there is a large class of cost functions for which the cost of a parameter set given data can be computed by summing its costs across the individual data points. As a consequence, once the costs of the individual data points are known, the parameter set's cost can be obtained for any of the cost functions related to one of the replicated datasets. The synergized bootstrap method put forward in this paper exploits these two features to accelerate the optimization procedures for re-sampling methods. It is applied to the non-parametric bootstrapping of the parameters of a univariate mixture model, of which the min-log-likelihood function can be shown to have multiple local minima, using the differential evolution heuristic as global optimizer. It is demonstrated that the synergized method can lead to incredible accelerations (up to 100-500 times faster) while being more accurate than the standard DE method.

Keywords Differential evolution · Finite mixture models · Maximum-likelihood optimization · Non-parametric bootstrapping · Re-sampling methods

T. Loossens (✉) · S. Verdonck · F. Tuerlinckx
KU Leuven, Leuven, Belgium
e-mail: tim.loossens@kuleuven.be

© Springer Nature Switzerland AG 2020
M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in
Mathematics & Statistics 322, https://doi.org/10.1007/978-3-030-43469-4_18

227

1 Introduction

Re-sampling methods can be used for a wide variety of statistical inference tasks such as uncertainty assessment (e.g., estimating confidence intervals), hypothesis testing (estimating tail-area probabilities), or cross-validation (estimating prediction errors to perform model selection). These methods prove their usefulness even or specifically in the absence of analytic results or when specific conditions of the analytic method, such as normality, are unjustified (Adzhubei et al. 2010; Cox and Mann 2008; Persson et al. 2013; Ramaswamy et al. 2003; Turnbaugh et al. 2009). The flexibility of re-sampling methods has made them a popular tool. Most well-known methods are the bootstrap (Efron and Tibshirani 1994), the permutation test (Good 2000), and different kinds of cross-validation methods (Hastie et al. 2016).

Re-sampling methods typically follow the same scheme. First, multiple datasets are generated by re-sampling from an original dataset. Then, a model is fitted to or a statistical test is conducted on all the re-sampled datasets. Ultimately, the results are combined and used, for instance, to assess the uncertainty of an estimator or to assess the robustness of a statistical test.

Notwithstanding the present-day computing power, a main disadvantage of re-sampling methods is still their computational burden (Mestdagh et al. 2015). This burden is conspicuously present when iterative methods are required to maximize a likelihood or minimize a cost function for every re-sampled dataset. Typical variance and bias calculations require between 50 and 200 bootstrap replications. Confidence limits are more costly; they may require between 1,000 and 2,000 replications (Efron 1987). Optimizers, especially global optimizers, go through multiple cost function evaluations during each iteration to update. Whenever the evaluation of the cost function becomes too costly, re-sampling methods become impractical.

Continued effort has been invested in alleviating the computational burden of re-sampling methods. The majority of these efforts focus on optimizing the process of fitting the model. This is achieved by, for example, reducing the required number of iterations for a single optimization (Andrews 1999; Bringmann et al. 2013; Cawley and Talbot 2008; Crainiceanu and Ruppert 2004; Davidson and MacKinnon 1999; Efron 1990; Halekoh and Højsgaard 2014; Lippert et al. 2011; Maho et al. 2014; Samuh et al. 2012; Shaw et al. 2006; Zhou and Stephens 2012, 2014), or splitting up the cost function into smaller, simpler parts (Hu and Kalbfleisch 2000). Another approach that was proposed was aimed at improving the post-processing of bootstrap estimates which in favorable situations allows a significant reduction of the necessary number of bootstrap replications (Efron 1990). Other improvements are custom-made for specific problems (Kleiner et al. 2011; Stamatakis et al. 2008; Zeng and Lin 2008). In general, these methods are not generic.

A special feature of re-sampling methods that can be more universally exploited is the large degree of overlap between re-sampled datasets. The re-sampled datasets all stem from the same origin and therefore have more or less the same characteristics. Therefore, when fitting the model to the re-sampled datasets, a similar task has

to be executed over and over again. The fingerprint re-sampling method (Mestdagh et al. 2015) exploits this feature by learning the relation between data characteristics and the corresponding model estimates. Once this relation has been established, it can be used to suggest more appropriate starting points for the iterative estimation process. After a few estimations, the fingerprint method can lead to a significant acceleration of the estimation procedures for the remaining re-sampled datasets. When the relation becomes sufficiently accurate, it may replace the estimation processes altogether.

The fingerprint method uses limited information to establish the relation between data characteristics and model parameters. Data characteristics are mapped to the final result of an optimization procedure; any information regarding the cost function obtained during the estimation is lost. This makes the fingerprint method susceptible to local minima. If the mapping between the data space and the model space is non-smooth or has discontinuities, the applicability of the fingerprint method can be called into question. The method has in fact only been tested for smooth, convex min-log-likelihood functions using local optimizers (Mestdagh et al. 2015). For such min-log-likelihood functions, the estimates of the re-sampled datasets typically lie closely together in the parameter space, and the relation between data characteristics and estimates can nicely be established. Under these conditions, the bootstrap distributions are unimodal. If a cost function has multiple local minima, it is possible that two datasets with similar features (they lie close together in the data space) map to very different parameter configurations (they are far apart in the model space). This results in multimodalities in the bootstrap distributions and can significantly obscure the relation between data characteristics and model parameters. Local optimizers are not adequate for such problems, and the way to build the relation between the data characteristics and the model estimates becomes unclear.

The prepaid method (Mestdagh et al. 2018) is another recent method that has been proposed for speeding up optimization procedures that also exploits the similarity of fitting procedures, but in a much broader context. The key idea of the prepaid method is that when the same model is used multiple times for different datasets, similar computations are required to arrive at estimations. The prepaid method capitalizes on this by creating a prepaid database in which, for a regular grid across the parameter space, the model is extensively simulated. Once this database has been constructed and stored, any estimation problem can be solved on the spot using advanced interpolation techniques. The main advantage of the prepaid method over the fingerprint re-sampling method is the humongous amount of available information in the database to construct the relation between the data space and the model space, making it less susceptible to local minima. Unfortunately, setting up the prepaid grid is computationally too burdensome for a one-time re-sampling task.

In this paper, the key ideas of the fingerprint and prepaid method are exploited for re-sampling methods in a slightly different fashion. For a large class of cost functions, the cost of a specific set of model parameters can be computed for all cost functions in the bootstrap without the need for extra computations (e.g., because all cost functions consist of additive contributions of the same data points). By evaluating the costs of the original data points, the cost of any re-sampled dataset can be obtained. Hence, when a specific cost function in the bootstrap is being optimized, we can keep track of the different locations in the parameter space that have been visited and save those that are of interest for the optimization of the other cost functions. In other words, we can construct prepaid databases on the fly which will help us with the initialization of the other optimization problems. In addition, when multiple cost functions in the bootstrap are being optimized at the same time, all visited locations in the parameter space and their corresponding costs for each of the functions can be made collectively available. As such, information can be exchanged between the different optimization procedures. The main advantage of this synergy is that due to the availability of collective information, the different cost functions can be scanned more rapidly and less cost function evaluations are required to reach the different global optima. Unlike the fingerprint method, the proposed method uses the prepaid databases that are constructed on the fly and that are optimized for each cost function independently. Hence, the method can be applied to cost functions with multiple local minima. As an example, the synergized bootstrap method described in this paper will be applied to the non-parametric bootstrapping of the parameter estimates of finite mixture models. There is no closed form expression for the optima of the min-log-likelihood function of a finite mixture model, and the function can have multiple minima (see e.g., McLachlan and Peel 2000). Finite mixture models therefore provide an ideal test case for the method. The differential evolution (DE) heuristic (Storn and Price 1997) will be used as global optimizer to fit the mixture models.

2 Non-parametric Bootstrapping

When fitting a model to data, a non-parametric bootstrap can be used to evaluate the uncertainty of the parameter estimates. The idea is to generate a number of new datasets (replications) based on the original and fit the model to each of them. That way, distributions of parameter estimates are obtained. These can then be used to compute standard errors or confidence bounds.

2.1 *Re-sampling: Generating New Datasets*

For a non-parametric bootstrap, a replicated dataset is created by sampling data points from the original dataset $\mathbf{y} = \{y_i\}_{i=1, \dots, n}^n$ with replacement until the sample

size n is reached. Note that we will focus in this paper on i.i.d. data (i.e., $y_i \sim F$ for all i and all y_i s are independent from one another). The replicated datasets will be denoted as \mathbf{y}^j , $j = 1, \dots, B$, where B refers to the number of replications.

The constituents of the replicated datasets are the same as those of the original dataset, namely, the data points y_i . The frequency of occurrence of the constituents, however, differs across replicated datasets. A replicated dataset \mathbf{y}^j is fully defined in terms of the original data points y_i and the frequencies f_i^j with which they appear. The notation f_i^j is used to refer to the frequency of occurrence of the data point y_i in the replicated dataset \mathbf{y}^j .

2.2 Optimization: Finding the Estimates

To find the parameter estimates of a model given data \mathbf{y} , a cost function (like a least-squares function or a min-log-likelihood function) $\mathcal{F} : \boldsymbol{\theta} \mapsto \mathcal{F}(\boldsymbol{\theta} | \mathbf{y})$ has to be optimized. A cost function \mathcal{F} relates a cost $\mathcal{F}(\boldsymbol{\theta} | \mathbf{y})$ to every parameter set $\boldsymbol{\theta}$ given the data \mathbf{y} . The parameter estimates $\hat{\boldsymbol{\theta}}$ correspond to the parameter set $\boldsymbol{\theta}$ for which the cost is optimal. This set can be found at the global extremum of the cost function. For least-squares and min-log-likelihood functions, this extremum corresponds to a global minimum.

The functional form of the cost function depends on the particular model. In this paper, we are primarily concerned with cost functions that have multiple local extrema (minima or maxima, depending on the context). For such functions, local optimization algorithms are inadequate and instead global optimization heuristics are required.

We will rely on the differential evolution (DE) heuristic for global optimization because it is a simple, reliable, and all-round optimizer (Storn and Price 1997) that has got the required properties for the synergized method proposed in this paper to work. Furthermore, it has been used for fitting finite mixture models (Boonthiem et al. 2017; Kwedlo 2014). The ideas underlying the DE method are nevertheless more broadly applicable to similar kinds of optimizers. Here, we briefly explain the ideas behind DE. For a more elaborate description, see Storn and Price (1997).

2.2.1 Differential Evolution

Differential evolution is a parallel, stochastic, direct search method for solving continuous optimization problems (Storn and Price 1997). It relies on a population $\mathcal{P}(g)$ of NP model parameter vectors or “agents” $\boldsymbol{\chi}_a(g)$, $a = 1, \dots, NP$. These agents are updated over generations g . The population size NP does not change during the optimization.

A DE optimization starts by sampling an initial population $\mathcal{P}(0)$ of NP agents $\chi_a(0)$ from a prior distribution on the parameter space. This prior distribution should cover the entire search space (the part of the parameter space that is of interest). The cost $\mathcal{F}(\chi_a | \mathbf{y})$ of each agent is also computed.

Subsequently, the populations are iteratively updated. Updating a population happens in different stages. First, NP mutant vectors are constructed by adding the weighted difference of two agents to a third. All of the triplets that constitute a mutant are unique. In total, there are as many mutant vectors as there are agents in the population. Second, a crossing-over between agents and mutant vectors takes place. During this procedure, the parameters of an agent are intermixed with the parameters of one of the mutant vectors. The resulting vectors make up the offspring and will be referred to as children. Third, the cost of each child is compared to that of the agent to which it is kin (the parent, i.e., the agent of which it inherited part of its parameters). If the child's cost is more optimal, it replaces its parent in the population. Otherwise, the parent lives on.

The total cost of the population $\mathcal{P}(g)$ can only improve from one generation to the next. Agents that are more optimal will attract other agents. In doing so, the population increasingly focuses on regions in the parameter space that are more interesting for finding a global optimum. The spread of the population naturally shrinks across generations and so do the weighted differences between the agents. As such, the optimization scheme naturally adapts from that of a global search into that of a local search.

Algorithm 1: standard DE

```

input : data and cost function
output: estimates, cost

1 agents = random_initialization(NP);
2 agents_costs = cost_function(data, agents);
3 for  $g = 1$  to  $n\_iter$  do
4   | children = mutate_and_crossover(agents);
5   | children_costs = cost_function(data, children);
6   | agents = compare_and_eliminate(agents, children, agents_costs, children_costs);
7   | agents_costs = update_costs(agents_costs, children_costs);
8 end
9 best_agent_position = find_optimal(agents_costs);
10 estimates = get_agent(agents, best_agent_position);
11 cost = get_cost(costs, best_agent_position);

```

2.2.2 Multiple Cost Functions

In the context of non-parametric bootstrapping, there is a cost function $\mathcal{F}^j : \theta \mapsto \mathcal{F}(\theta | \mathbf{y}^j)$ associated with every replicated dataset \mathbf{y}^j . These cost functions are entirely independent from one another. To find the optimum of each function, B

different optimization procedures have to be run – exactly as many optimizations as there are replications.

Given an agent χ_a^j of the population \mathcal{P}^j . In a traditional DE setup, the only relevant cost to know is $\mathcal{F}(\chi_a^j | \mathbf{y}^j)$. When offspring is created, it is this cost that is used during the selection stage of the updating step. The costs $\mathcal{F}(\chi_a^j | \mathbf{y}^k)$ with $j \neq k$ are never computed. Yet, such “external” costs can hold relevant information.

For one, if both the usual (internal) costs and the external costs are available to all populations, the amount of information on the different cost functions increases by a factor B . Instead of only NP function values (costs), $B \times NP$ function values are known per cost function. This significantly augments the sheer search power, making it easier to uncover the relevant regions for the different cost functions.

An additional benefit in the context of non-parametric bootstrapping is that the cost functions all look alike. The functional form of the cost functions is the same for all replicated datasets, since it is determined by the model. They only differ in the dataset they are associated with; the function \mathcal{F}^j is associated with the replicated dataset \mathbf{y}^j . Replicated datasets show a lot of overlap since they are all derived from the same origin \mathbf{y} , hence the similarity between cost functions.

As a consequence, the different populations will be attracted toward the same regions in the parameter space, and agents from a specific population can therefore learn a lot from external agents, especially during the global search stage of the optimization. Only when the populations start homing in on their specific global minimum will the gains that can be made by considering external agents reduce.

Yet, computing the agents’ costs is for many practical applications the most time-consuming task during the optimization. It is therefore generally not a good strategy to compute costs from obtained less promising external agents.

2.3 Collective Information

For a specific class of cost functions, however, external costs can be obtained for free given that the re-sampled datasets are constructed from the same original data points. There is a large class of cost functions with the property that the cost of a parameter set θ for an entire dataset \mathbf{y} is equal to the sum of the costs of the individual data points that make up the dataset (here we rely on the i.i.d. assumption made above):

$$\mathcal{F}(\theta | \mathbf{y}) = \sum_{i=1}^n \mathcal{F}(\theta | y_i).$$

By extension, the cost $\mathcal{F}(\boldsymbol{\theta} | \mathbf{y}^j)$ can straightforwardly be constructed from the costs $\mathcal{F}(\boldsymbol{\theta} | y_i)$ of the individual data points as well:

$$\mathcal{F}(\boldsymbol{\theta} | \mathbf{y}^j) = \sum_{i=1}^n f_i^j \mathcal{F}(\boldsymbol{\theta} | y_i).$$

The cost of the parameter set $\boldsymbol{\theta}$ for a re-sampled dataset \mathbf{y}^j is simply a weighted sum of the costs of the individual data points that make up the original dataset, where the weights are determined by the frequency of occurrence of the data points.

In order to update all DE populations of an entire non-parametric bootstrap, $n \times B \times NP$ individual costs have to be computed – n individual costs per agent. From these, the costs $\mathcal{F}(\boldsymbol{\chi}_a^j | \mathbf{y}^j)$ which are necessary for the selection procedure can be constructed. The external costs $\mathcal{F}(\boldsymbol{\chi}_a^k | \mathbf{y}^j)$ with $j \neq k$ can also be constructed from these, without the need for extra computations. In a traditional DE optimization, this external information is ignored, but in the next section we will show how this information can be put to good use.

3 The Synergized Bootstrap

Instead of initializing a different population \mathcal{P}^j for every cost function \mathcal{F}^j (as would be the standard procedure for the traditional DE), only one population \mathcal{P} of NP agents is initialized. The costs of these agents are computed for all the cost functions \mathcal{F}^j , resulting in $B \times NP$ costs. Note that only $n \times NP$ computations are required for this (not $n \times B \times NP$, as for the traditional DE). The agents of the population \mathcal{P} serve as initial generation for each optimization: $\forall j \in \{1, \dots, B\} : \mathcal{P}^j(0) = \mathcal{P}$.

The agents of one of the populations, say \mathcal{P}^1 , start updating like they would normally do. However, for each generation $\mathcal{P}^1(g)$, the children \mathbf{v}_a^1 are evaluated in all the cost functions (again, this does not require extra computations). If the cost $\mathcal{F}(\mathbf{v}_a^1 | \mathbf{y}^j)$ of the child \mathbf{v}_a^1 is better than the cost $\mathcal{F}(\boldsymbol{\chi}_a^j | \mathbf{y}^j)$ of the current agent $\boldsymbol{\chi}_a^j$ in population \mathcal{P}^j , the agent is replaced (the agent $\boldsymbol{\chi}_a^j$ and the child \mathbf{v}_a^1 have the same sub-index a – they have the same position within their own respective populations – but they are not parent and child because they come from different populations). So, although only NP agents are actively evolving, all agents of all populations are being updated, and it is possible that some agents appear multiple times across populations.

This initial updating scheme is sufficient for establishing the regions in the parameter space that are of interest for the different cost functions. The populations \mathcal{P}^j only retain agents that optimize their current cost. The agents that are being

retained can vary across populations, which still enables the populations to focus on their particular problem. Essentially, a prepaid database is being constructed for each cost function, each consisting of its own NP parameter sets and their related costs. The initial global search of the DE optimizations, which is quasi-random anyways, is shared across populations, and thus the computational cost of this “burn-in” is significantly reduced.

At some point, there will be populations that no longer adopt children from the population \mathcal{P}^1 because its agents are drifting away from their particular regions of interest. Typically, the populations whose cost functions differ most from \mathcal{F}^1 will diverge earlier. The first time a population, say \mathcal{P}^2 , is unable to adopt children for several generations in a row, its agents are also made to actively evolve. From then onward, there are two DE populations independently breeding children. All these children are still evaluated in all cost functions (because this does not require any extra computations). The supply of children which the populations can adopt to improve their prepaid database doubles. The idea is that populations corresponding to cost functions that have more in common with \mathcal{F}^2 might still be able to learn from the agents of the population \mathcal{P}^2 .

Once there is another population that can no longer learn from the two populations, its agents are also made to evolve. One after another, the populations start their own individual evolution. The rate at which populations launch their own evolutionary updating strongly depends on the similarity between cost functions. In the worst case, agents cannot learn much from external information, and the populations all start their own search almost immediately. In that case, B different DE optimizations are being done. Hence, the computation time of the synergized bootstrap is at worst equal to that of a traditional bootstrap where all B optimizations are done independently.

It can happen that multiple populations fail to update for several generations in a row. This implies that these populations require information that is different from the information provided by the currently active populations. It is however possible that several of these populations require similar information. Hence, it would suffice to just activate a selection of them. The other populations that were not activated can then learn from those that have been activated. As such, the rate at which the computational burden increases can be limited.

Populations that have started evolving can also still adopt children from other populations. This is an additional advantage over the traditional DE. Populations that have converged to a local minimum can still be corrected through the adoption of external children.

Algorithm 2: synergized DE

```

input : data, bootstrap frequencies and cost function
output: estimates, cost

1 agents = random_initialization(NP);
2 agents = replicate_agents(B);
3 individual_costs = cost_function(data, agents);
4 agents_costs = weighted_sum(individual_costs, frequencies);
5 n_active_populations = 1;
6 for  $i = 1$  to  $B$  do
7   | failed_counter( $i$ ) = 0;
8 end
9 for  $g = 1$  to  $n\_iter$  do
10  | for  $j = 1$  to  $B$  do
11    | updated( $j$ ) = false;
12    end
13    temporary_agents = agents;
14    for  $j = 1$  to  $n\_active\_populations$  do
15      | active_agents = get_population(temporary_agents,  $j$ );
16      | children = mutate_and_crossover(active_agents);
17      | children_individual_costs = cost_function(data, children);
18      | for  $k = 1$  to  $B$  do
19        | local_frequencies = get_frequencies(frequencies,  $k$ );
20        | children_local_costs = weighted_sum(children_individual_costs,
21          | local_frequencies);
22        | local_agents2 = get_population(temporary_agents,  $k$ );
23        | local_agents_costs = get_costs(agents_costs,  $k$ );
24        | local_agents = compare_and_eliminate(local_agents2, children,
25          | local_agents_costs, children_local_costs);
26        | local_agents_costs = update_costs(local_agents_costs, children_local_costs);
27        | if not equals(local_agents, local_agents2) then
28          | | updated( $k$ ) = true;
29        | end
30        | agents = insert_population(local_agents, agents,  $k$ );
31        | agents_costs = insert_costs(local_agents_costs, agents_costs,  $k$ );
32      | end
33    | end
34    n_activated_populations = 0;
35    for  $j = n\_active\_populations + 1$  to  $B$  do
36      | if updated( $j$ ) == false then
37        | | failed_counter( $j$ ) = failed_counter( $j$ ) + 1;
38        | else
39        | | failed_counter( $j$ ) = 0;
40        | end
41        if failed_counter( $j$ ) > max_fails and n_activated_populations < max_activations
42        | then
43          | n_active_populations = n_active_populations + 1;
44          | agents = swap_populations(agents, n_active_populations,  $j$ );
45          | agents_costs = swap_costs(agents_costs, n_active_populations,  $j$ );
46          | n_activated_populations = n_activated_populations + 1;
47        | end
48      | end
49    | end
50  | end
51 end

```

Algorithm 2: synergized DE (continued)

```

46 for  $i = 1$  to  $B$  do
47   local_agents = get_population(agents,  $i$ );
48   local_agents_costs = get_costs(agents_costs,  $i$ );
49   best_agent_position = find_optimal(current_agents_costs);
50   local_best = get_agent(local_agents, best_agent_position);
51   estimates = concatenate(estimates, local_best);
52   cost( $i$ ) = get_cost(costs, best_agent_position);
53 end

```

4 Application: Finite Mixtures

The ideas described above were applied to the non-parametric bootstrapping of the parameters of univariate finite mixture models. The probability density function $p(x)$ of a univariate finite mixture is defined as

$$p(x) = \sum_{m=1}^N \omega_m \phi_m(x),$$

where $\phi_m(x)$ denote univariate normal density functions (the components),

$$\phi_m(x) = \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{(x-\mu_m)^2}{2\sigma_m^2}},$$

and $\omega_m \geq 0$ denote weights which sum up to one. The min-log-likelihood function of the mixture model is given by

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\omega} | \mathbf{y}) = - \sum_{i=1}^n \ln \left(\sum_{m=1}^N \frac{\omega_m}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{(y_i-\mu_m)^2}{2\sigma_m^2}} \right). \quad (1)$$

By minimizing this function, the parameter estimates of the finite mixture can be found for a given dataset $\mathbf{y} = \{y_i\}_{i=1, \dots, n}$.

The min-log-likelihood function (1) can have multiple local minima. In Fig. 1, an example of such a local minimum is shown. In Panel (a), the two-component mixture distribution $p(x)$ is drawn from which 50 data points were simulated. A histogram of the data points is shown in Panel (b). The distribution corresponding to the global minimum is indicated as well. In Panel (c), the histogram is drawn again, but this time in combination with a distribution corresponding to a local minimum of the min-log-likelihood function. This distribution was obtained using a local minimizer.

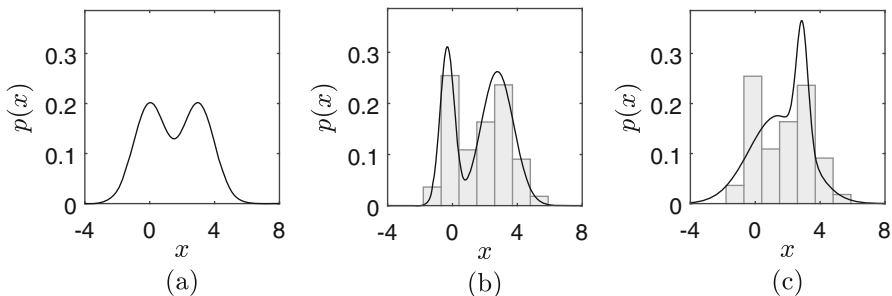


Fig. 1 Example of a local minimum. Panel (a) depicts the probability density function $p(x)$ of the finite mixture model that was used to simulate data. Panel (b) depicts the histogram of the 50 simulated data points $\{y_i\}_{i=1,\dots,50}$. The probability density function corresponding to the global optimum of the min-log-likelihood function is drawn as well. Panel (c) again depicts the histogram of the 50 simulated data points. Here, the probability density function corresponding to a local optimum of the min-log-likelihood function is shown

Because of the occurrence of local minima and the low computational cost for optimization, the univariate finite mixture model provides a suitable application to demonstrate the synergized bootstrap method.

4.1 Simulating Data

For our application, we simulated 100 different datasets \mathbf{y}^j ($j = 1, \dots, 100$) from 100 distinct two-component finite mixture distributions $p^j(x)$. Each of the datasets consisted of $n^j = 50$ data points.

4.1.1 Generating Model Parameters

The first component of each finite mixture distribution corresponded to the standard normal distribution ($\mu_1^j = 0, \sigma_1^j = 1$). The weights ω_1^j were sampled from a uniform distribution

$$\omega_1^j \sim U(.25, .75).$$

The means μ_2^j and the standard deviations σ_2^j of the second components were also sampled from uniform distributions:

$$\begin{aligned} \mu_2^j &\sim U(2, 4), \\ \sigma_2^j &\sim U(.5, 1.5). \end{aligned}$$

To ensure that the functions $p^j(x)$ were proper probability density functions, the weights ω_2^j of the second components were constructed as follows:

$$\omega_2^j = 1 - \omega_1^j.$$

4.1.2 Generating Data

To generate 50 data points from the two-components finite mixture distribution $p^j(x)$, first of all 50 numbers r_i^j ($i = 1, \dots, 50$) were randomly drawn from a uniform distribution $U(0, 1)$. For each of these r_i^j values, a normally distributed random number y_i^j was sampled either from the first component of the finite mixture $p^j(x)$ or from the second component. If $r_i^j < \omega_1^j$, then y_i^j was sampled from the first component; otherwise it was sampled from the second component.

Doing this for all 100 sampled parameters led to 100 datasets \mathbf{y}^j of which the data points y_i^j were appropriately distributed according to their corresponding two-component finite mixture distribution $p^j(x)$.

Algorithm 3: generating data

input : number of required data points and model parameters (means, standard deviations, weights)

output: data \mathbf{y} distributed according to mixture distribution

```

1 for  $i = 1$  to  $n\_data\_points$  do
2    $r = \text{uniform\_random\_number}(0,1);$ 
3   if  $r < \text{weights}(1)$  then
4      $y(i) = \text{mean}(1) + \text{standard\_deviation}(1) * \text{normal\_random\_number}();$ 
5   else
6      $y(i) = \text{mean}(2) + \text{standard\_deviation}(2) * \text{normal\_random\_number}();$ 
7   end
8 end
```

4.2 Simulation Study

4.2.1 Non-parametric Bootstrapping

For each of the 100 datasets \mathbf{y}^j , 1,000 bootstrap replicates \mathbf{y}^{jk} ($k = 1, \dots, 1000$) were created by drawing 50 data points (exactly as many as there are in the original datasets) with replacement from the original datasets \mathbf{y}^j . In doing so, 100,000 datasets were constructed in total.

Algorithm 4: non-parametric bootstrapping

```

input : data
output: re-sampled dataset y

1 for  $i = 1$  to  $n\_data\_points$  do
2    $r = \text{uniform\_random\_integer}(1, n\_data\_points);$ 
3    $y(i) = \text{data}(r);$ 
4 end
  
```

4.2.2 Optimization

Associated with every replicated dataset \mathbf{y}^{jk} , there is a min-log-likelihood function $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\omega} | \mathbf{y}^{jk})$. The parameter estimates that are obtained by minimizing the 1,000 min-log-likelihood functions \mathcal{L}^{jk} corresponding to a specific original dataset \mathbf{y}^j can be used to construct bootstrap confidence intervals around the parameter estimates $(\hat{\boldsymbol{\mu}}^j, \hat{\boldsymbol{\sigma}}^j, \hat{\boldsymbol{\omega}}^j)$ obtained by minimizing the min-log-likelihood function $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\omega} | \mathbf{y}^j)$. For our simulation study, the 1,000 min-log-likelihood functions \mathcal{L}^{jk} corresponding to a specific original dataset \mathbf{y}^j were minimized in parallel. In what follows, an optimization (procedure) will refer to the minimization of the 1,000 min-log-likelihood functions corresponding to a single original dataset \mathbf{y}^j .

All optimizations were done using the standard DE optimizer as well as the synergized DE method. For both, the crossover rate was set equal to .3. Moreover, they were run using the traditional *DE/rand/2/bin* configuration (see Storn and Price 1997). In other words, we opted for the random DE mutation strategy and the binomial crossover scheme. The 2 refers to the number of agents that made up the weighted differences of the mutants. The weighing factors of the weighted differences for the mutations were randomly sampled from a uniform distribution $U(0, 2)$, as was done by Mohamed et al. (2012).

Every optimization procedure was done with $NP = 100$ agents per population which is amply sufficient according to DE standards (see Storn and Price 1997). To study the consequence of significantly reducing the population size, each simulation procedure was also repeated with $NP = 10$ agents per population. Hence, in total 400 different optimization procedures were run, 4 for each original dataset \mathbf{y}^j .

The synergized DE required some additional specifications. Populations were activated if they failed to update for three consecutive iterations. Furthermore, only five populations could simultaneously be activated.

4.2.3 Accuracy and Benchmarks

There are 1,000 DE populations at play during a single optimization procedure. The accuracy of a DE population is defined as the difference between its best agent's min-log-likelihood (the smallest value in the population) and the min-log-likelihood of the global minimum of the function that is being minimized by the agents of

this population. The bootstrap accuracy is the average of all 1,000 accuracies in the optimization. Both the accuracy and the bootstrap accuracy are functions of the number of DE iterations; they become smaller the more the populations evolve toward their global minimum.

In order to be able to determine the accuracies, the global minima of the min-log-likelihood functions have to be known prior to doing the optimizations. We determined these by taking the best of three independent standard DE optimizations for each of the 100,000 min-log-likelihood functions in the simulation study. The setup of these DE optimizers was the same as described above. The number of agents NP was chosen to be 100 and the maximum number of iterations was set equal to 3,000.

4.2.4 Comparing Methods

Every optimization procedure was repeated four times: the standard DE algorithm was run with $NP = 100$ and $NP = 10$ agents per population and so was the synergized DE algorithm. For each of them, the bootstrap accuracy was tracked as a function of the number of cost function evaluations that were required to attain this specific accuracy. At every instance in the evolution, the accuracy of the different procedures could be compared. In all cases, the optimizations ran until a satisfactory accuracy was reached.

5 Results

The results of the simulation study are summarized in Fig. 2 which depicts the bootstrap accuracies as a function of the number of executed cost function evaluations. Red lines correspond to the optimizations with the standard DE method and blue lines correspond to the optimizations with the synergized method. Lighter and darker lines correspond to populations with $NP = 100$ and $NP = 10$ agents, respectively. The standard DE bootstraps ran for 4,000 iterations; the synergized bootstraps ran for 2,000 iterations.

Looking first at the red lines, we can see that reducing NP is not favorable for the standard DE method for this particular problem. The cost of updating an entire population is smaller for the darker lines, and initially the darker lines converge faster, but they typically do not reach the same level of accuracy as the lighter lines. It is known that smaller populations can evolve more rapidly, but for non-convex cost functions with many local minima, this increases the risk of getting stuck in local minima because the spread of the agents in the population shrinks too quickly. In the remainder of this discussion, we will only compare to the standard DE with $NP = 100$ (light red).

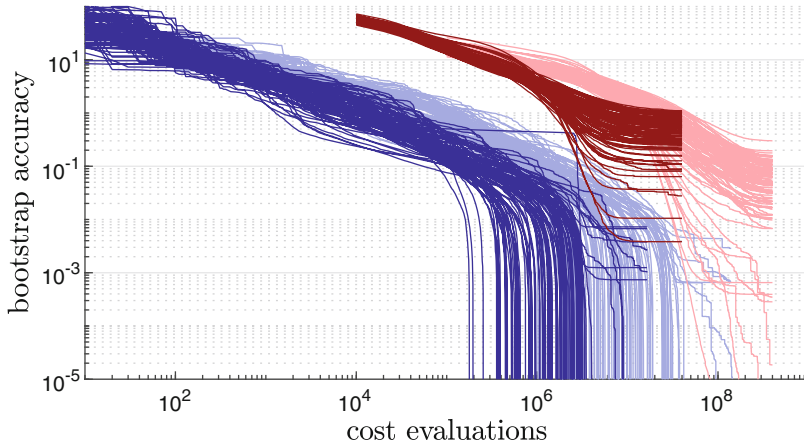


Fig. 2 Comparison of the standard DE bootstrap and the synergized bootstrap. The bootstrap accuracy is depicted in function of the number of cost function evaluations for the standard DE method (red) and the synergized method (blue). The bootstrap accuracy refers to the average difference between the min-log-likelihoods of the best agents of the populations and those of the corresponding global minima. Lighter lines correspond to populations of $NP = 100$ agents and darker lines to populations of $NP = 10$ agents

The synergized method converges a lot faster than the standard DE method, regardless of the population size. The lines corresponding to synergized optimizations start earlier, because less cost function evaluations are required for initialization and also during the global search stage of the optimization process (not yet all populations are actively evolving). For the red lines, $B \times NP = 10^5$ evaluations are needed to initialize the optimization procedure, which determines the starting points of the lines. Moreover, for the standard DE every update of a generation requires 10^5 evaluations. The light blue lines ($NP = 100$) reach an accuracy of about 10^{-1} approximately 100 times faster than the red lines. The dark blue lines ($NP = 10$) reach this accuracy approximately 500 times faster.

Not only do the synergized bootstraps converge faster, they are more accurate than the standard method. The bootstrap accuracy of the DE method only drops to about 10^{-1} to $2 \cdot 10^{-2}$, indicating that for each simulation several of the bootstrap populations get stuck in local minima. For the synergized method, the accuracy generally drops below 10^{-5} . Reducing the number of agents per population does not affect this accuracy for the synergized method (unlike for the standard DE). The fact that populations that got stuck in local minima can still be corrected by others seems to have an advantageous effect on the global performance of the optimizer.

In summary, the synergized method is both faster and more accurate than the standard DE method. It also requires less agents per population to work, which reduces the computation time even more.

6 Discussion

In this paper, we introduced a new method for speeding up the optimization procedures in the context of re-sampling methods, and we applied the method to the non-parametric bootstrapping of the parameters of a univariate two-component mixture model. Just like the fingerprint method (Mestdagh et al. 2015), the proposed synergized bootstrap method exploits the fact that the characteristics of the re-sampled datasets are very similar, and so are the related optimization problems. Moreover, it relies on the fact that, for a large class of cost functions, the costs of a parameter set for the different cost functions can be obtained at a very low computing cost. This allows combining resources across the different optimization procedures during run time without extra burden. By letting the different optimization processes communicate with one another, the computation time can significantly be reduced since the optimization problems are very similar.

For the application on the mixture models, the results indicated that the synergized differential evolution (DE) can be approximately 100 to 500 times faster than the standard DE. Furthermore, because the synergized DE can correct populations that got stuck in local minima by relying on information from the other populations, it can generally achieve a better accuracy.

The synergized bootstrap method is more widely applicable than the fingerprint method because it is not limited to smooth, convex cost functions. A limitation of the synergized method is, however, the need for cost functions of which the costs of the different data points can be computed independently. This class of functions is nonetheless very large, and these functions are frequently encountered in practical applications because independence of data points is a common assumption. Although this requirement is sufficient for the procedure to work, it is not strictly necessary. Another case where the synergized method could be used is for optimization routines where there is separability of the initial model calculations and the final comparison to data in the objective function. If, for instance, the entire probability density function is calculated for an agent (e.g., as a sufficiently fine discrete distribution) independent of the data, it can be recycled for use with other (re-sampled) datasets (and hence for other cost functions) for practically no extra cost – here the bottleneck is the preparatory calculation of the probability density function, not its evaluation. Some examples of algorithms that first calculate/simulate the entire probability density function and then in a second separable step compare this with the data are D*M optimization for diffusion models of choice response time (Verdonck and Tuerlinckx 2016) and simulated likelihood approaches (Verdonck and Tuerlinckx 2014).

References

- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), 248–249.
- Andrews, D. W. K. (1999). *Higher-order improvements of a computationally attractive-step bootstrap for extremum estimators* (Tech. Rep. No. 1230). Cowles Foundation for Research in Economics, Yale University.
- Boonthiem, S., Boonta, S., & Klongdee, W. (2017). A differential evolution algorithm with adaptive controlling weighted parameter for finite mixture model of some fire insurance data in Thailand. *SNRU Journal of Science and Technology*, 9, 491–501.
- Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., et al. (2013). A network approach to psychopathology: new insights into clinical longitudinal data. *PLOS ONE*, 8(4), e60188.
- Cawley, G. C., & Talbot, N. L. C. (2008). Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*, 71(2–3), 243–264.
- Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12), 1367–1372.
- Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 66(1), 165–185.
- Davidson, R., & MacKinnon, J. G. (1999). Bootstrap testing in nonlinear models. *International Economic Review*, 40(2), 487–508.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185.
- Efron, B. (1990). More efficient bootstrap computations. *Journal of the American Statistical Association*, 85(409), 79–89.
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton: CRC Press. (Google-Books-ID: gLlpIUxRntoC).
- Good, P. I. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. New York: Springer.
- Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models: The R Package pbkrtest. *Journal of Statistical Software*, 59(9), 1–32.
- Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.
- Hu, F., & Kalbfleisch, J. D. (2000). The estimating function bootstrap. *Canadian Journal of Statistics*, 28(3), 449–499.
- Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2011). A scalable bootstrap for massive data. arXiv:1112.5016 [stat], (arXiv: 1112.5016).
- Kwedlo, W. (2014). Estimation of parameters of Gaussian mixture models by a hybrid method combining a self-adaptive differential evolution with the EM Algorithm. *Advances in Computer Science Research*, 11, 109–123.
- Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10), 833.
- Maho, Y. L., Whittington, J. D., Hanuise, N., Pereira, L., Boureau, M., Brucker, M., et al. (2014). Rovers minimize human disturbance in research on wild animals. *Nature Methods*, 11(12), 1242.
- McLachlan, G. & Peel, D. (2000). *Finite mixture models* (1 ed.). New York: Wiley-Interscience.
- Mestdagh, M., Verdonck, S., Duisters, K., & Tuerlinckx, F. (2015). Fingerprint resampling: A generic method for efficient resampling. *Scientific Reports*, 5, 16970.

- Mestdagh, M., Verdonck, S., Meers, K., Loossens, T., & Tuerlinckx, F. (2018). Prepaid parameter estimation without likelihoods. arXiv:1812.09799 [stat]. (arXiv: 1812.09799).
- Mohamed, A. W., Sabry, H. Z., & Khorshid, M. (2012). An alternative differential evolution algorithm for global optimization. *Journal of Advanced Research*, 3(2), 149–165.
- Persson, F., Lindén, M., Unoson, C., & Elf, J. (2013). Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nature Methods*, 10(3), 265.
- Ramaswamy, S., Ross, K. N., Lander, E. S., & Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, 33(1), 49–54.
- Samuh, M. H., Grilli, L., Rampichini, C., Salmaso, L., & Lunardon, N. (2012). The use of permutation tests for variance components in linear mixed models. *Communications in Statistics – Theory and Methods*, 41(16–17), 3020–3029.
- Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., et al. (2006). Intellectual ability and cortical development in children and adolescents. *Nature*, 440(7084), 676.
- Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Systematic Biology*, 57(5), 758–771.
- Storn, R. & Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4), 341–359.
- Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228), 480.
- Verdonck, S., & Tuerlinckx, F. (2014). The Ising decision maker: A binary stochastic network for choice response time. *Psychological Review*, 121(3), 422–462.
- Verdonck, S., & Tuerlinckx, F. (2016). Factoring out nondecision time in choice reaction time data: Theory and implications. *Psychological Review*, 123(2), 208–218.
- Zeng, D., & Lin, D. Y. (2008). Efficient resampling methods for nonsmooth estimating functions. *Biostatistics (Oxford, England)*, 9(2), 355–363.
- Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7), 821–824.
- Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4), 407.

Synchronized Time Profile Similarity in Applications to Nearest Neighbor Classification



Qimin Liu

Abstract One of the existing approaches to time series classification exploits the time profiles using the original data with synchronization instead of model-implied data. Synchronization aligns inter-individual data from different time points to account for potential phase offsets and nonstationarity in the data. Such synchronization has been applied in psychology: For example, coordinated motion between two individuals exchanging information was used as a predictor and outcome of psychological processes. Synchronization also affords better classification outcomes, as discussed in the data mining community, through aligning the data to reveal the maximally shared profile underlying two compared data sequences. For inter-individual comparison of univariate time series data, existing similarity indices include Euclidean distances and squared correlations. For synchronization, we introduce dynamic time warping and window-crossed lagging. The current study compares the Euclidean distance and the squared correlation before and after synchronization using window-crossed lagging and dynamic time warping in applications to one-nearest-neighbor classification tasks. Discussion, limitations, and future directions are provided.

Keywords Classification · Time series · Dynamic time warping

1 Introduction

Advancement in and accessibility to technology have enabled psychologists to harvest big data at greater complexity and larger scale. Intensive longitudinal data are often the product of many repeated measurements through approaches such as daily diary, experience sampling, and burst measurements. The pervasive use of smartphones, fitness trackers, and the Internet of Things has made the use of

Q. Liu (✉)
Vanderbilt University, Nashville, TN, USA
e-mail: qimin.liu@vanderbilt.edu

intensive longitudinal data increasingly accessible. The latitude of information in intensive longitudinal data can enable data-driven investigation of novel research questions (Harlow and Oswald 2016). In this paper, we examine classification of time-invariant membership based on individual trajectories across time.

More attention is due on the use of intensive longitudinal data from a big data perspective. Intensive longitudinal data are no stranger to psychological researchers. Daily diary data help developmental psychologists to explore the process of aging (Birditt et al. 2005). Clinical psychologists utilize daily data of multiple individuals to explore mental health problems (Laurenceau et al. 2005). Brain activity measures, such as EEG and fMRI methods, that are frequently employed by cognitive neuroscientists, result in data that are collected at high frequency (Kounios and Beeman 2009). In addition, organizational and industrial psychologists make use of daily diary data as well to study psychological issues at workplace (Conway and Briner 2002). As data of intense repeated measures have already shown great promise in the field of psychology, application of classification and prediction methods may afford researchers and practitioners additional viewpoints and practical utility.

One novel research problem that psychologists may be interested in is time series classification, i.e., to assign time-invariant class membership to test individuals given training time series data of multiple individuals with their respective time-invariant class membership and the time series data of the test individual. For example, cognitive neuroscientists attempted to use the brain activity data to predict the success of word recall tasks (Ezzyat et al. 2017). In particular, Ezzyat and colleagues used time series classification in hope of uncovering the mechanism of episodic memory performance with respect to brain activities and also to help decide whether intracranial brain stimulation should be induced to improve memory performance. As shown in the example, the reason for psychologists to explore the time series classification problem can be twofold. On the one hand, classification, to some extent, extracts the signals behind the noise from the data and thus can represent the useful information in the data. For example, consider a potential personality theory where personality categorization is based upon time profile. To validate such theory, time series classification can be applied where the classification accuracy may serve to explore the utility and the practicality of such theory. On the other hand, the classification problem can readily translate academic research to real-world applications. For example, clinical psychologists may utilize routine outcome data from behavioral care to aid in diagnosis or in decisions of posttreatment care regime. While approaches for time series classification exist, such as through autoregressive logistic or multinomial models, these conventional methods have limitations.

One limitation in commonly used conventional modeling of intensive longitudinal data, e.g., autoregressive moving average models (Box et al. 2008), lies in its various assumptions. One such assumption is stationarity. Admittedly, stationarity assumption is beneficial both in computation and interpretation: The collected times series information is assumed to be representative of the entire behavior, and thus relevant statistics can be meaningfully obtained. In other words, for each person, any interval of time is assumed to be representative for the intraindividual changes. To

this end, data preprocessing techniques (West and Hepworth 1991) exist in shaping the data to better fit the stationarity assumption. However, such techniques may mask the important substantive information that the nonstationarity in behavioral time series data would convey. For example, nonstationarity can imply nonconstant dynamics of within-person changes (Boker et al. 2016). Moreover, given the big data nature in intensive longitudinal data, the enormous amount of measurement occasions may render the stationarity assumption for the whole series unlikely. Therefore, innovative methods that do not require such assumptions may glean the information that conventional methods often omit due to their assumptions.

While the big data aspect of intensive longitudinal data may be new, psychologists have been interested in class membership in relevance to individual profiles across time. Admittedly, meager efforts have been paid to developing methods for time series classification particular to social and behavioral science: A brief search in selected methodological journals for psychology – *Structural Equation Modeling*, *Psychological Methods*, *Psychometrika*, *Multivariate Behavioral Research*, *Behavioral Research Methods* – with keyword “time series classification” returned no exact match. However, quantitative psychologists have made relevant discussions. For example, Gates et al. (2017) discussed community detection within group iterative multiple model estimation for clustering time series data. The goal was to simultaneously detect homogeneous subgroups and to classify individuals to identified groups. This is often referred to as a clustering problem that is “unsupervised” as it does not require a training data with assigned class membership information (Gates et al. 2017). The focus of this paper, however, is on classification problems that are supervised: The training data contain both time series information and the time-invariant class membership labels; the goal is to predict the membership for new or test data given time series information. In other words, provided with time profile and their subgroup labels, we aim to “learn” the difference between predefined subgroups and apply this to classify individuals into predefined subgroups.

Interindividual differences in intraindividual time profiles are central to discerning between predefined classes. For classification algorithms to exploit the interindividual differences of the training data and to discover the pattern to apply to the test data, interindividual similarity measures of time series data can be highly relevant. The intricacy in time series data asks methodological researchers not only to consider simple similarity indices but also to take into account of the potential nonsynchrony between time profiles. That is, two time profiles can look different just because the processes are not in sync or the individuals in comparison are not aligned at their respective phases. For example, in psychotherapy process research, individuals may be studying different cognitive behavioral skills modules. The daily diary data between individuals may show different patterns despite the fact that they would be the same if the individuals were undergoing the same module.

The alignment and synchrony of time series data between individuals can convey substantive meaning to psychologists. For example, interpersonal communication can often come with certain levels of coordinated behaviors, such as nonverbal synchrony. Nonverbal synchrony can be that where eye gaze or facial expressions

are in sync between individuals. For example, lack of eye contact in communications can be seen as lack of nonverbal synchrony and can be interpreted as insincere (Ramseyer and Tschacher 2006). Nonverbal synchrony has been employed to investigate association between nonverbal behaviors and psychotherapeutic outcomes (Ramseyer and Tschacher 2011). Although the idea of nonsynchrony seemed to have only applied to study dyadic time series data, the idea should transcend into studying the inter-individual similarity of time profiles in general. Furthermore, it is worth noting that quantitative psychologists have proposed windowed cross-correlation (WCC) to measure such nonsynchrony, which breaks the data into segments of “window” and computes a correlation matrix across windows (Boker et al. 2002). In fact, phase misalignment or nonsynchrony has also been entertained in the machine learning community (Jeong et al. 2011).

Machine learning community has offered a great amount of discussion in time series classification in general with over 100 methods before 2003 and even more dedicated efforts in recent years (Bagnall et al. 2017). One benchmark method, dynamic time warping (DTW) one nearest neighbor (1NN), pays special attention to phase alignment before computing similarities between individuals (Sakoe and Chiba 1978). In short, dynamic time warping involves maximally aligning the two times series prior to comparison. Moreover, dynamic time warping, combined with nearest neighbor classification, has shown success for real-world applications such as in human physical activity detection (Sempena et al. 2011). However, the performance of DTW 1NN in social and behavioral time series classification tasks, to the author’s knowledge, has not been investigated.

The present paper aims to continue exploring the big data aspect of intensive longitudinal data in social and behavioral sciences. In particular, we provide discussions on supervised time series classification. Since the stationarity assumption may not hold for intensive longitudinal data, we limit most of our discussions to methods that do not assume stationarity. Moreover, given the potential substantive meaning of phase nonsynchrony and misalignment in social and behavioral sciences, we only discuss techniques that include such consideration. The rest of the paper is organized as follows. First, we review two general inter-individual similarity measures, Euclidian distance and squared correlation, as the former is used in DTW and the latter in WCC. Then, we introduce two techniques for phase alignment, DTW and windowed cross lagging (which is derived from WCC). After that, we provide details on pairing phase alignment methods with similarity measures in applications to 1NN for time series classification. Furthermore, we provide a simulation study to show the potential of discussed methods. The paper aims to motivate psychologists to consider time series classification problems. In addition, since dynamic time warping has mostly been applied only to Euclidean distances and windowed cross lagging only to correlations, the present paper also innovates by disaggregating phase alignment methods to similarity indices and introduces some example indices via mixing-and-matching existing techniques. Lastly, we hope to promote the use of dynamic time warping to psychologists and windowed cross lagging to the machine learning community.

2 Methods

2.1 Similarity Measures

Euclidean distance, d , between two points x_i and x_j is defined as the length of the shortest straight-line segment that connects the two points, $d = |x_i - x_j|$. The Euclidean distance between two vectors $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(T)})$ and $X_j = (x_j^{(1)}, x_j^{(2)}, \dots, x_j^{(T)})$ is subsequently defined as $d(X_i, X_j) =$

$$\sqrt{(x_i^{(1)} - x_j^{(1)})^2 + (x_i^{(2)} - x_j^{(2)})^2 + \dots + (x_i^{(T)} - x_j^{(T)})^2} = \sqrt{\sum_{t=1}^T (x_i^{(t)} - x_j^{(t)})^2}.$$

Here, i and j denote individuals i and j , T is the total time points, and t represents specific time points. It is worth noting that, first, to calculate Euclidean distance between two vectors, two vectors should share the same length; second, the Euclidean distance of two vectors sums the distance between concurrent points before the square root. Euclidean distance can be used to measure data points of two individuals or data sequences of two individuals. The magnitude of Euclidean distance is inversely related with the similarity between two individuals.

Squared correlation, r^2 , between two vectors $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(T)})$ and $X_j = (x_j^{(1)}, x_j^{(2)}, \dots, x_j^{(T)})$ can be given by $r^2(X_i, X_j) = \left(\frac{\sum (x_i^{(t)} - \bar{x}_i)(x_j^{(t)} - \bar{x}_j)}{T \times s_{x_i} s_{x_j}} \right)^2$. The squared correlation of data sequences between individuals can represent the similarity. Similar to Euclidean distance, the above formula requires two data sequences to be of equal length. While Euclidean distance varies with the absolute value of the data, squared correlation is invariant to linear transformation.

2.2 Phase Alignment: Dynamic Time Warping

The goal is to compare two time-dependent sequences $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(T)})$ of length $T_i \in \mathbb{N}$ and $X_j = (x_j^{(1)}, x_j^{(2)}, \dots, x_j^{(T)})$ of length $T_j \in \mathbb{N}$. Graphically, for two sequences to appear similar in shape and thus phase-aligned, each pair of concurrent points should only show minimal differences in Euclidean distance, which would in turn make two sequences “close” to each other visually. Euclidean distance helps measure the visual proximity between two points because the Euclidean distance is the shortest straight-line distance. Dynamic time warping first originates from speech recognition literature (Sakoe and Chiba 1978). In speech recognition, the speech data of the same words may appear different for different utterance, for example, the pace of how one talks or the length of pause between words. Thus, the speech data sequence of two individuals may appear out of phase

due to the difference in pause, and the length of the phase for two individuals may also differ owing to the difference in talking speed. Dynamic time warping attempts to adjust for such phase misalignment by “stretching” or “shrinking” each phase so that two data sequences are maximally in sync in a sense that two data sequences would appear similar graphically.

Dynamic time warping thus attempts to align each point in one sequence to each closest point in the other sequence under some constraints. Let $d(m, n) := d(x_i^m, x_j^n)$ denote the Euclidean distance between a pair of elements in X_i and X_j . A (T_i, T_j) -warping path between X_i and X_j , $p = (p_1, \dots, p_L)$ where $p_l = (m_l, n_l) \in [1 : T_i] \times [1 : T_j]$ for $l \in [1, L]$. That is, each element in the warping path, p , records the pairing of an element in X_i to in X_j . The warping path satisfies the following conditions:

1. Boundary condition: $p_1 = (1, 1)$ and $p_L = (T_i, T_j)$
2. Monotonicity condition: $n_1 \leq n_2 \leq \dots \leq n_L$ and $m_1 \leq m_2 \leq \dots \leq m_L$, i.e., $n_t \leq n_{t'}$ and $m_t \leq m_{t'} \forall t, t' \in [1, L] \cap \mathbb{N}$
3. Step size condition: $p_{l+1} - p_l \in \{(1, 0), (0, 1), (1, 1)\}$ for $l \in [1, L - 1] \cap \mathbb{N}$

The first condition requires that the first and the last element in X_i and in X_j are always respectively matched to each other so that all elements of two sequences are aligned. The second condition helps the alignment proceed forward. The third condition further specifies the alignment process and requires that no elements can be omitted and that no path step can involve the same pair.

After alignment, X_i and X_j are augmented into $X_i^* = (x_i^1, x_i^{m_i}, \dots, x_i^{T_i})$ and $X_j^* = (x_j^1, x_j^{n_j}, \dots, x_j^{T_j})$. The distance between X_i^* and X_j^* can be defined as $\Delta(X_i^*, X_j^*) = \sum_{l=1}^L d(m_l, n_l)$. An optimal warping path is that which $\Delta(X_i^*, X_j^*)$ is at its minimum, that is, $DTW(X_i^*, X_j^*) = \text{Argmin} \Delta(X_i^*, X_j^*)$ given that p satisfies the conditions of a warping path. Let $\Delta^*(m, n)$ denote the optimized DTW distance between X_i^* and X_j^* up to point the m th and the n th element, respectively, and the optimization can be realized in the following algorithm:

```

Result:  $p^* = (p_1, \dots, p_L)$ 
initialization:  $\Delta^*(1, 1) = d(1, 1)$ ;
while  $L < T_i$  or  $L < T_j$  do
  |  $\Delta^*(m, n) = \min\{\Delta^*(m, n-1) + d(m, n), \Delta^*(m-1, n-1)$ 
  |    $+d(m, n), \Delta^*(m-1, n) + d(m, n)\}$ ;
end

```

The algorithm effectively yields the optimal path $p^* = (p_1, \dots, p_L)$ from the reverse order of the indices starting with $p_L = (T_1, T_2)$. Let the previous step l be known, $p_l = (m_l, n_l)$. Note that the optimal path is defined backward: The initial step is p_L , which is the alignment of last elements in both sequences. The optimal next step $l-1$ is $p_{l-1} = \text{Argmin}\{d(m_{l-1}, n_{l-1}), d(m_{l-1}, n_l), d(m_l, n_{l-1})\}$. In the case where $m_l = 1$, $p_{l-1} = (1, n_{l-1})$. Similarly, $p_{l-1} = (m_{l-1}, 1)$ if $n_l = 1$. The optimal path is defined when l reaches 1, that is, $p_1 = (1, 1)$. It is worth noting that the minimum value may not be unique. That is, multiple optimal warping paths may

exist. However, while the specific paths differ, the resulting distance $DTW(X_i^*, X_j^*)$ would be the same for all optimal warping paths.

We illustrate the algorithm with a numerical example. Consider

$$\begin{aligned} X_1 &= (1, .77, .17, -.5, -.94, -.94, -.5, .17, .76, 1) \\ X_2 &= (.33, .68, 1.04, .96, .36, -.25, -.77, -.92, -.58, 0) \end{aligned}$$

The raw sequences are plotted in Fig. 1. To find an optimal warping path p^* , we start with $p_L = (10, 10) = d(10, 10) = 1$. Thus, the next step is

$$\begin{aligned} p_{L-1} &= \text{Argmin}\{d(10, 9), d(9, 10), d(9, 9)\} \\ &= \text{Argmin}\{|1 - (-.58)|, |.76 - 0|, |.76 - (-.58)|\} \\ &= \text{Argmin}\{1.58, .76, 1.34\} \\ &= (9, 10). \end{aligned}$$

A similar procedure can be repeated till the optimal warping path p^* is fully defined. One potential optimal path is $p^* = \{(10,10), (10,9), (10,8), (9,7), (8,6), (8,5), (7,5), (6,4), (5,3), (4,2), (3,1), (2,1), (1,1)\}$. The DTW aligned sequences are also plotted in Fig. 1. As shown in Fig. 1, the DTW results in phase alignment that reveals a maximally shared and synced graphical pattern. Subsequently, to compare two sequences, the Euclidean distance, for example, can be calculated between DTW aligned X_1^* and X_2^* given p^* . DTW can be readily implemented in R using the package `dtw` (Giorgino 2009). Given sequences x and y , `dtw(x, y)` computes Euclidean distance after dynamic time warping. To extract the distance, we can use `dtw(x, y)$normalizedDistance`. Let `align=dtw(x, y)`; we can use `cor(x[align$index1], y[align$index2])^2` to compute squared correlation with dynamic time warping alignment.

2.3 Phase Alignment: Windowed Cross Lagging

Windowed cross lagging (Boker et al. 2002) assumes stationarity in short durations of the time series. By breaking time-dependent data sequences into smaller “windows,” stationarity assumption is more likely to be met than for the entire sequence. Thus, this may prove particularly beneficial for intensive longitudinal data. Given stationarity, any collection of measurement occasions within a window is representative of the window. Moreover, all occasions of measurement within a window share an underlying expected value that does not vary across occasions within the window. As such, common statistics such as means and variances can be meaningfully computed. In particular, a window, W_{x_i} , is defined as a sequential measurements sampled from the time series X_i . For a data sequence of length T and windows of size T_W , we can identify $T - T_W + 1$ windows.

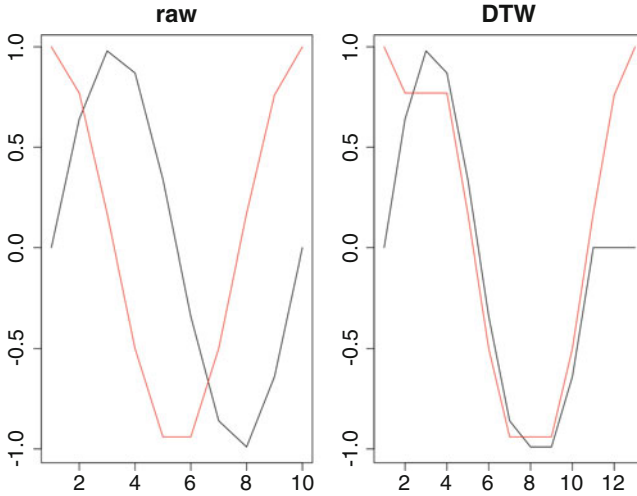


Fig. 1 Illustrative example of DTW. The left plot shows data sequences before alignment and the right plot shows data sequences after alignment. The red line represents X_1 and the black line represents X_2 . The X-axis is the index of the number. Note that DTW “stretches” the sequence, resulting in greater maximum index. The Y-axis represents the value of the datum

The window size can be theoretically determined. For example, a 13-week window, the median mood cycle (Solomon et al. 2010), can be chosen so that the windowing is informative given its substantive underpinning. This can be both advantageous and disadvantageous: The theoretical basis can make the interpretation of statistics from windowing more meaningful; however, the strength of the theory may also relate to the quality of similarity metrics from windowing. Windowing let researchers specify the length of the “phase” given theoretical reasons. It is worth noting the specified “phases” are of the same length and at times can be at odds with theoretical considerations.

After obtaining the windows from data sequences in comparison, instead of one single comparison to be made in other regular comparison like in Euclidean distance, a matrix of similarity metrics can be obtained through cross-comparison. Given a desired lag range of size v , the a th window in the first sequence can be compared to $(a-v)$ th, $(a-v+1)$ th, . . . $(a+v)$ th window of the other sequence. Each window can be compared to $2v + 1$ windows in the other sequence. Specifically, for example, a $v = 1$ cross-lagged comparison would be comprised of three comparisons: the comparison between the a th window of the first sequence to the $a - 1$ th window of the second sequence, the comparison between the a th windows of both sequences, and the comparison between the $a + 1$ th window of the first sequence and the a th window of the second sequence. The original paper used correlation as the metric (Boker et al. 2002) where each element of the matrix is the

Table 1 Left: example windowing with $T_w = 8$; right: WCC with $v = 1$ and $T_w = 8$

	X_1	X_2	lag		
Window 1	1, .77, .17, -.5, -.94, -.94, -.5, .17	.33, .68, 1.04, .96, .36, -.25, -.77, -.92	-1	0	1
Window 2	.77, .17, -.5, -.94, -.94, -.5, .17, .76	.68, 1.04, .96, .36, -.25, -.77, -.92, -.58	-.59	.00	.69
Window 3	.17, -.5, -.94, -.94, -.5, .17, .76, 1	1.04, .96, .36, -.25, -.77, -.92, -.58, 0	-.69	-.14	.59

correlation of respectively lagged windows. The correlation between two windows can be expressed as $r(W_{x_i}, W_{x_j}) = \frac{1}{T_w} \sum_{t=1}^{T_w} \frac{(W_{x_i}^{(t)} - \bar{W}_{x_i})(W_{x_j}^{(t)} - \bar{W}_{x_j})}{s_{W_{x_i}} s_{W_{x_j}}}$

We illustrate the windowed cross lagging procedure in terms of the windowed cross correlation. Recall prior example of two data sequences X_1 and X_2 . We consider a lag range of size 1 and window size of 8 for the sake of convenience. First, the windowing step resulted in three windows for both data sequences. The specific windows are displayed in Table 1 (left). Then, the cross lagging procedure resulted in a 3×2 matrix as shown in Table 1 (right).

Since windowed cross lagging results in a matrix of similarity metrics, summarizing the matrix into a single index may be preferable especially in applications to nearest neighbor classifier. A number of strategies can be considered. One strategy is to compute the average of all matrix elements. Such averages can represent the average similarity between two sequences across phase alignments. Another is to pick meaningful elements from the matrix. For example, with WCC, one may desire to choose the maximum squared correlation across all elements. The maximum squared correlation may represent the “strongest signal” across lags and between phases. Similarly, if Euclidean distance is coupled with windowed cross lagging, then the minimum may be chosen as small Euclidean distance indicates high similarity. The following R scripts implement aforementioned indices:

```
wcc<-function(x,y,win.size,lag.max){
  temp=sapply(1:(length(x)-win.size+1)-lag.max, function(r) {
    sapply(1:(lag.max*2+1), function(c) {
      s1=sapply(1:(length(x)-win.size+1),
        function(a){x[a:(a+win.size-1)]})
      [,ifelse((r+c-1)<r+lag.max,r+c-1,lag.max+r)]
      s2=sapply(1:(length(x)-win.size+1),
        function(a){y[a:(a+win.size-1)]})
      [,ifelse(c<lag.max+1,lag.max+r,lag.max+r-(c-lag.max-1))]
      return(c(ifelse((sd(s1)==0|sd(s2)==0),0,cor(s1,s2)^2),
        sum((s1-s2)^2)))
    })
  })
  temp=matrix(as.numeric(temp),ncol=2,byrow=T)
  return(c(max(temp[1,]), # maximum window-crossed correlation
    min(temp[2,]), # minimum window-crossed distance
    mean(temp[1,]), # average window-crossed correlation
    mean(temp[2,]))) # average window-crossed distance
}
```

2.4 Nearest Neighbor Classification

Given a similarity measure of two time-dependent sequences that handles phase misalignment, nearest neighbor classification can be readily used. Consider a training dataset of size n , $S = \{X_1, X_2, \dots, X_n\}$ where any element in S is a time-dependent sequence measured across T occasions, $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(T)})$. Moreover, for each X_i , we are also given the class membership information Y_i . Now we wish to determine the class membership of a test data sequence X_j . The logic behind nearest neighbor (NN) classification is that observations of the same class ought to be similar to each other. In the context of time series classification, we expect time-dependent sequences of the same class to show similar patterns after adjusting for phase misalignment. The algorithm of NN classification can be summarized as follows:

1. Compute the distance δ between X_j and each element in S , i.e., X_i for all i in $[1, n] \cap \mathbb{N}$. That is, we obtain $D_j = \{\delta(X_j, X_1), \delta(X_j, X_2), \dots, \delta(X_j, X_n)\}$ where δ can be any distance measure such as Euclidean distance, squared correlation, DTW distance, DTW-aligned squared correlation, mean WCC, or mean windowed cross-lagged distance (WCD).
2. Rank the vector of distance, D_j , in order to identify the nearest neighbor(s). Denote the ranked distance vector as D_j^* . Based on the choice of δ , the k nearest neighbor(s) are those with k largest or smallest δ . We denote N as the set of indices of these nearest neighbors. For example, because small Euclidean distance indicates high similarity, the k nearest neighbor would be the k X_i s with smallest δ .
3. Assign the most frequently counted class membership among the identified nearest neighbors as the predicted class membership for X_j . That is, let F denote the frequencies for each class within $\{Y_i\}$ with $i \in N$, $\hat{Y}_j = \mathop{\text{argmax}} F$.

In the context of time series classification, 1NN is especially popular. 1NN is a nearest neighbor classifier where only the single nearest neighbor is identified and the class membership of the new observation is assigned as the class membership of the single nearest neighbor. Such popularity is not unjustified. On the one hand, the computation of the similarity index between time-dependent sequences is costly. For example, the computational cost of DTW is $O(T_1 T_2)$ and of WCC is $O((2v + 1)(T - T_w + 1))$. Using kNN with k from cross-validation would further the computational cost and render the classification task resource-, time-, and computation-consuming. On the other hand, it has been shown that cross-validated kNN, at least with DTW and with Euclidean distances, did not show significant improvement over 1NN (Bagnall and Lines 2014). Admittedly, it hasn't been tested if cross-validated kNN with windowed cross-lagged metrics would perform significantly better than 1NN. For comparability, we only consider 1NN in the present paper for all distance metrics.

To implement 1NN in R, let `dist` be a distance matrix where each column represents a case from the training dataset and each row represents a case from

the test dataset. Let y be the labels from the training dataset. $\text{dist}[i, j]$ is the distance between i th observation in the test dataset and j th observation in the training dataset. To determine the label for the i th observation in the test dataset, we use $y[\text{which.min}(\text{dist}[i])]$ as the predicted label.

3 Simulation Design and Data Generation

Our first simulation study was designed to show the performance of INN with “mix-and-match” similarity measures with and without phase alignment with AR-based classes. In particular, we considered two similarity measures without phase alignment: squared correlation (sqrC) and Euclidean distance (EuCD). We considered two similarity measures with DTW: DTW distance (DTWd) where Euclidean distance was calculated after DTW alignment and DTW squared correlation (DTWc). We also considered four similarity measures with windowed cross lagging: the average of all elements in the WCC matrix (WCCmean), the maximum of the WCC matrix (WCCmax), the average of all elements in windowed cross-lagged Euclidean distance matrix (WCDmean), and the minimum of the windowed cross-lagged Euclidean distance matrix (WCDmin).

In addition, we included a simple method based on an autoregressive (AR) model: We first fitted an AR(1) model to all X_i in the training dataset S and obtained the AR coefficient ρ_i s and then computed the empirical density function (edf) in each class. After that, we applied the AR(1) coefficient ρ_j of the test case X_j to the edfs for each class, obtained the probability of ρ_j being in the distribution of AR(1), and subsequently assigned the class membership for X_j as from that where the edf had the highest probability from fitting ρ_j . The AR-based method differs from INN with aforementioned similarity measures in that AR-based methods assume a parametric form and stationarity for the whole time series.

In particular, three classes were present in our simulation study: (1) random noise class, the time series data were simply random noises without any time-dependent pattern; (2) AR(1) class, the times series data in this class were generated with realistic model parameters from an AR(1) model (Wang et al. 2012); and (3) the nonstationary class where the data still showed a time-dependent pattern but do not meet the stationarity assumption. The data were generated using the model $x_i^{(t+1)} = x_i^{(t)} \rho_i Y_i + \varepsilon^{(t+1)}$ where the AR coefficient $\rho_i \sim N(.2, .2)$ and the error term $\varepsilon^{(t+1)} \sim N(0, 1)$. Moreover, $Y_i \in 0, 1, 5$, respectively, correspond to the random noise, AR(1) process, and a nonstationary class.

One variable was manipulated: the balance between training samples per class. For the balanced condition, training datasets with a sample size of 150 were generated with each class of 50 observations. For the imbalanced condition, the random noise, AR(1), and the nonstationary class, respectively, had 30, 80, and 40 observations. For both conditions, a test dataset of 30 observations with 10

observations for each class was used to evaluate the performance of different classification algorithms. Five hundred replications were conducted.

Performance was evaluated via the average overall accuracy and the average accuracy per class, which is computed via the mean of percentages of the true positives and the true negatives per class.

4 Results and Discussion

The results for the balanced condition and the imbalanced condition are summarized respectively in Table 2. The balance of training samples per class had some effect on the overall accuracy with the imbalanced condition having higher overall accuracy in general. For example, the overall accuracy for all methods was above 70% for the imbalanced condition, while only WCDmin exceeded 70% accuracy for the balanced condition. This observation encourages future study to examine the effect of balance in greater detail where different setup of imbalance should be investigated.

For the balanced condition, WCDmin showed the highest overall performance (76% accuracy), and DTWc performed the worst (54% accuracy). The result is encouraging because WCD with 1NN seems to show promise above and beyond the DTW-based measures such as DTWd, which is held as the state-of-the-art method in the machine learning community. Admittedly, only one data-generative process was examined in the simulation study. Nevertheless, the result from the simulation study invites researchers to test the performance of similarity measures with windowed cross lagging in broader contexts. With respect to particular classes, WCDmin and WCDmean performed best in identifying the random noise class (67%), DTWd for the AR(1) class (100%), and WCDmin for the nonstationary class. It appears that DTWd performed worst for both the random noise class (45%) and for the nonstationary class (1%). The performance of DTWd is surprising, and future research should study the performance of DTW-based similarity measures specific to nonstationary time profile patterns. Moreover, it is worth discussing that the AR-based method showed only mediocre performance overall (65%) even with respect to identifying the AR(1) class (81%).

For the imbalance condition, the AR-based method showed superior performance (78%) overall and in identifying the AR(1) class (100%). The second best method with regard to overall accuracy is DTWd (77%) owing to its good performance in identifying the AR(1) class (98%). All other methods showed similar performance both in overall accuracy (around 70%) and in accuracy per class. The results from the imbalance class should be generalized with caution as only one imbalance setup was considered and the setup may have shown favor toward the AR(1) class. Because more training samples are allocated to the AR(1) class, methods that are good at identifying AR(1) class can reach saturated performance. Similarly, with fewer training samples in other classes, it is possible not enough training samples exist to discriminate performance between methods.

Table 2 Simulation results. The table includes results from both the balanced and the imbalanced design. The overall column provides the classification accuracy across all classes. Then per-class accuracy is provided in subsequent columns

	Balanced				Imbalanced			
	Overall	Random noise	AR(1)	Nonstationary	Overall	Random noise	AR(1)	Nonstationary
EucD	0.61	0.48	0.92	0.43	0.71	0.67	0.78	0.67
DTWd	0.45	0.33	1	0.01	0.77	0.67	0.98	0.67
WCDmin	0.76	0.67	0.77	0.82	0.7	0.67	0.76	0.67
WCDmean	0.68	0.67	0.77	0.6	0.7	0.66	0.78	0.67
sqrC	0.55	0.44	0.92	0.3	0.72	0.66	0.83	0.66
DTWc	0.54	0.45	0.84	0.33	0.71	0.67	0.78	0.67
WCCmax	0.68	0.66	0.76	0.61	0.71	0.67	0.78	0.67
WCCmean	0.68	0.66	0.76	0.61	0.71	0.67	0.78	0.67
AR	0.65	0.66	0.81	0.49	0.78	0.67	1	0.67

References

- Bagnall, A., & Lines, J. (2014). *An experimental evaluation of nearest neighbour time series classification* (Technical Report). Retrieved from <http://arxiv.org/abs/1406.4757>.
- Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3), 606–660. <https://doi.org/10.1007/s10618-016-0483-9>.
- Birditt, K. S., Fingerman, K. L., & Almeida, D. M. (2005). Age differences in exposure and reactions to interpersonal tensions: A daily diary study. *Psychology and Aging*, 20(2), 330–340. <https://doi.org/10.1037/0882-7974.20.2.330>.
- Boker, S. M., Staples, A. D., & Hu, Y. (2016). Dynamics of change and change in dynamics. *Journal for Person-Oriented Research*, 2(1–2), 34–55. <https://www.ncbi.nlm.nih.gov/pubmed/29046764>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5642952/>.
- Boker, S. M., Xu, M., Rotondo, J. L., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, 7(3), 338–355.
- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control*. Wiley. <https://books.google.com/books?id=IJnnPQAACAAJ>.
- Conway, N., & Briner, R. B. (2002). A daily diary study of affective responses to psychological contract breach and exceeded promises 23(3). <https://doi.org/10.1002/job.139>.
- Ezzyat, Y., Kragel, J. E., Burke, J. F., Levy, D. F., Lyalenko, A., Wanda, P., et al. (2017). Direct brain stimulation modulates encoding states and memory performance in humans. *Current Biology*, 27(9), 1251–1258. <http://linkinghub.elsevier.com/retrieve/pii/S0960982217303263>.
- Gates, K. M., Lane, S. T., Varangis, E., Giovanello, K., & Guiskewicz, K. (2017). Unsupervised classification during time-series model building. *Multivariate Behavioral Research*, 52(2), 129–148. <https://doi.org/10.1080/00273171.2016.1256187>.
- Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, 31(7), 1–24. [http://www.jstatsoft.org/v31/i07/](http://www.jstatsoft.org/v31/i07/%5Cnhttp://www.jstatsoft.org/v31/i07/).
- Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21(4), 447–457. <https://doi.org/10.1037/met0000120>.
- Jeong, Y.-S., Jeong, M. K., & Omitaomu, O. A. (2011). Weighted dynamic time warping for time series classification. *Pattern Recognition*, 44, 2231–2240.
- Kounios, J., & Beeman, M. (2009) The Aha! Moment. *Current Directions in Psychological Science*, 18(4), 210–216. <https://doi.org/10.1111/j.1467-8721.2009.01638.x>.
- Laurenceau, J. P., Barrett, L. F., & Rovine, M. J. (2005). The interpersonal process model of intimacy in marriage: A daily-diary and multilevel modeling approach. *Journal of Family Psychology*, 19(2), 314–323.
- Ramseyer, F., & Tschacher, W. (2006). Synchrony: A core concept for a constructivist approach to psychotherapy. *Constructivism in the Human Sciences*, 11(1–2), 150–171. http://www.researchgate.net/publication/215507443_Synchrony_A_Core_Concept_for_a_Constructivist_Approach_to_Psychotherapy/file/3606d2cad7a8d399c757cbb48c1e8ec.pdf.
- Ramseyer, F., & Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology*, 79(3), 284–295.
- Sakoe, H., & Chiba, S. (1978). Dynamic Programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), 43–49.
- Sempena, S., Maulidevi, N. U., & Aryan, P. R. (2011). Human action recognition using dynamic time warping. In *2011 International Conference on Electrical Engineering and Informatics (ICEEI)* (pp. 1–5).

- Solomon, D. A., Leon, A. C., Coryell, W. H., Endicott, J., Li, C., Fiedorowicz, J. G., et al. (2010). Longitudinal course of bipolar I disorder. *Archives of General Psychiatry*, *67*(4), 339. <https://doi.org/10.1001/archgenpsychiatry.2010.15>.
- Wang, L. P., Hamaker, E., & Bergeman, C. S. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*, *17*(4), 567–581. <https://doi.org/10.1037/a0029317>.
- West, S. G. & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: daily experiences. *Journal of personality*, *59*(3), 609–662.

Topic Modeling of Constructed-Response Answers on Social Study Assessments



Jiawei Xiong, Hye-Jeong Choi, Seohyun Kim, Minho Kwak,
and Allan S. Cohen

Abstract Topic models were used to detect the latent thematic structure of examinees' answers to constructed-response items. Results for two different topic models, latent Dirichlet allocation (LDA) and supervised LDA, were compared for their utility in detecting different latent thematic patterns in examinees' responses on US History and Economics tests. LDA and sLDA results suggested both a four-topic model for the US History item and a three-topic model for the Economics item. For the US History item, Topic 1 consisted of use of everyday language and was negatively correlated with the rubric-based score. Topic 4, use of academic language focusing on government and politics, was positively correlated with the score. For the Economics test, Topic 3 consisted of use of technical vocabulary and had a positive correlation with item score. Complete results are discussed in the paper.

Keywords Topic models · LDA · sLDA · Constructed-response items

1 Introduction

Constructed-response (CR) answers are used on many educational tests as a means of having examinees show their reasoning (Attali 2014). CR answers are typically scored on one or more traits and the scores taken as measures of achievement in the domain being tested. Once answers have been scored, however, little, if any, attention is paid to the text in examinees' answers. Recent qualitative evidence has suggested there is useful information in the text of answers to CR items (Buxton et al. 2014). A comparison of qualitative evidence with statistical evidence from

J. Xiong (✉) · H.-J. Choi · M. Kwak · A. S. Cohen
University of Georgia, Athens, GA, USA
e-mail: jiawei.xiong@uga.edu; hjchoi1@uga.edu; minho.kwak25@uga.edu; acohen@uga.edu

S. Kim
University of Virginia, Charlottesville, VA, USA
e-mail: seohyun@uga.edu

topic modeling suggests that there may be instructionally useful information in the text not accounted for in the rubric-based scores (Kim et al. 2017).

Topic modeling consists of a family of statistical models initially developed for indexing the text of large corpora of documents (Blei 2012). The topic models provide a tool for mining of textual data in an effort to detect the latent semantic structures in the textual data. Blei presents a brief summary of several more commonly used topic models. In this study, we examined the use of two of these models, latent Dirichlet allocation (LDA) and supervised LDA (sLDA), for use in analyzing the text of CR answers to items on two social studies tests.

1.1 Latent Dirichlet Allocation and Supervised Latent Dirichlet Allocation

1.1.1 Latent Dirichlet Allocation (LDA)

LDA (Blei et al. 2003) is one of the simplest topic models. LDA has been applied to students' written responses in educational assessments (Choi et al. 2017) and detects latent topics in a corpus of text documents (Bolelli et al. 2009). LDA can also be applied to detect latent topics in the text of students' journal writings (Chen et al. 2016) and to analyze middle grades students' answers to CR items (Kim et al. 2017). LDA assumes a document is composed of a random mixture of topics, and a topic is a random mixture of words.

A document is assumed to consist of a random mixture of K topics. Each topic is a collection of V words, $\mathbf{t} = (w_1, w_2, \dots, w_V)'$, with the vector of probabilities of $\beta_{\text{topic-word}} = (\beta_{z,1}, \beta_{z,2}, \dots, \beta_{z,V})'$. The generative model of LDA describes θ_d , the vector of topic proportions in document d , $\theta_{\text{document-topic}} = (\theta_{d,1}, \theta_{d,2}, \dots, \theta_{d,K})'$. These proportions follow a Dirichlet distribution with parameter α as $\theta_d \sim \text{Dirichlet}(\alpha)$, and then for each of the N_d words, it chooses a topic z_{dn} from $\text{Multinomial}(\theta_d)$. Then it chooses a word w_{dn} from $p(w_{dn}|z_{dn}, \beta)$. For a collection D of M documents, given the parameters β and α , the probability of the corpus can be written as (Blei et al. 2003):

$$p(D|\alpha, \beta) = \prod_{d=1}^M \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d \quad (1)$$

where M is the number of documents.

1.1.2 Supervised Latent Dirichlet Allocation (sLDA)

LDA is considered an unsupervised method as it only uses the text in a corpus of documents to determine the latent topic structure. In the supervised LDA, external

information is used to help guide the LDA. Supervised latent Dirichlet allocation (sLDA; Mcauliffe and Blei 2008) is an extension of the LDA model that includes additional information, referred to as labels. In the context of CR answers, the labels are the rubric-based scores of examinees' answers. Other kinds of covariates could include political preferences, movie ratings, etc. As an example, sLDA has been used to analyze textual data in an attribute selection task on adjective-noun phrases (Hartung and Frank 2011).

Little research has yet been reported comparing results from LDA and sLDA in the context of answers to CR items. In this paper, we use topic modeling to analyze the text of examinees' answers to CR items on two social studies tests: a US History test and an Economics test. Topic models have been used in the past in social science research. Grimmer (2010) used a topic model to detect the latent structure in political rhetoric. Roberts et al. (2013) used a topic model to detect the latent structure in open-ended responses to a social science survey. Topic modeling also has been used to detect historical trends in newspapers (Yang et al. 2011).

In this study, we analyzed examinees' responses to CR items on a high school tests of Economics and US History. The purpose of this study was to compare results obtained from LDA and from sLDA

2 Methods

2.1 Participants and Instruments

2.1.1 Participants

Two corpora were analyzed in this study: CR answers to standardized assessments, a US History test and an Economics test. The US History test was administered to 722 examinees in Grade 9 to Grade 12. Economics test was administered to 663 examinees in Grade 9 to Grade 12.

2.1.2 Instruments

Both tests were developed to be aligned to the state standards in the respective subjects for a large Southeastern state. There were 22 multiple-choice items, 2 short answer CR items, and 1 extended CR on each test. The CR items were designed to require extended reasoning and critical thinking. The two short answer CR items were scored from 0 to 2 points, and the extended response item was scored from 0 to 4 points. Only the extended response items in each assessment were analyzed for purposes of this study. For both tests, the extended response item consisted of a question followed by two passages describing the context for the response.

2.2 Data Cleaning

The process of data cleaning of a corpus is typically done in a topic model analysis (Boyd-Graber, Mimno and Newman (2014)). Initial data cleaning included removal of white spaces, changing numerical digits to text, changing uppercase letters to lowercase, removal of punctuation characters, etc. Correcting typos or other types of stemming was also done before the LDA analysis.

Stop words are words that tend to have high frequencies but low information. These include words such as “the,” “a,” “of,” etc. The stop words for this study are shown in Table 1. Stop words were not included in the analysis as they can overwhelm the latent thematic structure, making interpretation difficult.

Documents with less than ten words after data cleaning were also excluded from the analysis. Thus, the number of documents was reduced following data cleaning. Descriptive statistics of the numbers of words and documents and average document length are given in Table 2.

2.3 Model Selection

2.3.1 Deviance Information Criterion

Exploratory use of a topic model typically consists of estimating models with different numbers of latent topics. The best fitting model of these candidate models

Table 1 Stop words for the US History test and the Economics test

US History item						Economics item					
Next	Into	Every	Not	Their	This	Next	Not	Their	This	Only	One
Only	One	Much	Can	Yet	For	Much	Can	Yet	For	And	Are
Could	And	Are	That	What	Him	That	What	Him	With	But	Out
With	But	Out	His	Who	From	His	Who	From	Will	They	Also
Will	They	Also	Which	Other	You	Which	Other	You	Still	Our	All
Still	Our	All	How	Than	Two	How	Than	Two	After	Many	Have
After	Many	Have	Both	There	According	Both	There	Just	Now	Every	Into
Now	Just					Its	When	While	Then	About	Yes

Table 2 Number of documents, number of words, and average document length

	US History		Economics	
	Before	After	Before	After
	Processing	Processing	Processing	Processing
Number of documents	722	416	663	482
Number of unique words	583	296	332	145
Number of total words	22,203	9,726	19,526	9,143
Average length	53	23	40	19

then needs to be determined. As topic models are not nested, selecting the best fitting model typically is informed using one or more information criterion indices. When the topic model is estimated using a Bayesian algorithm, the deviance information criterion (DIC; Spiegelhalter et al. 2002) is usually used to inform model selection as follows:

$$DIC = D(\bar{\theta}) + 2p_D \tag{2}$$

where $D(\bar{\theta}) = -2 \log p(y|\bar{\theta}) + 2 \log f(y)$ and $p_D = D(\bar{\theta}) - D(\hat{\theta})$. Here y denotes the data, θ denotes the parameter, $f(y)$ is some fully specified standardizing term which is a function of the data alone, and $D(\theta)$ is the ‘‘Bayesian deviation’’. The DIC could be calculated through the MCMC algorithm and smaller value of DIC indicates a better fitting model.

3 Results

3.1 Latent Dirichlet Allocation

In this study, DIC suggested a four-topic model for US History item (left) and a three-topic model for Economics (right). The plots of DIC values for topic models with two to ten latent topics are given in Fig. 1. The lowest DIC for each test is taken as the suggested model.

3.1.1 US History Test Results

The correlation between the rubric-based score and the topics estimated in the four-topic model is given in Table 3. The 15 highest probability words for each topic are also given in Table 3. Inspection of these high probability words for each topic can

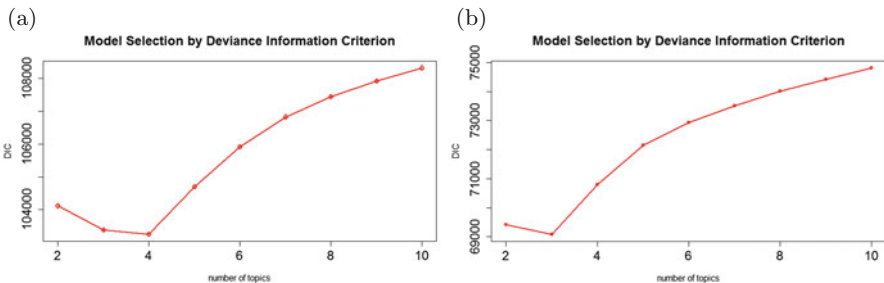


Fig. 1 Plots of model selection by DIC form the LDA. (a) DIC values for US History item. (b) DIC values for Economics item

often help to interpret the latent theme captured by the topic. Table 3 lists the top 15 most probably words co-occurring for each of the four topics. Topic 3 contained high-frequency words that could be characterized as use of *everyday words* in their answers to the item. Topic 4 consisted of words about *US presidents and civil rights*. Examinees who used these words followed directions in the prompt and tried to integrate information in the passages for the item and to use this information evidences to support their answer.

Correlations that are shown in Table 3 between the topic and the rubric-based score are given in the heading for each topic. Topic 1 had a moderate negative correlation ($r = -0.361$) with the rubric-based score, and Topic 4 had a moderate positive correlation ($r = 0.443$) with the rubric-based score. It is also sometimes useful to examine the answers by examinees who make the highest use of each topic. For example, examinees who had the highest use of Topic 1 typically wrote answers to the question that simply copied the information in the stem or passages. Topic 2 had some important words from the item, but examinees using this topic tended to take sentences directly from the item question or passage without trying to integrate them into an answer. Topic 3 consisted of an integrated structure of both everyday words with language from the passages; however, the answer did not include a clear argument. Examinees who made most use of Topic 4 typically used words from the passages and integrated them to provide evidence for their conclusions.

Table 3 Fifteen highest probability words for the four-topic model for US History using LDA

Topic 1 $r = -0.361^a$		Topic 2 $r = -0.162$		Topic 3 $r = -0.010$		Topic 4 $r = 0.443$	
Part	0.274	Randolph	0.040	Right	0.093	March	0.057
Know	0.049	Labor	0.025	Part	0.081	Martinluther king	0.049
Help	0.036	Black	0.025	Civil	0.042	Randolph	0.048
Want	0.017	America	0.023	People	0.032	President washington	0.040
Follow	0.010	African	0.020	Movement	0.024	America	0.039
Work	0.009	Racial	0.014	Protest	0.022	Leader	0.025
Learn	0.009	Social	0.013	Get	0.021	African	0.024
Non	0.008	First	0.011	Fight	0.019	Protest	0.021
Get	0.008	Civil	0.011	Want	0.017	Right	0.019
Like	0.008	World	0.010	Equal	0.012	Discrimination	0.018
People	0.008	Brotherhood	0.009	Make	0.011	Civil	0.016
African	0.007	War	0.009	Randolph	0.011	Industry	0.016
White	0.007	Movement	0.009	Because	0.010	Lead	0.016
President washington	0.007	During	0.009	Impact	0.010	War	0.014
Could	0.006	Philip	0.009	Start	0.010	Federal	0.013

^aCorrelation is between item score and logit of proportion of topic usage

Table 4 Fifteen highest probability words for the three-topic model for the Economics test using LDA

Topic 1 $r = -0.403^a$		Topic 2 $r = 0.124$		Topic 3 $r = 0.272$	
Part	0.385	Interest	0.236	Interest	0.196
Money	0.047	Compound	0.055	Compound	0.079
Compound	0.034	Principal	0.053	Amount	0.068
Because	0.025	Simple	0.051	Pay	0.057
Dollar	0.022	Rate	0.038	Money	0.049
Interest	0.020	Loan	0.035	Simple	0.049
Rate	0.019	Time	0.026	Year	0.048
Know	0.019	Calculate	0.023	Because	0.036
Simple	0.019	Retire	0.016	Beneficial	0.022
Time	0.016	Deposit	0.012	Part	0.015
Take	0.013	Save	0.012	Rate	0.011
Make	0.012	Period	0.012	Add	0.011
Add	0.010	Addition	0.010	Save	0.010
Retire	0.009	Get	0.010	Investment	0.009
Good	0.009	Good	0.010	Time	0.009

^aCorrelation is between item score and logit of proportion of topic usage

3.1.2 Economics Test Results

Table 4 presents results for the Economics test. Topic 1 had a moderate negative correlation ($r = -0.403$) with the score, and Topic 3 had a positive correlation ($r = 0.272$) with the score. The correlation for Topic 2 ($r = 0.124$) is low albeit positive. Topic 2 and Topic 3 can both be characterized as use of academic language related to interest calculation. Examinees who use words mainly from Topic 2 were typically repeating the definitions in the passages while calculating the simple interest posed in the question. Examinees who use more words from Topic 3 provided answers that included choices and computation of the principle. Their responses also provided a convincing rationale. Topic 1 contains several simple words but did not provide a clear answer to the question. Some of the words for Topic 1 were actually not relevant for the answer to the item.

3.2 Supervised Latent Dirichlet Allocation Modeling

The supervised LDA model uses a linear regression model to predict an outcome variable using the topic model proportions. In our case, the outcome variable is the rubric-based score, and it is regressed on the topic proportions. The following is the regression model for the sLDA:

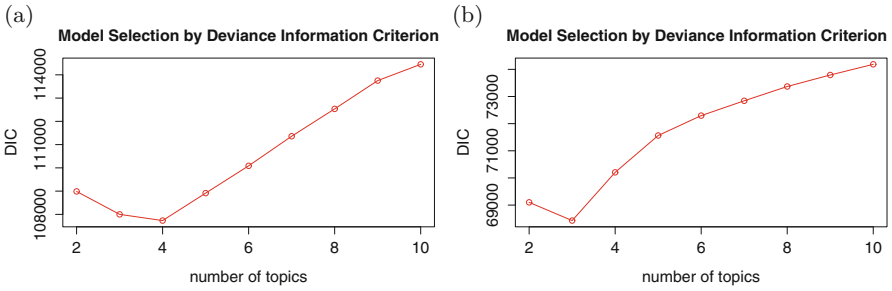


Fig. 2 Plots of model selection by DIC from the sLDA. (a) DIC values for US History item. (b) DIC values for Economics item

$$Y_i = \beta X \tag{3}$$

where the observation $Y_i = [y_1 \ y_2 \ \dots \ y_n]'$, regression coefficients $\beta = [\beta_1 \ \dots \ \beta_n]'$, and the topic proportion matrix $X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$, while

k is the number of the topics.

The plots of DIC are given in Fig. 2 for topic models with two to ten latent topics. The DIC results are similar to the LDA results and suggested a similar four-topic model for US History (left) and a three-topic model for Economics (right).

3.2.1 US History Test

There is no intercept β_0 in the regression as the topic proportions sum to 1, i.e., $\sum_{k=1}^4 x_{nk} = 1$ here for the US History item. The topic structure from the sLDA is given in Table 5 for a four-topic model, which shows a pattern of topic proportions similar to those for the LDA model. Only the order of some words changed slightly.

Topic 1 has a coefficient of $\beta = -0.159$ which means examinees who mostly use words from Topic 1 tend to have a low score. Similarly, Topic 2 has a coefficient of $\beta = -0.015$, which also means examinees who use words mostly from Topic 2 also have a low score. Topic 4 has a coefficient of $\beta = 2.530$, which means examinees who made most use of this topic tended to have a score of 2.53 points.

Differences between the observed score and the predicted score from the sLDA model are shown in the scatter plot in the left graph of Fig. 3. The mean for these differences is given by $\mu = \sum_{i=1}^n |y_i - \hat{y}_i| / n = 0.598$, and the standard deviation is 0.538, which indicated a relatively good fit to the data.

Table 5 Fifteen highest probability words for the four-topic model for US History using sLDA

Topic 1 $\beta = -0.159^a$		Topic 2 $\beta = -0.015$		Topic 3 $\beta = 1.169$		Topic 4 $\beta = 2.530$	
Part	0.533	Randolph	0.063	Right	0.153	March	0.098
Know	0.078	Labor	0.041	Civil	0.100	Randolph	0.082
Help	0.058	Black	0.039	Protest	0.065	America	0.079
Want	0.054	America	0.030	People	0.061	African	0.063
Get	0.048	Racial	0.022	Movement	0.060	Discrimination	0.035
Make	0.028	War	0.021	Because	0.034	Lead	0.033
Give	0.022	First	0.020	Fight	0.031	Work	0.033
Thing	0.014	African	0.020	Equal	0.029	President	0.031
Same	0.012	Union	0.019	Impact	0.028	Industry	0.027
Null	0.012	During	0.019	Influence	0.027	Leader	0.025
Everyone	0.012	Social	0.019	Direct	0.023	Federal	0.024
Stand	0.009	World	0.018	Leader	0.022	Order	0.019
Cause	0.008	Car	0.017	Like	0.020	Threat	0.019
Good	0.008	Philip	0.017	Follow	0.020	Equality	0.018
Man	0.008	Group	0.016	Peace	0.018	Government	0.018

^aRegression coefficients for regression of observed score on topic proportions

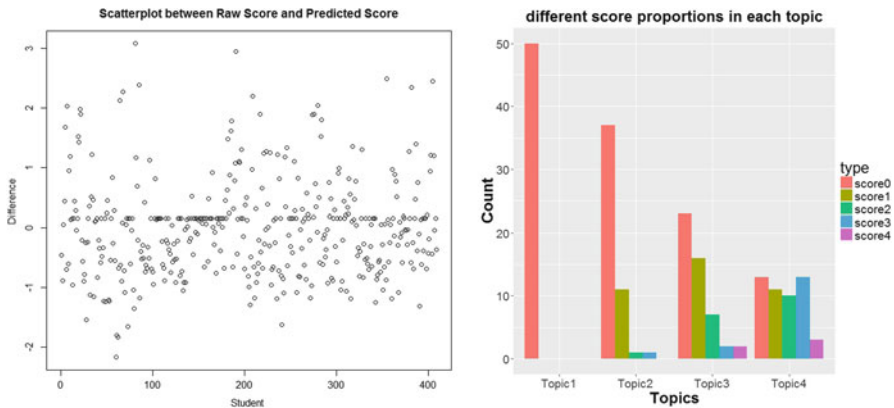


Fig. 3 Plots of sLDA model prediction and score proportion in each topic for US History item

Observed scores of the 50 examinees who had the highest percentages of use of words from each topic are plotted in the histogram in the right graph in Fig. 3. As is evident from the regression coefficients in Table 5, examinees who used words mainly from Topic 3 or Topic 4 had higher scores than examinees who used words mainly from Topic 1 or Topic 2. A full credit score of 4 did not occur for examinees who used words mostly from Topics 1 or 2. In addition, by comparing the quantity of zero scores among all the topics, the number of examinees who use Topic 4 is the lowest.

Table 6 Fifteen highest probability words for the three-topic model for the Economics test using sLDA

Topic 1 $\beta = -0.281^a$		Topic 2 $\beta = 0.400$		Topic 3 $\beta = 3.671$	
Part	0.485	Interest	0.305	Interest	0.203
Simple	0.092	Pay	0.070	Compound	0.190
Because	0.076	Rate	0.068	Amount	0.092
Money	0.054	Principal	0.067	Year	0.068
Save	0.033	Simple	0.052	Money	0.060
Know	0.028	Loan	0.042	Time	0.056
Get	0.027	Retire	0.034	Add	0.030
Take	0.026	Calculate	0.029	Beneficial	0.029
Make	0.018	Good	0.020	Over	0.023
Bank	0.018	Charge	0.017	Principle	0.021
Back	0.015	Deposit	0.017	Account	0.017
Null	0.014	Period	0.016	Earn	0.017
Little	0.010	Investment	0.013	Build	0.014
Double	0.009	Long	0.011	Borrow	0.013
Help	0.009	Sum	0.011	End	0.013

^aRegression coefficients for regression of observed score on topic proportions

3.2.2 Economics Test Results

The topic structure of the Economics test in Table 6 response shows similar characteristics with the results of LDA’s. Topic 1 has a coefficient of $\beta = -0.281$, which indicates that the examinees who mostly used words from Topic 1 tended to have a lower score. Topic 2 has a coefficient of $\beta = 0.400$, which indicates the examinees who mostly use words from Topic 2 may get few points. Topic 3 has a coefficient of $\beta = 3.671$, which means with examinees using more words from Topic 3 tend to have higher scores.

Differences between the observed score and the predicted score from the sLDA model are shown in the scatter plot in the left graph of Fig. 4. The mean for these differences is given by $\mu = \sum_{i=1}^n |y_i - \hat{y}_i| / n = 0.695$, and the standard deviation was 0.530. These suggest a comparatively good fit to the data.

Observed scores of the 50 examinees who had the highest percentages of use of words from each of the three topics are plotted in the histogram in the right graph in Fig. 4. Examinees who used words mainly from Topic 3 tended to have higher scores than examinees who used words mainly from Topic 1 or Topic 2. Few examinees who used more words from Topic 3 had zero scores. A score of 4 does not appear for examinees who used words mostly from Topic 1.

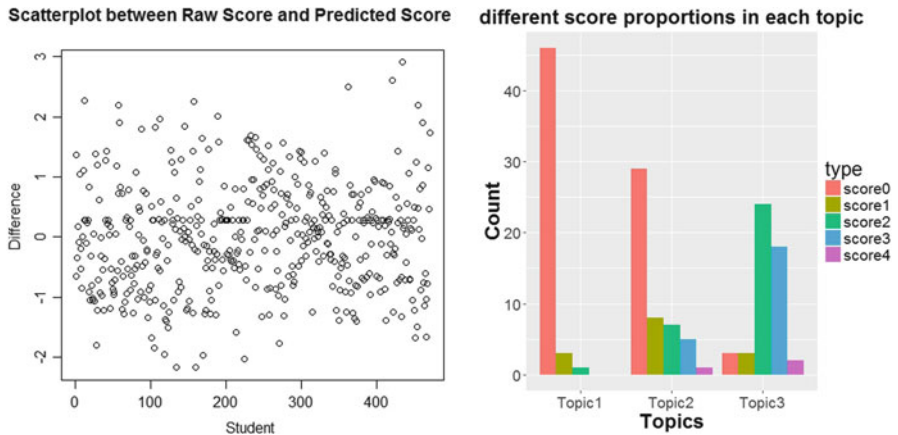


Fig. 4 Plots of sLDA model prediction and score proportion in each topic for Economics item

4 Discussion

Previous research has suggested that samples from a test of English and Language Arts as small as 150 documents could be analyzed with LDA (Kim et al. 2017). Sample sizes in this study were somewhat larger for each of the two tests.

The topic structures detected by LDA and sLDA differed for the US History test and the Economics test, but the topic structures for a given test were similar. Correlations for the US History test between the observed score and the topic proportions from the LDA model indicated that use of Topic 4 was modestly related to a higher score and use of Topics 1 or 2 was related to a lower score. The regression coefficients from the sLDA suggested a similar outcome as use of words from Topic 4 was associated with a higher predicted score than use of words from Topics 1 or 2. For the Economics test, correlations between topic proportions from the LDA and observed score suggested use of words from Topic 3 was moderately related to higher scores, and use of Topic 1 was moderately related to lower scores. Similarly, the use of words from Topic 3 was associated with a high predicted score, and use of words from Topic 1 was associated with a low score of effectively zero.

What is clear from the topic modeling of the results from both tests is that information about the latent thematic structure of the text of answers can extend what can be learned from analysis of CR tests. The topic structure can provide information of instructional effects. Previous research has suggested that instructional effects can be observed in the use of each topic even though these same effects may not be present in the scores (Kwak et al. 2017). Attali (2014) has suggested CR tests can tell us about examinee reasoning. What may be evident from the topic model results is that differences in examinee reasoning might be reflected in the words used to construct answers. It would be useful to examine this conjecture in future research.

References

- Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement*, 74(5), 795–808.
- Blei, D. M. (2012). Surveying a suite of algorithms that offer a solution to managing large document archives. *Communication of the ACM*, 55(4), 77–84.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Bolelli, L., Ertekin, S., & Giles, C. L. (2009). Topic and trend detection in text collections using latent Dirichlet allocation. In *European Conference on Information Retrieval* (pp. 776–780). Springer, Berlin/Heidelberg.
- Boyd-Graber, J., Mimno, D., & Newman, D. (2014). *Care and feeding of topic models: Problems, diagnostics, and improvements. Handbook of mixed membership models and their applications*, 225255. Boca Raton, FL: CRC Press.
- Buxton, C., Alleksaht-Snyder, M., Aghasaleh, R., Kayumova, S., Kim, S. H., Choi, Y. J., & Cohen, A. (2014). Potential benefits of bilingual constructed response science assessments for understanding bilingual learners' emergent use of language of scientific investigation practices. *Double Helix*, 2.
- Congdon, P. (2007). *Bayesian statistical modelling* (Vol. 704). Chichester: Wiley.
- Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1–5). ACM.
- Choi, H. J., Kwak, M., Kim, S., Xiong, J., Cohen, A. S., & Bottge, B. A. (2017). An application of a topic model to two educational assessments *Quantitative psychology: The 83rd annual meeting of the psychometric society*, 265, (pp. 449–459). Cham: Springer.
- Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1), 1–35.
- Hartung, M., & Frank, A. (2011). Exploring supervised LDA models for assigning attributes to adjective-noun phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 540–551). Association for Computational Linguistics.
- Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & Cohen, A. S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics*, 1, 82–102.
- Kwak, M., Kim, S., & Cohen, A. S. (2017). *Mining students' constructed response answers*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio.
- Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121–128). Red Hook, NY: Curran Associates, Inc.
- Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: Computation, application, and evaluation* (pp. 1–20).
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- Yang, T. I., Torget, A. J., & Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 96–104). Association for Computational Linguistics.

Impact of Measurement Bias on Screening Measures



Oscar Gonzalez , William E. Pelham III , and A. R. Georgeson 

Abstract In psychology and medicine, diagnostic and screening measures are often used to make decisions by comparing the observed score to a cut score. If these measures contain items that exhibit measurement bias, then there might be systematic inaccuracies of who gets “caught” by the selection process. Traditionally, approaches that flag items for bias do not provide guidance about how measurement bias affects the decisions from the overall measure. In previous work on the area of selection, Millsap and Kwok developed a procedure that described the impact of ignoring measurement bias as changes in test sensitivity and specificity across groups. Recently, the Millsap and Kwok procedure has been extended to handle discrete items and make less stringent distributional assumptions. In this chapter, we discuss a version of the Millsap and Kwok procedure that accommodates discrete items and illustrate the use of this approach to evaluate how measurement bias affects the sensitivity and specificity of a measure comprised of binary items.

Keywords Differential item functioning · Measurement bias · Screening · Sensitivity · Specificity

1 Impact of Measurement Bias on Screening Measures

In psychology and medicine, diagnostic measures and screening procedures are valuable tools used to obtain observed scores which can then supplement a clinician’s diagnosis or identify respondents who may be at risk for a mental health disorder. Examples include the Child Behavior Checklist (CBCL; Achenbach and

O. Gonzalez (✉) · A. R. Georgeson
Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill,
Chapel Hill, NC, USA
e-mail: ogonza13@unc.edu; georgeson@unc.edu

W. E. Pelham III
Department of Psychology, Arizona State University, Tempe, AZ, USA
e-mail: wpelham@asu.edu

Rescola 2001) to screen children for ADHD; the Child and Adolescent Symptom Inventory – Revised (CASI-4R; Gadow and Sprafkin 2005) to screen children for pediatric bipolar spectrum disorders (Ong et al. 2017); and the K6 scale (Kessler et al. 2002) to screen for nonspecific psychological distress in the general population (Kim et al. 2016). Diagnostic measures or screeners are commonly comprised of binary items (i.e., a symptom is present or not) or polytomous items (i.e., Likert-type categories), and an observed score is estimated by aggregating item responses or counting symptoms. A researcher or an assessment specialist would then classify a respondent by examining if the observed summed score is above or below a predetermined cut score (Youngstrom 2013).

In practice, if the scores of respondents from multiple (g) groups are going to be compared to the same cut score, then measurement invariance is assumed (Millsap 2011). Formally, measurement invariance can be expressed as,

$$P(X|\theta) = P(X|\theta, g).$$

In other words, the probability of observing score X given the respondent's standing on the latent variable θ assessed by the items does not depend on the background characteristic(s) that define group g (Millsap 2011). However, in some cases, a certain respondent group (e.g., males) systematically rates itself higher or lower on a subset of items than does a different respondent group (e.g., females), independent of the θ being assessed. In this case, the subset of the items exhibits measurement bias across groups, also referred to as differential item functioning (DIF; Millsap 2011), which violates the assumption of measurement invariance. Assuming that two groups have the same distribution of θ , the presence of measurement bias can lead to three different outcomes: (1) respondents from a certain group have a *higher* likelihood of being caught (i.e., flagged, identified) by the screener, (2) respondents from a certain group have a *lower* likelihood of being caught by the screener, or (3) there are no systematic effects on the likelihood of being caught by the screener across groups. The practical implications of the first two outcomes are of concern as they could lead to incorrectly screening *in* persons who do not actually exceed the cut score, resulting in lost time and resources for all parties involved, or incorrectly screening *out* persons who actually exceed the cut score, resulting in a lack of proper services for individuals in need. A case for the third outcome is that bias in one item could lead to a higher item score, and bias in a different item could lead to a lower item score, so bias could cancel out once items scores are aggregated. As such, it is important that assessment specialists test for measurement bias before administering measures to priority groups (i.e., groups defined by race, ethnicity, language of origin, or gender) in order to prevent placing any group at a disadvantage. However, work assessing the cultural equivalence of screening or diagnostic measures has been sparse (Manly 2006; Teresi et al. 2006).

While significance testing procedures have traditionally been used to flag items that might exhibit bias (Millsap 2011), the statistical significance obtained in these procedures is often not a meaningful proxy for practical significance. Moreover, there is also a lack of guidance about how the bias affects classification

decisions. Previous research has investigated several approaches to study the effect of measurement bias on the decisions made from the overall measure, such as graphical approaches, differences between item parameters across groups, or effect sizes that describe the change in expected observed summed score after accounting for item bias (Kleinman and Teresi 2016; Meade 2010; Steinberg and Thissen 2006). However, these approaches do not shed light on how measurement bias affects the *screening performance* of the measure, making it difficult for a practitioner to decide if the bias is tolerable in the context of selection. Millsap and Kwok (2004) proposed an approach to evaluate if the presence of measurement bias materially changes the sensitivity and specificity of the measure in each respondent group. This procedure has the distinct advantage of communicating the effect of measurement bias in terms familiar to assessment specialists (e.g., changes in sensitivity and specificity), which empowers the specialists to decide how much bias they can tolerate. Below, we describe the Millsap and Kwok (2004) procedure and its current extensions.

1.1 Millsap and Kwok (2004) Procedure

Suppose that the examined measure has a unidimensional linear factor structure. Millsap and Kwok (2004) indicated that if the items are continuous and the relation between the items and the factor is linear, then the relation between the observed summed score on the whole measure X and the latent variable θ assessed by the measure is a bivariate normal distribution. Suppose now that we have two groups that have the same latent mean and variance. When measurement invariance holds, there is one bivariate normal distribution for both groups. When only a subset of the items are invariant across groups (i.e., *partial invariance* holds), the relation between θ and X is a mixture of two bivariate normal distributions, defined by the two group-specific bivariate normal distributions. Under partial invariance, θ scores are on the same metric and can be directly compared, but X are not in the same metric—any observed differences in X could either reflect true differences in θ or measurement bias. Millsap and Kwok (2004) used the previous relations to study the classification agreement between selecting individuals based on their estimated θ score and selecting individuals based on their observed summed score, X . One could determine an expected X from a model in which measurement bias in the measure is accounted for (i.e., groups have group-specific item parameters) and an expected X from a model in which measurement bias of the measure is ignored (i.e., groups have the same item parameters). To study the classification agreement, cut scores are imposed on X and θ in order to define four quadrants of the bivariate distribution of θ and X . Then, the cut scores can be used to define four different types of cases: true positives (respondents who are above the cut score in both θ and X), true negatives (respondents who are below the cut score of both θ and X), false positives (respondents who are below the cut score of θ , but above the cut score of X), and false negatives (respondents who are above the cut score of θ , but below the cut score of X). The Millsap and Kwok (2004) procedure communicates the effect

of measurement bias on screening performance by comparing how the functions of the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) change when the expected X comes from a measure assumed to be invariant and when the expected X comes from a measure that accounts for measurement bias. Functions to automate the Millsap and Kwok (2004) procedure in the R statistical software environment are provided by Lai et al. (2017).

A limitation of the Millsap and Kwok (2004) procedure is the assumption that items from a measure are continuous. In fact, items in many screening measures are discrete. Lai et al. (2019) extended the Millsap and Kwok procedure to binary items using an analytical solution. However, the current implementation of the analytical procedure by Lai et al. (2019) does not accommodate polytomous items, nor does it have the potential to handle mixed item-types. Recently, Gonzalez and Pelham (in press) outlined an approach inspired by Millsap (2013) that is analogous to the Millsap and Kwok (2004) procedure to handle binary and polytomous items. Their procedure accommodates discrete items by using item response models, such as the two-parameter logistic (2PL) model or the graded response model (GRM), to represent item responses (Thissen and Wainer 2001). Also, instead of analytically deriving the relation between the X and θ , Gonzalez and Pelham approximated that relation using Monte Carlo simulation. Gonzalez and Pelham (in press) focused on illustrating how their simulation-based procedure can be used to examine how measurement bias affects screening performance on measures with polytomous items, but applications for measures with binary items have not been discussed.

1.2 Present Study

The goal of this chapter is to illustrate how the simulation-based procedure can be used to evaluate the impact of measurement bias on screening performance of a measure comprised of binary items. First, we introduce an empirical example in which one suspects that there might be measurement bias due to biological sex in the AQ-10, a brief measure used to screen for autism spectrum disorder (Murray et al. 2019). Second, we illustrate the use of the simulation-based approach to evaluate the impact of measurement bias on sensitivity and specificity of the AQ-10. Finally, we discuss the results of the procedure and consider future directions.

2 Method

2.1 Empirical Example

The AQ-10 is a brief screening measure used to identify individuals with possible autism spectrum disorder (ASD). The AQ-10 consists of ten items with four

response options ranging from *strongly disagree* to *strongly agree*; for scoring, item responses are dichotomized to *agree* and *disagree*. Previous research suggests that the AQ-10 is unidimensional and that a cut score at or above 6 could adequately screen participants with ASD and refer them a full diagnostic assessment (Murray et al. 2019). For this illustration, we used similar AQ-10 item parameters to those reported by Murray et al. (2019), but modified the item parameters of two items to fit our illustration. In this case, we induced measurement bias in two item parameters for females by making the *a*-parameters smaller (i.e., items were less representative of the construct for females) and the *b*-parameters larger (i.e., it takes more of the construct for females to endorse the item). Item parameters are presented in the top part of Table 1.

Table 1 Item parameters and diagnostic classification statistics for the two illustrative examples

	Item parameters			
	a_M	b_M	a_F	b_F
Item 1	.041	.454	.041	.454
Item 2	.717	.038	.717	.038
Item 3	1.432	-.215	1.432	-.215
Item 4	1.037	-.012	1.037	-.012
Item 5	2.532	-.045	2.532	-.045
Item 6	2.147	-.264	1.203	-.100
Item 7	1.146	.210	1.146	.210
Item 8	1.031	-.357	1.031	-.357
Item 9	2.902	-.191	2.029	.567
Item 10	1.422	.034	1.422	.034

	Classification accuracy			
	Ignoring DIF		Accounting for DIF	
	Males	Female	Males	Females
Sensitivity	.84	.83	.88	.80
Specificity	.87	.87	.84	.88
Classification rate	.85	.85	.86	.84
True positive %	.38	.38	.40	.36
True negative %	.47	.47	.46	.48
False positive %	.07	.07	.09	.06
False negative %	.07	.08	.06	.09
Prop. selected	.46	.46	.46	.43

Note: In bold are the item parameters modified (for illustration purposes) to induce some measurement bias across males (M) and females (F), and in turn affect screening performance when measurement bias is ignored

2.2 General Procedure

To carry out the simulation-based procedure, users need three pieces of information: the mean and standard deviation of the θ distribution for each group (assumed to be normally distributed), group-specific item parameters that are in the same metric, and the proportion of cases in each group. The simulation-based procedure assumes that the measure is unidimensional and that a subset of the items have been correctly flagged with DIF using any DIF procedure that yields item parameters, such as the IRT-LR-DIF procedure or a Wald test. The simulation-based procedure can be described in six steps:

1. Sample a large number of θ values from group-specific latent variable distributions. The number of cases sampled is directly related to the stability of the solution. Previous research suggests that the simulation-based procedure yields *stable* estimates (i.e., classification accuracy estimates within 0.01 from analytical estimates) with $N = 25,000$ (Gonzalez and Pelham, [in press](#)). Match the number of cases sampled per group to the population proportions for each group (e.g., males and females would be 50–50%). If population proportions are not known, sample proportions would be the best estimate. For the AQ-10 example, the mean and variance of the latent variable that the AQ-10 measures for males and females were not reported by Murray et al. (2019), so we assumed that both groups had a standard normal θ distribution (mean = 0, variance = 1).
2. Use a 2PL model to generate item responses for person i in group g using the θ_i , and group-specific a - and b -parameters per item (from the top part of Table 1).
3. Sum the simulated item responses to estimate X_i when the item response model accounts for measurement bias.
4. Plot the relation between θ and X , impose cut scores on θ and X , and estimate the proportion of respondents in each of the quadrants defined by the cut scores (see Fig. 1 for reference). For the AQ-10 example, the summed score cut score is 6, so a θ cut score is selected to choose the same proportion of respondents as a summed score at or above 6 in the mixed distribution of male and female respondents (e.g., if a summed score of 6 selects 30% of respondents, pick a θ that selects 30% of respondents).
5. Fit a 2PL model to the simulated item responses, assuming the item parameters are invariant across group, and save the estimated item parameters.
6. Repeat steps 2, 3, and 4 using θ from step 1 and the item parameters from step 5.

The proportions of respondents per quadrant estimated in steps 4 and 6 can be used to estimate classification accuracy. For example, sensitivity would be the proportion of respondents above the θ cut score that are also above the X cut score (i.e., $TP/[TP + FN]$). Similarly, specificity would be the proportion of respondents below the θ cut score that are also below the X cut score (i.e., $TN/[TN + FP]$). The false negative rate and the false positive rate are the complements of sensitivity and specificity, respectively. Comparing the expected sensitivity and specificity of the measure under a model that accounts for measurement bias (e.g., estimates from

step 4) and the expected sensitivity and specificity of the measure that ignores measurement bias (e.g., estimates from step 6) would provide an estimate of how measurement bias affects screening performance. Functions to automate this process in the R statistical software environment are presented in Gonzalez and Pelham ([in press](#)). The contribution of this chapter is to illustrate how the simulation-based procedure by Gonzalez and Pelham ([in press](#)) can be used to examine the effect of measurement bias on classification accuracy of screeners comprised of binary items.

3 Results

The relationship between the simulated AQ-10 θ score and the estimated AQ-10 observed X summed score under the model that allows for measurement bias is presented in Fig. 1. We selected cut scores to evaluate screening performance using the recommended AQ-10 cut score at or above 6, which indicates possible ASD. In the mixed distribution of males and female respondent, an observed X summed score of 6 or higher selected the top 44.3% of respondents. In the mixed distribution of male and female respondent, a θ cut score of 0.112 selected the same proportion (44.3%) of respondents.

Based on the θ and X cut scores, we calculated assessment sensitivity, specificity, and other classification statistics (see the bottom part of Table 1). We focus on interpreting sensitivity and specificity for this example. For sensitivity, if measurement invariance were to hold (i.e., items had equal discrimination and location parameters across groups), the expected sensitivity of the measure for males and females would be approximately 0.84. However, given that measurement invariance did *not* hold (i.e., some items exhibit DIF—the discrimination parameter, the location parameter, or both parameters differed across groups), the expected sensitivity of the measure was for 0.88 for males and 0.80 for females. Thus, when there is measurement

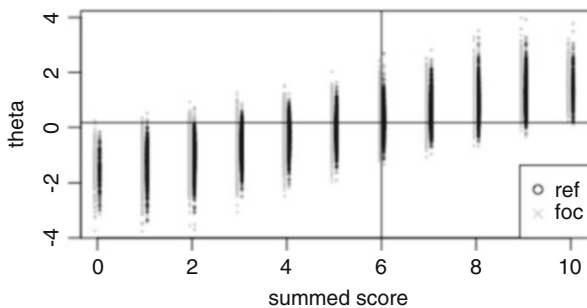


Fig. 1 Relation between AQ10 summed score X and AQ10 θ score. Note: “ref” is for the male reference group and “foc” is for the female focal group. For readability, the reference scores were dodged +0.05 points along the x-axis and the focal scores were dodged -0.05 points along the x-axis (rather than having the points overlap)

bias, the measure is *better* at identifying male respondents with possible ASD than female respondents with possible ASD. For specificity, if measurement invariance were to hold, the expected specificity of the measure for males and females would be 0.87. Given that measurement variance did not hold, the expected specificity of the measure was 0.84 for males and 0.89 for females. Thus, when there is measurement bias (also referred to as DIF), the measure is *better* at ruling out possible ASD in female respondents than in male respondents. Both results suggest that the presence of two items with measurement bias in the AQ-10 affected screening performance. Assessment specialists would have to decide if the differences across groups (e.g., 0.08 difference in sensitivity and a 0.05 difference in specificity) are practically important and/or acceptable in the specific screening application.

4 Discussion

The goal of this chapter was to describe how the simulation-based procedure proposed by Gonzalez and Pelham (in press) can be used to examine how item bias affects scale-level screening (or selection) decisions when the measure is comprised of binary items. Broadly, the procedure uses Monte Carlo simulations to determine the relation between the latent variable θ and the observed summed score X derived under two conditions: (1) a model that accounts for measurement bias in the items and (2) a model that ignores measurement bias. The classification agreement between θ and X across the two scenarios quantifies how measurement bias affects the sensitivity and specificity of the measure.

One limitation is that the procedure assumes that item parameters have been accurately estimated and that the model fits the data well. If this is not the case for a specific application, the procedure may not yield meaningful results. Perhaps a Bayesian approach to the simulation-based procedure could be used to accommodate the uncertainty in item parameter estimates. Another limitation is that the simulation-based procedure assumes that the items with measurement bias have been accurately identified and that the amount of bias is not substantively changing the way that the respondents are interpreting what is being measured (Millsap & Everson 1993).

Future directions include extending the simulation-based procedure to investigate the likelihood for a person from a specific θ to be reclassified. Data could be simulated per θ value, and a percentage of cases above and below the X cut score could be estimated. Also, it would be interesting to investigate how many categories are needed before discrete items could be treated as continuous (Rhemtulla et al. 2012). At that point, researchers would not need to use the simulation-based procedure and they could simply use the Millsap and Kwok (2004) procedure. Lastly, it would be interesting to incorporate the analysis of mixed item-types in the simulation-based procedure. The simulated respondents came from a 2PL model, but the procedure could be extended so that item parameters could be simulated from several models (e.g., some items coming from a 2PL model and

others from a GRM). Overall, we believe that the simulation-based procedure by Gonzalez and Pelham (in press) can be a useful complement to traditional significance-testing procedures for conducting a DIF or measurement bias study, and this chapter illustrates how to use the simulation-based procedure to examine the effect of measurement bias on screeners comprised of binary items. We encourage researchers to quantify the practical impact of measurement bias on screening decisions made based on their measures.

References

- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington: University of Vermont.
- Gadow, K. D., & Sprafkin, J. (2005). *Child and adolescent symptom inventory-4 revised (CASI-4R)*. Stony Brook: Checkmate Plus.
- Gonzalez, O. & Pelham, W. E. III. (in press). When does differential item functioning matter for screening? A method for empirical evaluation. *Assessment*. <https://doi.org/10.1177/1073191120913618>.
- Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., et al. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, 32, 959–976. <https://doi.org/10.1017/s0033291702006074>.
- Kim, G., DeCoster, J., Bryant, A. N., & Ford, K. L. (2016). Measurement equivalence of the K6 scale: The effects of race/ethnicity and language. *Assessment*, 23, 758–768. <https://doi.org/10.1177/1073191115599639>.
- Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling*, 58(1), 79–98.
- Lai, M. H., Kwok, O. M., Yoon, M., & Hsiao, Y. Y. (2017). Understanding the impact of partial factorial invariance on selection accuracy: An R script. *Structural Equation Modeling: A Multidisciplinary Journal*, 24, 783–799. <https://doi.org/10.1080/10705511.2017.1318703>.
- Lai, M. H., Richardson, G. B., & Mak, H. W. (2019). Quantifying the impact of partial measurement invariance in diagnostic research: An application to addiction research. *Addictive Behaviors*, 94, 50–56. <https://doi.org/10.1016/j.addbeh.2018.11.029>.
- Manly, J. J. (2006). Deconstructing race and ethnicity: implications for measurement of health outcomes. *Medical Care*, S10–S16. <https://doi.org/10.1097/01.mlr.0000245427.22788.be>
- Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, 95(4), 728. <https://doi.org/10.1037/a0018966>
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied psychological measurement*, 17(4), 297–334. <https://doi.org/10.1177/014662169301700401>
- Millsap, R. E. (2013, October 17–19). *The impact of violations of measurement invariance on selection: The discrete case*. Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, St. Pete Beach.
- Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9, 93–115. <https://doi.org/10.1037/1082-989X.9.1.93>.

- Murray, A. L., Booth, T., Auyeung, B., McKenzie, K., & Kuenssberg, R. (2019). Investigating sex bias in the AQ-10: A replication study. *Assessment, 26*, 1474–1479. <https://doi.org/10.1177/1073191117733548>.
- Ong, M. L., Youngstrom, E. A., JJX, C., Halverson, T. F., Horwitz, S. M., Storfer-Isser, A., Frazier, T. W., Fristad, M. A., Arnold, L. E., Phillips, M. L., Birmaher, B., Kowatch, R. A., Findling, R. L., & the LAMS Group. (2017). Comparing the CASI-4R and the PGBI-10 M for differentiating bipolar spectrum disorders from other outpatient diagnoses in youth. *Journal of Abnormal Child Psychology, 45*, 611–623. <https://doi.org/10.1007/s10802-016-0182-4>.
- 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*, 354–373. <https://doi.org/10.1037/a0029315>.
- Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological methods, 11*(4), 402. <https://doi.org/10.1037/1082-989X.11.4.402>
- Teresi, J. A., Stewart, A. L., Morales, L. S., & Stahl, S. M. (2006). Measurement in a multi-ethnic society: Overview to the special issue. *Medical Care, 44*(11 Suppl 3), S3. <https://doi.org/10.1097/01.mlr.0000245437.46695.4a>
- Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah: Lawrence Erlbaum.
- Youngstrom, E. A. (2013b). A primer on Receiver Operating Characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology, 38*(10), 1111–1119. <http://dx.doi.org/10.1093/jpepsy/jst062>

Reliability and Structure Validity of a Teacher Pedagogical Competencies Scale: A Case Study from Chile



Juan I. Venegas-Muggli 

Abstract The following paper examines the reliability and structure validity of a quantitative observational teacher pedagogical competencies scale implemented at one of Chile's largest higher education institutions. In a context in which new students accessing post-secondary education are challenging traditional teaching methods, the evaluation of this instrument is presented as a relevant case study for those interested in promoting teachers' pedagogical competencies. Reliability analyses considered the KR-20 coefficient and corrected item-total correlations. Structure validity was assessed through an exploratory factor analysis in which the concept's theoretical and latent structures were compared. The results suggest that the scale has high levels of internal consistency. Additionally, although the scale's theoretical and latent structures do not match exactly, relevant common elements are found. The considerations for applying these types of educational measurement instruments are discussed.

Keywords Pedagogical Competencies · Reliability · Validity · Higher Education

1 Introduction

Teachers' pedagogical competencies in the context of higher education have been an issue of great interest during the last decades. As higher education systems grow and become more diverse, there is increasing concern about the quality of teaching, with it being stated that universities do not only require academically well-prepared teachers and researchers but also pedagogically skilled educators (Apelgren and Giertz 2010; OECD 2010). Furthermore, the fact that current students accessing post-secondary education have been raised in a digital era has also heightened the importance of teachers' pedagogical competencies. These have become relevant to develop innovative teaching practices that are capable of engaging this new and

J. I. Venegas-Muggli (✉)
Universidad Tecnológica de Chile INACAP, Santiago, Chile
e-mail: jvenegasm@inacap.cl

more challenging type of student (Demirbilek 2015; Johnson et al. 2016; O'Flaherty and Phillips 2015).

Against this background, several methods to assess and promote post-secondary teachers' pedagogical competencies have been implemented worldwide. From national policies in which quality assurance agencies support institutions in order to enhance teaching quality to the implementation of schemes to promote good teaching practices within universities, different types of initiatives have had the shared goal of improving educators' skills (Fry and Ketteridge 2008; OECD 2010). Likewise, several institutions have developed a set of instruments to measure teachers' pedagogical competencies or teaching quality (see Apelgren and Olsson 2010; Berk 2005; Wingrove et al. 2017).

In this context, it is also observed that there are no clear guidelines for higher education institutions on how to develop and evaluate more complex instruments to assess teachers' pedagogical abilities. In light of this, this paper presents a reliability and structure validity analyses of an observational teacher pedagogical competencies scale developed at INACAP, one of Chile's largest higher education institutions. In this respect, the psychometric evaluation is presented as a relevant case study for those interested in applying these types of educational measurement practices.

2 Teacher Pedagogical Competencies in Higher Education

2.1 Defining Pedagogical Competencies

When reviewing the literature, it is possible to observe that various terms have been used to describe teachers' skills, including effective teaching behavior, teaching quality, teaching effectiveness, and pedagogical competencies (see Burnett and Meacham 2002; Maulana et al. 2017; Opdenakker et al. 2012; van de Grift et al. 2014). All these concepts make reference to how teachers apply their skills to promote learning among their students as it is argued that teacher behavior is effective when it has a significant influence on student outcomes such as academic achievement and academic engagement (Maulana et al. 2017; van de Grift et al. 2014).

From the existing perspectives on this issue, one of the most complete ones is the framework provided by Van de Grift, Maulana, and Helms-Lorenz. They describe six domains of teaching behavior that promote effective teaching in the context of primary and secondary education: safe and stimulating learning climate, efficient classroom management, clarity of instruction, activating learning, adaptive teaching, and teaching learning strategies (Maulana et al. 2017). Considering these dimensions, they sustain that some specific teacher behaviors that positively influence student outcomes are creating a relaxing learning atmosphere, ensuring that lessons begins and ends on time, giving clear instructions, and promoting pupils to learn actively (Maulana et al. 2017; van de Grift et al. 2014; van de Grift 2014).

2.2 Higher Education and Pedagogical Competencies

Even though teaching quality has been principally studied in primary and secondary education, it is also important to highlight its relevance for tertiary education. As previously highlighted, the expansion and diversification of higher education systems have promoted the necessity of universities of having pedagogically skilled lecturers. This is highlighted, for example, by the 2010 OECD report on quality teaching in higher education that sustains that higher education institutions should implement evaluation mechanisms to identify and promote good teaching practices in order to respond to increasing societal concerns about the quality of programs offered to students. Likewise, several other reports have provided guidelines to improve teaching practices in higher education (see Jenkins et al. 2003; Fry et al. 2008; Ryegård et al. 2010).

In terms of specific teaching practices to be promoted in higher education, an important match is observed between these reports and previously described teaching strategies for primary and secondary education. Perhaps the most distinctive element of higher education is the comparatively higher importance of promoting active learning strategies. This is the case as current tertiary students have been raised in a digital era. This makes them to be more challenging students, which requires teachers to develop more innovative practices in order to engage them (Demirbilek 2015; Johnson et al. 2016; O’Flaherty and Phillips 2015).

2.3 Measuring Pedagogical Competencies

Several attempts at determining the extent to which educators are able to encourage effective learning in students can be found. These have taken on a variety of different forms, from self-reported surveys about teachers’ abilities to teaching portfolios and more sophisticated quantitative observational methods (Apelgreen and Olsson 2010; Berk 2005; Wingrove et al. 2017).

Within this background, the first discussion is whether pedagogical competencies can be measured or not and which of its components can be measured. This is a relevant debate as several authors have highlighted that most instruments to measure effective learning present significant limitations. Tschannen-Moran and Woolfolk (2001) argue that the study of teacher efficacy presents relevant measurement problems. For example, in the case of those instruments that measure this construct through teachers’ self-efficacy beliefs, they sustain that they present social desirability bias. Likewise, Coe et al. (2014) sustain all teaching quality measures provide at best poor approximations of how much students actually learn. Among their main problems, they highlight two of them: having too specific measures and defining indicators that are too general to be empirically testable. Finally, Burnett and Meacham (2002) also reflect on the limitations associated with the process of measuring teacher quality. They argue that one of the main problems

is that over-simplistic approaches to assess teaching skills have led to a focus on elements that may not be necessary for effective teaching. Equally, it is criticized that instruments based on checking if teachers exhibit or not specific behaviors identified by theorists might not be accurate, as they do not consider students' perspectives concerning what it means to be a good teacher.

Within this discussion, this paper's position is that, despite all limitations associated with the process of measuring pedagogical competencies, it is worth using these types of instruments as they deliver valuable information. In this same line, it is considered that these measures have to be used carefully considering mostly a formative perspective. As highlighted by Coe et al. (2014), only if teaching effectiveness indicators have emphasis on feedback, support, and professional learning they may lead to improvements in student learning, although these indicators are in some ways insufficient.

Based on the previously stated, it is also argued that in order to obtain more valuable information in this context of measurement difficulties, observational scales are the most valid option. Van de Grift (2007) argues that although questionnaires are cheaper and more efficient, they are problematic, as they require correcting for socially desirable responses. Similarly, Garnett (1983) sustains that instruments that do not consider behavioral descriptors provide only a general framework and lack the structure required to guide instructional improvement.

Finally, in terms of these instruments' psychometric properties, it is important to highlight there is little evidence on how valid and reliable these teaching quality measurements are in higher education. Even though some studies have considered these subjects, most of them have only focused on teachers' self-reported instruments and/or students' surveys (see Ramsden 1991; Spooen et al. 2007). In addition, the evaluation of quantitative observational scales is more frequent in primary and secondary teaching (Burnett and Meacham 2002; Brookhart and Durkin 2003; O'Leary 2015). Thus, the evaluation of these types of instruments in higher educational contexts emerges as a relevant research field.

3 Method

3.1 Instrument

The instrument evaluated is part of INACAP's *Classroom Accompanying Program for Teachers* or *Programa de Acompañamiento Docente en Aula* (ADA), which is aimed at assessing teachers' pedagogical competencies. It was developed by the intuition considering a quantitative observational form in which teachers are evaluated according to 20 items using the following four-response category scale: "Totally Agree" (4), "Agree" (3), "Disagree" (2), and "Totally Disagree" (1).

Each item describes a statement with a positive pedagogical attitude or behavior (e.g., *He/she generates conditions that favor student motivation and/or openness to learn.*). To this effect, based on the supporting material that describes what each

response category means for each item, evaluators assign the teachers a value from the four-response category scale for each of the 20 items, depending on how they carry out their classes.

This instrument is based on INACAP's AAVC Pedagogical Method (Aprendizaje Activo, Vinculado y Colaborativo/Active, Engaged and Collaborative Learning) (Mundo INACAP 2018). This method considers that learning processes in the context of higher education adhere to three main principles. First, they have to promote active learning strategies. This means engaging students through innovative and meaningful actions in which they are encouraged to reflect on their own actions autonomously and to avoid absorbing knowledge passively and with resignation (Walder 2017). The second element is associated with encouraging students to engage with their social environment. If students are to be successfully integrated into the job market, they must connect with relevant stakeholders, such as productive sector organizations or local community groups (Jongbloed et al. 2008; Vernon and Ward 1999). Finally, the AVVC pedagogical method states that learning processes also have to be based on collaboration among students. This involves promoting cooperation between classmates and students from other study programs, since this generates a more interdisciplinary perspective (Bruffee 1993; Boud et al. 2014).

Based on this method, the *Classroom Accompanying Program for Teachers* developed a quantitative observational scale aimed at measuring the extent to which teachers have the pedagogical competencies required. The scale is formed by the following 5 dimensions:

- *Structure of the Class*: development of a class based on a defined structure whose purpose is initially explained and which activates previously learned contents.
- *Methodological Strategies*: application of didactic strategies in a framework in which the acquisition of knowledge is encouraged through participants exchanging experiences.
- *Pedagogical Resources*: use of resources relevant to the class's aim and linked to the didactics of the specialty, based on different sources of information.
- *Evaluative Process*: development of evaluations using explicit mechanisms. Feedback is given during the class and reflection among the students is encouraged.
- *Generic Competencies*: promotion of activities that strengthen the institution's hallmark competencies, integrating generic subject proficiencies into learning activities.

Table 1 presents the pedagogical competencies scale's theoretical structure:

3.2 Participants

This study considers information from 736 higher education teachers to whom this observational scale was applied during the second semester of 2017. All participants are teachers at INACAP, one of Chile's largest higher education institutions, with

Table 1 Theoretical structure of the pedagogical competencies scale

#	Dimension	Item
i_1	<i>Structure of the class</i>	He/she explains the purpose of the class according to what students expect to learn
i_2		He/she activates students' previous learning
i_3		He/she generates conditions that favor student motivation and/or openness to learn
i_4		He/she effectively uses the time available for the class
i_5		He/she implements learning activities relevant to the class's purpose
i_6		He/she contextualizes the learning and/or the contents of the class within the study program's career field and/or graduate profile
i_7		He/she ends the class by highlighting the main ideas
i_8		He/she generates opportunities for student comments and/or questions
i_9		He/she encourages students to carry out activities outside the classroom according to what they expect to learn
i_10	<i>Methodological strategies</i>	He/she promotes an environment conducive to learning and exchange among students
i_11		He/she develops activities using the "learning by doing" methodological approach in accordance with what students expect to learn
i_12		He/she promotes student disposition and responsibility in the learning process
i_13	<i>Pedagogical resources</i>	He/she uses resources relevant to the class's purpose
i_14		He/she effectively uses selected resources relevant to the class structure
i_15		He/she encourages the use of different sources of information promoted by INACAP
i_16	<i>Evaluative process</i>	He/she explains how and in what way the expected learning is evaluated
i_17		He/she delivers feedback during the class
i_18		He/she reinforces student reflection by incorporating mistakes as a means of learning
i_19	<i>Generic competencies</i>	He/she promotes activities that strengthen the development of INACAP's hallmark competencies
i_20		He/she integrates generic subject competencies into learning activities

approximately 120,000 students and 5000 teachers in 26 faculties throughout the whole country.

This instrument is applied yearly to both new teachers and teachers who obtained less than 75% in their teaching performance evaluation scores based on students' opinions the previous year. Observed classes correspond to what INACAP defines as *Milestone Subjects*. These subjects aim to integrate specialty, generic, and hallmark competencies into the practical learning central to each study program. Accordingly,

the number of each participant teacher's classes observed is defined by the amount of milestone subjects they teach each semester.

In the case of the semester considered for this study (the second semester of 2017), 1372 observation sessions were carried out. Other teachers from the same institution who had received an outstanding performance evaluation were trained to carry out the evaluations, each of which lasted 45 min, as per the length as an academic class at the institution.

In terms of the institutional context of the place this study was carried out, it is important to state that INACAP is a private, non-profit, integrated higher education system made up of three different institutions: a University, a Technical Formation Center, and a Professional Institute. It offers two-year technical degrees and bachelor's degrees. It is also a non-selective institution with no entry requirements apart from graduating from high school. This means the institution receives an important number of underrepresented social groups, especially first-generation students from low- and middle-income families. In this setting, teachers' pedagogical competencies are relevant, since they are particularly important for supporting the learning of socially underrepresented groups.

3.3 Data Analysis

In order to evaluate the reliability and structure validity of the described scale, different statistical analyses were applied. First, descriptive analyses were considered in order to examine response distributions for the scale's items. A second stage applied reliability indicators to the whole scale. Specifically, the Kuder-Richardson 20 (KR-20) coefficient was applied (Streiner 2003). This assessment was considered since preliminary analyses suggested that the items behave as binary variables, which led to the recoding of the initial four categories into two.

Finally, an exploratory factor analysis (EFA) was conducted to examine the instrument's structure validity (Byrne 1990). EFA was used instead of a Confirmatory Factor Analysis (CFA) given the preliminary nature of these analyses, as this scale has never been validated before. Following the advice of Bartholomew et al. (2008), a principal component analysis (PCA) was initially conducted to define the number of factors to be extracted. They recommend choosing the number of factors by considering the proportion of the total variation explained by the components (70–80%), the magnitude of eigenvalues (greater than 1), the form of the screen plot, and whether the components have useful interpretations. Then, an EFA was applied using the generalized least squares extraction method and by performing an oblique rotation (Oblimin). This rotation method was selected since certain dimensions of pedagogical abilities are hypothetically believed to be correlated. For these analyses, two-category recoded items were considered. For this purpose, a tetrachoric correlation matrix was initially estimated from the raw data set before conducting PCA and EFA using this matrix. All analyses were conducted using the software IBM SPSS Statistics 24.

4 Results

4.1 Descriptive Analyses

The first type of data to be examined is descriptive analyses. Specifically, the distribution of teachers' scores for each item on the four-category response scale is considered in order to explore both items' discriminating capacity and the fitness of the defined response scale.

Table 2 shows that, even though items allow for discrimination among evaluated teachers, differences mainly apply in their scores for the two highest categories of each item. For all items except item 15, the first two categories account for more than 90% of teachers' scores.

These results imply that the response scale did not work properly. Evaluators hardly considered the categories "Disagree" and "Totally Disagree" when evaluating teachers' pedagogical competencies. Thus, it seems plausible that instead of scoring the level of teachers' pedagogical skills on a 1- to 4-point scale, evaluators used this instrument as a dummy checklist.

Table 2 Distribution of scores of each item on the scale and reliability indicators

Items	Item distribution				Reliability indicators	
	Totally agree	Agree	Disagree	Totally disagree	KR-20 without item	Item-rest correlation
i_1	64.4%	27.6%	6.5%	1.5%	0.929	0.506
i_2	70.8%	24.3%	4.2%	0.8%	0.928	0.552
i_3	71.1%	25.4%	3.3%	0.2%	0.926	0.638
i_4	76.5%	20.2%	2.9%	0.4%	0.926	0.611
i_5	77.1%	18.4%	3.9%	0.6%	0.925	0.658
i_6	74.3%	18.8%	5.5%	1.4%	0.929	0.487
i_7	61.9%	28.1%	7.7%	2.3%	0.926	0.644
i_8	79.7%	17.5%	2.3%	0.5%	0.927	0.590
i_9	70.8%	24.6%	3.5%	1.1%	0.926	0.635
i_10	77.3%	19.8%	2.6%	0.3%	0.926	0.611
i_11	75.3%	19.4%	4.6%	0.7%	0.926	0.614
i_12	76.1%	20.7%	2.7%	0.5%	0.925	0.675
i_13	81.5%	16.5%	1.8%	0.2%	0.926	0.635
i_14	79.1%	18.3%	2.3%	0.4%	0.926	0.637
i_15	63.5%	25.3%	7.7%	3.6%	0.928	0.553
i_16	63.9%	27.0%	7.1%	2.0%	0.926	0.625
i_17	83.8%	14.1%	1.9%	0.2%	0.927	0.594
i_18	74.1%	21.9%	3.5%	0.6%	0.926	0.652
i_19	62.5%	30.0%	5.9%	1.6%	0.926	0.630
i_20	65.7%	27.1%	5.3%	1.9%	0.926	0.650
Total	–	–	–	–	0.927	0.610

Based on these assumptions, it was decided that the original scale should be recoded into a two-category response scale. Specifically, the categories “Agree”, “Disagree”, and “Totally Disagree” were grouped into a new category. This meant the items on the scale implemented in subsequent analyses were treated as binary variables or checklist indicators. Those who scored four points are defined as teachers who have the pedagogical competencies evaluated, while those who scored three, two, or one point are understood as educators who do not have sufficient levels of these.

4.2 Reliability Analyses

In Table 3, reliability indicators for each of the instrument’s sub-scales and for the scale as a whole are shown. As previously explained, given the fact that the original response scale was recoded into two-category items, reliability analyses considered new binary indicators for the scale. In this respect, Table 3 shows the KR-20 coefficient.

The results indicate that the instrument applied is highly reliable. When considering the scale as a whole, a KR-20 coefficient of 0.930 is obtained. Likewise, the sub-scales coefficients are all over 0.7.

A second way in which reliability was examined was by estimating the KR-20 coefficient when individual items were removed from the scale. Table 2 presents the KR-20 coefficient when removing each item from the scale. Given that this coefficient was 0.930 when calculated for the whole 20-item scale, it can be seen that no item’s removal improves the overall reliability of the pedagogical skills scale. In other words, the internal consistency of the scale cannot be improved by removing a single item, which shows that all the scale’s items contribute to a high level of reliability.

Table 2 also shows each item’s correlation with the sum of all remaining items—what is called corrected item-total correlation. The results show an average item-total correlation of 0.61 (ranging from $r = 0.487$ to $r = 0.675$). According to Clark and Watson (1995), this is an indication that the scale has good internal consistency, since adequate values for this measurement of narrow constructs (such as pedagogical abilities) should range between 0.4 and 0.5.

Table 3 KR-20 coefficient by scale dimensions

Dimension	N° of Items	KR-20
Structure of the class	9	0.855
Methodological strategies	3	0.741
Pedagogical resources	3	0.747
Evaluative process	3	0.739
Generic competencies	2	0.897
Total	20	0.930

4.3 *Exploratory Factor Analyses*

Having examined the scale's reliability levels, subsequent analyses focus on evaluating the pedagogical competencies scale's structure validity. To be precise, this section analyzes whether the scale's previously described theoretical structure (see Methods section) matches the latent structure of this construct as observed when applying factorial analyses.

Structure validity was assessed by applying EFAs to the 20 items on the pedagogical skills scale. A tetrachoric correlation matrix was initially estimated from the raw data set of binary indicators to then estimate factorial analyses using this matrix as the main input.

A PCA was carried out first to determine the number of factors to be extracted. From the PCA of the 20 items, it was decided that four factors should be extracted as the first four components of the PCA explained 77.8% of the variance. Additionally, even though factors 3 and 4 presented eigenvalues lower than 1 (0.835 and 0.746), this rule was relaxed according to Joliffe's (1972) advice that retaining components with eigenvalues higher than 0.7 is better than a cut-off point of 1. This was also decided based on the degree of interpretability of the components extracted.

Next, an EFA was carried out, extracting four factors and using the generalized least squares method and an oblique rotation (Oblimin). Concerning this solution, an item-selection analysis was carried out first. Using a criterion that eliminates items exhibiting low factor loadings (<0.4), it was suggested that five of the 20 indicators should be removed as they did not accurately fit the scale structure. Therefore, a new EFA was carried out without these items in order to obtain a cleaner and more rigid structure.

Table 4 shows the rotated matrix of the new 15-item solution. This reveals a much clearer structure which, when analyzed, does not suggest eliminating any item. When examining this latent structure, it can be observed that, even though it does not exactly match the defined theoretical structure, there are several relevant common elements that indicate that the evaluated instrument is valid. First, factors 2 and 3 exactly match the "Generic Competencies" and "Pedagogical Resources" theoretical dimensions. Likewise, all of the items in factor 4 belong to the "Class Structure" dimension and they all refer to elements associated with how the beginning of a class is structured.

Although factor 1 contains eight items from three different theoretical dimensions (Class Structure, Methodological Strategies, and Evaluative Process), it provides a useful interpretation that suggests the evaluated pedagogical skills scale is a valid instrument. When going through these items, it can be seen that they are all associated with promoting student engagement during classes, either by encouraging direct participation or by delivering information aimed at increasing the usefulness of the class for students.

In summary, when carrying out an EFA, it can be concluded that the concept of pedagogical competencies is structured into four dimensions. The first and most important dimension (as shown by the 64.4% of explained variance) is associated

Table 4 Rotated matrix (Oblimin) of factor loadings (15 items)^a

Items	Factor (explained variance %)			
	1 64,4%	2 7,0%	3 5,2%	4 4,8%
i_17	0.959			
i_18	0.815			
i_8	0.800			
i_7	0.711			
i_9	0.623			
i_16	0.617			
i_6	0.481			
i_10	0.446			-0.301
i_19		-0.948		
i_20		-0.835		
i_14			-0.935	
i_13			-0.762	
i_3				-0.946
i_2				-0.592
i_1				-0.382

^a Only factors loadings higher than 0.3 are shown

with how teachers encourage student engagement through both participation and by delivering useful and relevant information, which can be called “Student Engagement”. The second and third dimensions highly match to the previously defined theoretical concepts of “Generic Competencies” and “Pedagogical Resources”. Finally, a fourth dimension, understood as a sub-dimension of the subscale of “Structure of the Class”, emerges. This is specifically associated with how the beginning of a class is structured and can be called the “Class Introduction Structure”.

5 Discussion

In this paper, the reliability and structure validity of a teacher pedagogical competencies scale was examined. This instrument emerges as a relevant way of measuring pedagogical skills in a context in which new students accessing higher education and raised in a digital era are challenging traditional teaching methods that do not manage to engage them successfully (Demirbilek 2015; Johnson et al. 2016).

The results showed that the initially defined four-response category did not work properly. Thus, an initial recommendation is for observational scales measuring teachers’ skills to be based on checklist indicators. This has also been the most common way these phenomena have been studied in elementary and secondary teaching (see O’ Leary 2015), also justifying their replication in the higher education context.

In relation to reliability analyses, the evaluated scale showed high levels of internal consistency. Likewise, item-retest correlations were also high and it was seen that getting rid of an item did not increase the scale’s internal consistency, reinforcing the opinion that the scale developed is reliable.

Regarding the scale’s structure validity, the 20-item solution was not very clear, because five items presented low factor loadings in each of the four factors. Thus, a new model was fitted which only considered 15 items. Although this new solution did not exactly match the scale’s theoretical structure, it matched some of its most important elements. By suggesting certain considerations to guide how this instrument is applied, such as using binary items and discarding items that do not accurately fit the scale structure, it can be concluded that the evaluated scale is also a valid measure of teachers’ pedagogical skills.

In summary, the examined evidence provides relevant information for those interested in measuring pedagogical competencies in the context of higher education. Despite the validity results showing some discrepancies between the scale’s theoretical and latent structures, these analyses also deliver valuable information by revealing the complexity of measuring pedagogical abilities. Accordingly, one of the main contributions of this research is to encourage the development of more complex instruments to measure teachers’ pedagogical skills in higher education contexts where self-reported or student surveys have had a more predominant role (Berk 2005). Similarly, this paper also presents evidence to justify the proper

importance and viability of measuring teacher quality as it shown how the adequate application of an observational scale can be a significant asset when supporting teachers to improve their teaching abilities.

References

- Apelgren, K., & Giertz, B. (2010). Pedagogical competence – A key to pedagogical development and quality in higher education. In K. Apelgreen & T. Olsson (Eds.), *A Swedish perspective on pedagogical competence* (pp. 25–40). Uppsala: Uppsala University.
- Apelgren, K., & Olsson, T. (Eds.). (2010). *A Swedish perspective on pedagogical competence*. Uppsala: Uppsala University.
- Bartholomew, D. J., Steele, F., Galbraith, J., & Moustaki, I. (2008). *Analysis of multivariate social science data*. Boca Raton: Chapman and Hall/CRC.
- Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International journal of teaching and learning in higher education*, 17(1), 48–62.
- Boud, D., Cohen, R., & Sampson, J. (2014). *Peer learning in higher education: Learning from and with each other*. Oxon: Routledge.
- Brookhart, S. M., & Durkin, D. T. (2003). Classroom assessment, student motivation, and achievement in high school social studies classes. *Applied Measurement in Education*, 16(1), 27–54.
- Bruffee, K. A. (1993). *Collaborative learning: Higher education, interdependence, and the authority of knowledge*. Baltimore: Johns Hopkins University Press.
- Burnett, P. C., & Meacham, D. (2002). Measuring the quality of teaching in elementary school classrooms. *Asia-Pacific Journal of Teacher Education*, 30(2), 141–153.
- Byrne, B. M. (1990). Methodological approaches to the validation of academic self-concept: The construct and its measures. *Applied Measurement in Education*, 3(2), 185–207.
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309–319.
- Coe, R., Aloisi, C., Higgins, S., & Major, L. E. (2014). What makes great teaching? Review of the underpinning research. Center for Evaluation and Monitoring, Durham University.
- Demirbilek, M. (2015). Social media and peer feedback: What do students really think about using Wiki and Facebook as platforms for peer feedback? *Active Learning in Higher Education*, 16(3), 211–224.
- Fry, H., Ketteridge, S., & Marshall, S. (2008). *A handbook for teaching and learning in higher education*. New York: Routledge.
- Garnett, P. (1983). The use of the teacher performance assessment instruments for assessing pre-service secondary students in Western Australia. *The South Pacific Journal of Teacher Education*, 11(2), 40–53.
- Jenkins, A., Breen, R., & Lindsay, R. (2003). *Reshaping teaching in higher education: Linking teaching with research*. London: Kogan.
- Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., & Hall, C. (2016). NMC horizon report: 2016 higher (Education Edition). Austin: The New Media Consortium.
- Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: Artificial data. *Applied Statistics*, 21(2), 160–173.
- Jongbloed, B., Enders, J., & Salerno, C. (2008). Higher education and its communities: Interconnections, interdependencies and a research agenda. *Higher Education*, 56(3), 303–324.
- Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2017). Validating a model of effective teaching behaviour of pre-service teachers. *Teachers and Teaching*, 23(4), 471–493.
- Mundo INACAP. (2018). Docencia 2030. Retrieved from <http://portales.inacap.cl/revista-mundo-inacap/revista-mundo-inacap/docencia-2030>

- O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education, 25*, 85–95.
- O'Leary, M. (2015). *Classroom observation: A guide to the effective observation of teaching and learning*. Oxon: Routledge.
- OECD. (2010). *Learning our lesson: Review of quality teaching in higher education*. Paris: OECD Publishing.
- Opdenakker, M. C., Maulana, R., & den Brok, P. (2012). Teacher–student interpersonal relationships and academic motivation within one school year: Developmental changes and linkage. *School Effectiveness and School Improvement, 23*(1), 95–119.
- Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The course experience questionnaire. *Studies in Higher Education, 16*(2), 129–150.
- Ryegård, Å., Apelgren, K., & Olsson, T. (Eds.). (2010). *A Swedish perspective on pedagogical competence*. Uppsala: Uppsala University.
- Spooren, P., Mortelmans, D., & Denekens, J. (2007). Student evaluation of teaching quality in higher education: Development of an instrument based on 10 Likert-scales. *Assessment & Evaluation in Higher Education, 32*(6), 667–679.
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99–103.
- Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783–805.
- van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research, 49*(2), 127–152.
- van de Grift, W. J. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement, 25*(3), 295–311.
- van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation, 43*, 150–159.
- Vernon, A., & Ward, K. (1999). Campus and community partnerships: Assessing impacts and strengthening connections. *Michigan Journal of Community Service Learning, 6*(1), 30–37.
- Walder, A. M. (2017). Pedagogical innovation in Canadian higher education: Professors' perspectives on its effects on teaching and learning. *Studies in Educational Evaluation, 54*, 71–82.
- Wingrove, D., Hammersley-Fletcher, L., Clarke, A., & Chester, A. (2017). Leading developmental peer observation of teaching in higher education: Perspectives from Australia and England. *British Journal of Educational Studies, 66*(3), 1–17.



Joshua Chiroma Gandhi 

Abstract Psychoperiscope is a coined nomenclature which integrates periscope and psychometrics to mirror the mediating-moderating effect of cognitive coping strategies in the relationship between illness and quality of life. Coping refers to the effort toward mastering demands posed by harm, threat, or challenge being appraised and/or perceived as taxing available resources. It could be in terms of problem-focused versus emotion-focused as well as behavioral coping versus cognitive coping dimensions. The mediating-moderating effect of cognitive coping strategies, in the relationship between illness and quality of life, has not been clearly understood due to lack of a construct-relevant assessment scale. Therefore, this study developed a suitable scale using mixed methods embedded design. The mixed methods embedded design was opted for due to its advantageous measurement characteristics which would elucidate quality of life variance in relation to the effects of cognitive coping strategies on the variance of illness. Based on the Gandhi Psychometric Model, the term *psychoperiscope* was coined as a new psychometric nomenclature and adopted in this context as the scale name. Psychoperiscope was pilot-tested on a sample of 30*3 (i.e., $n = 30 \times 3$) participants, translated as consisting of 30 patients alongside their respective 30 family members and 30 clinical practitioners selected by the multistage sampling method. The final psychoperiscope, a 21-item (3-version) scale, proves significantly reliable for research and also serves as a valid screening tool. Following the useful data it elicited in this study, psychoperiscope would effectively generate more optimal and robust data if complemented with an experimental case study.

Keywords Quality of life · Psychoperiscope · Psychometrics · Mixed methods embedded design · Illness · Cognitive coping strategies

J. C. Gandhi (✉)

Department of General and Applied Psychology, University of Jos, Jos, Plateau State, Nigeria

The Psychometric Laboratories, Jos, Plateau State, Nigeria

1 Introduction

1.1 Background to the Study

Cognitions help us in regulating our emotion or any feeling in order not to be overwhelmed by the effects of negative or stressful life events. Since it seems the regulatory effect of cognitions would moderate life events, the perspective suggesting that cognitive processes can impact both illness and quality of life in some way is appropriate (in the circumstances). Garnefski et al. (2002) opine that “the regulation of emotions through cognitions is inextricably associated with human life.” Such cognitions include both conscious and unconscious cognitive processes. The unconscious cognitive processes, which include mental defence mechanisms such as projection, denial, daydreaming, and rationalization have been more significantly studied than the conscious ones which are popularly referred to as cognitive coping strategies (Garnefski et al. 2002).

Psychoperiscope is a coined nomenclature which integrates the principles of periscope and the ideals of psychometrics to more optimally define the mediating-moderating effect of cognitive coping strategies in the relationship between illness and quality of life. Coping, according to Monat and Lazarus (1991, p. 5), is defined as “an individual’s efforts to master demands (i.e., conditions of harm, threat, or challenge) that are appraised and/or perceived as exceeding or taxing available resources.” Coping could be classified in terms of problem-focused versus emotion-focused as well as behavioral coping (what you do) versus cognitive coping (what you know) dimensions (Gandi and Wai 2010; Garnefski et al. 2002). Considering that behavior includes actions, interactions, and reactions mostly in response to goals, motivations, needs, and problems, a complex interplay of psychological components has to form appropriate networks that determine and/or shape it. According to Epskamp et al. (2017), there are three measures of the network structure which include the strength, the closeness, and the between-ness of the node. Just as Lauritzen (1996) showed in the Gaussian graphical model, the node represents observed variables while the edges represent partial correlation coefficients between two variables after conditioning on all other variables in the dataset.

The strength quantifies how well a node is indirectly connected to other nodes, the closeness quantifies how well a node is directly connected to other nodes, and the between-ness quantifies how important a node is in the average path between two other nodes. These measures of the network structure (strength, closeness, and between-ness) corroborate and/or represent Schwartz and Rapkin’s (2004) three-stage quality of life (QOL) measurement model which include performance-based, perception-based, and evaluation-based measurements. QOL measures are either designed on assumptions that measurement scales are consistently used while scores are directly comparable across people over time or even designed to account for response shift phenomena. Schwartz and Rapkin (2004) insist on the inferred evidential suggestion supporting response shift phenomena that the underlying

processes of appraisal differ across people and over time. This can greatly affect how most of the QOL scale items are responded to. It could also be inferred, from the viewpoint of Schwartz and Rapkin (2004), that a more optimal assessment scale will be best suited if the ideals of clinimetrics and psychometrics are taken into consideration as an integrative whole.

Clinimetrics, which was conceptualized as the science of clinical measurement (Fava et al. 2011), refers to the domain concerned with indexes and other experiences that are used to describe or measure symptoms, physical signs, and any distinctly clinical phenomena (Mayo 2015; Feinstein 1987). Clinimetrics aims to develop seemingly heterogeneous measures with good face validity for clinical common sense and, therefore, principally rely on the opinions of patients and clinicians (Feinstein 1987; Upton and Upton 2007). Psychometrics, which on the other hand ensures quality and valued degree of homogeneity, rely more on statistical techniques and generally aims to develop measures that are mathematically valid and reliable (Upton and Upton 2007). Despite the obvious differences, there is some overlap “*ab initio*” between the ideals of clinimetrics and the techniques of psychometrics, which makes for easy complementary integration of their principles. To effectively integrate the ideals/techniques of clinimetrics (the science of clinical measurement) and psychometrics (the science of psychological measurement), an appropriate choice of suitable and more optimal research design seems cogently helpful.

De Vaus (2001) and Yin (2014) believe that any design that uses a more logical and comprehensive approach to investigating the research problem ensures that the evidence(s) obtained enables us to answer initial research questions as unambiguously as possible. Such designs would have to be adequately representative, by integrating different techniques, to accommodate various peculiarities toward attenuating/controlling extraneous (or confounding) variable effects. Combining different techniques this way has been described in terms of multimethod and mixed methods designs. A multimethod research involves combining multiple elements of either qualitative techniques or quantitative techniques, while mixed methods research involves combining the elements of both qualitative and quantitative techniques in one study. The major specific designs of mixed methods research include triangulation design, embedded design, explanatory design, and exploratory design (Creswell and Plano Clark 2007).

Mixed methods triangulation design aims at obtaining different but contemporary data on the same topic to best resolve the research problem (Morse 2003). It brings together the differing strengths and nonoverlapping weaknesses of quantitative methods (symbolized as QUAN: large sample size, trends, and generalization) with those of qualitative methods (symbolized as QUAL: small N, details and in depths). The small N (in this case) refers to one group or single-subject (single-case) designs that are particularly based on qualitative approach. Hence, mixed methods triangulation design is simply symbolized as QUAN + QUAL. It is used to directly compare and contrast quantitative results with qualitative findings (Schoonenboom and Johnson 2017).

Another mixed methods design is the embedded design which includes one data set that can provide a supportive secondary role in a study based primarily on the other data type (Creswell 2003). This was corroborated by Schoonenboom and Johnson (2017) who believe that it has been premised on three facts: that a single data set is not sufficient, that different questions need to be answered, and that each type of question requires different types of data. The need to include qualitative data to answer a research question within a largely quantitative study and vice versa is a cogent justification for using the mixed methods embedded design. Thus, it embeds a qualitative (qual) component within a quantitative (QUAN) design (symbolized as QUAN + qual), it compares quantitative (QUAN) and qualitative (QUAL) designs (symbolized as QUAN + QUAL), and then embeds a quantitative (quan) component within a qualitative (QUAL) design (symbolized as QUAL + quan). Therefore, the mixed methods embedded design is generally symbolized as QUAL + quan, QUAN + QUAL, and QUAN + qual.

The next mixed methods design is referred to as mixed methods explanatory design which is a two-phase design whose overall purpose is completely dependent on the fact that qualitative (qual) data helps explain or build upon initial quantitative (QUAN) results (Creswell 2003; Schoonenboom and Johnson 2017). It is most suitable for any study that requires qualitative data to explain three findings: (a) significant or nonsignificant results, (b) outlier results, and (c) surprising results (Morse 2003). The mixed methods explanatory design is, therefore, symbolized as QUAN + qual. The design (QUAN + qual) can be used to form groups based on quantitative results and follow-up with the groups through subsequent qualitative research or using quantitative participant characteristics to guide purposeful sampling for a qualitative phase (Creswell 2003).

The last mixed methods design is exploratory design, a two-phase design, in which results of the qualitative (QUAL) method can help develop or inform the quantitative (quan) method (Greene et al. 1989). It has been premised on the fact that an exploration is cogently needed for one of the several reasons which include nonavailability of instruments, unknown variables, lack of a guiding framework, or the need for appropriate theories. Because the mixed methods exploratory design (symbolized as QUAL + quan) begins qualitatively, it is most suitable for exploring a phenomenon. The design (QUAL + quan) is particularly useful in developing and testing new instruments (Creswell 2014) and for identifying and quantitatively studying important unknown variables. Mixed methods exploratory design is also appropriate in case of the need to generalize results to different groups, to test aspects of an emergent theory (or classification), or to explore a phenomenon in-depth and measure its prevalence (Creswell 2014; Morgan 1998).

Whatsoever may be the case, designs must be construct-relevant in order to help minimize or even avoid drawing incorrect causal inferences from data (De Vaus 2001). The mixed methods designs (triangulation, embedded, explanatory, and exploratory) have been found to appropriately approach investigating problem(s) in various logical ways that mostly lead to correct causal inferences. The overall goal of these designs has been to expand and strengthen a study's conclusions and make more empirical contributions to the published literature. This has been

more significantly demonstrated by the way and manner any mixed methods design answers research questions more effectively in empirical ways than the other designs. Johnson and Christensen (2014) subscribe to this by corroborating that the mixed methods approach heightens knowledge by providing sufficient quality to achieve more legitimate multiple validities. Although exploratory design is said to be particularly useful in developing and testing new instruments, the embedded design is found to be more useful and optimally suitable in developing and testing new instruments (especially 3-version scales) for complex or multifaceted mixed methods assessments. According to Schoonenboom and Johnson, the mixed methods embedded design has additional advantage of being implemented either sequentially or concurrently as the case may be. It is, therefore, more optimally suitable for empirical studies than other designs.

1.2 Statement of Problem and Purpose of the Study

It has been observed that “if the meaning of quality of life (QOL) rating depends upon any underlying appraisal processes, the relationship between the observed item and the underlying latent true score is far more complicated than assumed” (Schwartz and Rapkin 2004). This, as Gandi and Wai (2010) inferred, has been more obvious in some scale for assessing the impacts of cognitions in emotion regulation to determine QOL. Hence invoking the principles of performance-based, perception-based, and evaluation-based measurements is expected to lend credence to and adequately ensure a construct-relevant scale. While the clarion call by Upton and Upton (2007) to integrate clinimetric and psychometric strategies in developing a multi-item health outcome measure is apt, its suitability needs to be tested by integrating the principles of performance-based, perception-based, and evaluation-based measurements as a model.

The purpose of the study was because the cogent need to assess the mediating-moderating effects of cognitive coping strategies in the relationship between illness and quality of life has been hampered by lack of a suitable construct-relevant scale over the years. Therefore, the study was designed to develop and validate a suitable scale for assessing the mediating-moderating effects of cognitive coping strategies in the relationship between illness and perceived quality of life. Thus, the envisioned scale was developed as a measurement tool that assesses the mediating-moderating role(s) of cognitive coping strategies in the relationship between illness and quality of life.

1.3 Conceptual Framework

The new scale was conceptualized by and premised on lucid integration of Gandi Psychometric Model (2018) and Schwartz and Rapkin’s (2004) model (see Figs. 1 and 2). To clarify the complicated relationship between observed items and

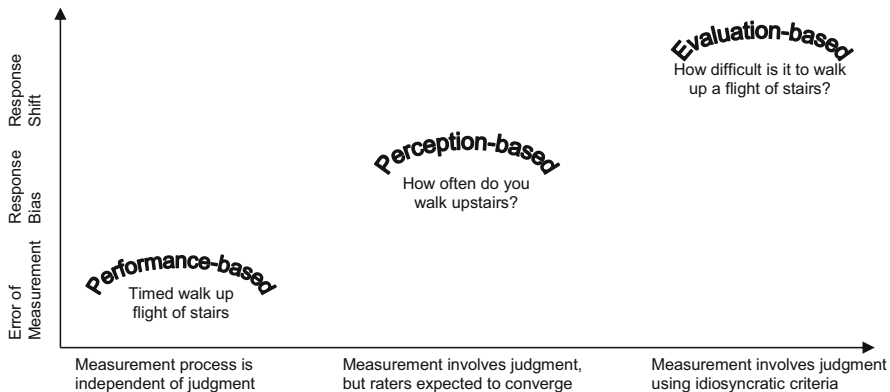


Fig. 1 Clarifying the discrepancy in performance-based, perception-based, and evaluation-based methods. (Adopted with copyright permission from the authors Schwartz and Rapkin 2004)

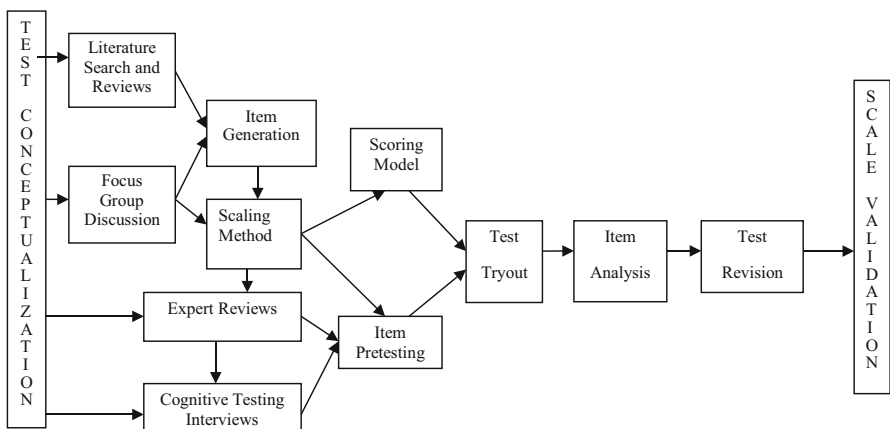


Fig. 2 Scale Development Framework based on Gandi Psychometric Model. (Adopted from Gandi 2018)

the underlying latent true score as the appraisal processes upon which the QOL rating depends, Schwartz and Rapkin (2004) propounded a model which integrated performance-based, perception-based, and evaluation-based measurement dimensions as presented in Fig. 1.

The performance-based, which yields measures reflecting the quantity and quality of effort, is independent of judgment and more susceptible to error of measurement. Perception-based, which yields measures of individual judgment concerning the occurrence of an observable phenomenon, involves judgment while raters are expected to converge due to response bias. Evaluation-based, which yields measures rating experience as positive or negative compared with an internal standard, involves judgment using idiosyncratic criteria that enhances cogent merits of response shift.

A test development conceptual framework, based on the nine stages according to Gandhi Psychometric Model (2018), which facilitates answering the “how” and “why” question(s), is presented in Fig. 2. These nine stages include test conceptualization, item generation, scaling methods, item pretesting, scoring models, test tryout, item analysis, test revision, and scale validation and standardization. It sets out clearly “how” the main stages through which the test development process moves (i.e., from left to right) and also reflects the systematic sequence of the process, i.e., from test conceptualization to scale validation (Fig. 2), including “why” particular stage(s) or variable(s) precedes and/or succeeds the other (Gandi 2018, 2019).

The creative integration of Figs. 1 and 2 (Schwartz and Rapkin 2004; Gandhi 2018) led to forming the coined term “psychoperiscope,” which is a combination of psychometrics and periscope, deliberately conceptualized and adopted as the scale name. While psychometrics refers to the science of psychological measurement, periscope is an instrument that consists of a tube attached to a set of mirrors or prisms by which an observer can see things that are otherwise out of sight (Soanes and Stevenson 2007). Mental periscope refers to the ability of the intellect to observe, understand, and initiate appropriate action(s) in which the self can re-energize, examine, reflect, and refine, or just be completely still. When the intellect uses its capacity as a periscope, it can find a balance between the inside and the outside worlds.

This integrative conceptual framework forms the premise upon which the envisioned psychoperiscope, a 3-version scale, aims to more effectively be mirroring the mediating-moderating effects of cognitive coping strategies in the relationship between illness and quality of life. More optimal designs such as mixed methods embedded designs and a 3-step formula for sample size determination would positively corroborate the conceptual framework.

2 Methods

2.1 Research Design and Study Setting

The study adopted a multifaceted mixed methods design, referred to as mixed methods embedded design, in which Creswell (2003) pointed out that “one data set provides a supportive secondary role in a study primarily based on the other data type.” It was premised on the fact that “a single data set is not sufficient, emphasizing that different questions need to be answered, and that each type of question requires different types of data.” The need to include qualitative data to answer a research question within a largely quantitative study and vice versa is a cogent justification for using the mixed methods embedded design (Creswell 2003; Schoonenboom and Johnson 2017). It embeds a qualitative (qual) component within a quantitative (QUAN) design (QUAN + qual), compares a quantitative

(QUAN) design with a qualitative (QUAL) design (QUAN + QUAL), and embeds a quantitative (quan) component within a qualitative (QUAL) design (QUAL + quan). Hence, the design (i.e., mixed methods embedded design) has been symbolized as QUAL + quan, QUAN + QUAL, and QUAN + qual which is implementable, both concurrently and sequentially, for developing 3-version scale (such as psychoperiscope). Mixed methods embedded design was also adopted because of the considered suitability for elucidating quality of life variance in relation to effects of cognitive coping strategies on the variance of illness. This corroborates the embedded design's optimal suitability for eliciting three sets of data (from three sources) on the same target subjects-of-assessment.

The study was conducted at Jos University Teaching Hospital (JUTH) in Plateau State of Central Nigeria. JUTH's diversity added impetus to the resulting data in terms of helping to "adequately prevent and avoid any perceived social desirability or other unwanted influence(s) that could amount to raping the psychometric quality of the scale" under consideration (Gandi 2019). This is because diverse professional and ethnic peculiarities as well as different ideological leanings within JUTH (the study setting) helped in achieving a study sample that more optimally met the requirements for adequate participant representativeness. Jos city is a miniature Nigeria and one of the settlements in the country where men and women exist as co-equals, without a significant gender bias or discrimination. This also had positive implication for research participation and the collected data characteristics.

2.2 Target Population and Sample Participants

The study essentially targeted clinical population, comprising patients alongside their family members and respective clinical practitioners, at the Jos University Teaching Hospital (JUTH). It must be noted that patients are the primary target participants (i.e., the subjects of assessment) for whom the new scale (herein referred to as psychoperiscope) was developed. The participating patients' family members (spouse, parent, child, sibling, or others) and the clinical practitioners (doctor, nurse, or psychologist), as significant others in this case, serve the purpose of providing relevant data to adequately complement, supplement, and even validate the individual patient's respective responses.

Since psychoperiscope is a 3-version scale that can generate data from three sources and is suitable for studies that adopt mixed methods embedded design, the pilot sample size consisted of 30*3 participants. Thus, it comprises 30 patients (the target subjects of assessment) alongside 30 family members (family source of embedded data) and 30 clinical practitioners (professional source of embedded data), respectively. This sample size (30*3) was systematically determined by forming a computation formula that determines the appropriate sample size for any mixed methods design that adopts concurrent data collection procedure, such as the mixed methods embedded design. The steps of the integrated formula include: determining sample size for infinite population (n_i), determining attenuating sample

size (n_a) adjusted to facilitate avoiding nonresponse effects (n_e), and converting the attenuating adjusted sample size (n_a) to a sample size for finite population (n_f).

First step – Determining sample size for infinite population by using confidence level, population proportion (P), and error margin (E):

$$n_i = \frac{Z^2 \times P \times (1 - P)}{E^2}, \quad (1)$$

where in this case $Z = 2.576$ which is the Z value corresponding to adopted confidence level which was set to 99%, P was assumed to be 50%, and E was set to 1% in this case.

Second step – Determining attenuating sample size adjusted to facilitate avoiding nonresponse effects:

$$n_a = n_i + \frac{n_i}{1 - n_e}, \quad (2)$$

where n_e was set to 5%.

Third step – Determining sample size for finite population by using the adjusted sample size (a_n):

$$n_f = 1 + \frac{n_i}{\frac{n_i - 1}{N}} \quad (3)$$

where N = Population size.

Using the aforementioned adopted formulae (1), (2), and (3) helped to systematically determine the study sample size as 30*3 participants, which translates as 30 patients alongside 30 family members and 30 clinical practitioners, respectively. The 30 selected patients included male ($n = 15$) and female ($n = 15$) aged 16–73 years across different ethnicity, religion, education levels, occupations, and socioeconomic status.

The 30*3 participants were selected by multistage sampling across the study setting, Jos University Teaching Hospital (JUTH). Although it is a more complex method, the choice of multistage sampling premised on reliability and validity of its combined techniques. Trochim and Donnelly (2008, p. 47) describe multistage sampling as a method that combines several probability sampling techniques to create a more reliable and efficient or effective sample than the use of just any one sampling type can achieve on its own. The sampling techniques that constituted the multistage sampling method for this study include cluster sampling technique (phase 1) and stratified random sampling technique (phase 2). Cluster sampling helped in determining the specific study sites (i.e., units/wards) within the hospital and then stratified random sampling helped in selecting the individual study participants, i.e., patients (the subjects of assessment) alongside family members (the family embedded source of data) and clinical practitioners (the professional embedded source of data) at each of the participating units/wards.

2.3 *Materials and Procedure*

Materials

Psychoperiscope, a 21-item scale primarily developed for research and screening, consists of three versions namely version A (the patient or target participant version), version B (the patient family member version), and version C (the clinical practitioner version). Materials used in developing psychoperiscope have been similar to the instruments and conditions used in Rumor Scale Development chapter by Gandhi (2019). Thus, the materials for psychoperiscope development include informed consent forms, interview schedule forms, demographic data forms, focus group discussions checklist, expert reviews rating rubrics, cognitive testing feedback sheets, video camera, writing materials, SPSS software, and the processing/analysis system (computer). The conditions considered as significant materials in the study include the basic necessary and sufficient conditions as well as a great deal of miscellaneous conditions that Mackie (1965) referred to as “insufficient but nonredundant part of unnecessary but sufficient (inus) conditions”.

Procedure

Psychoperiscope was developed by applying the nine stages of the Gandhi Psychometric Model, which include test conceptualization, item generation, scaling methods, item pretesting, scoring models, test tryout, item analysis, test revisions, and scale validation (Gandhi 2018). Having conceptualized psychoperiscope to be developed as a 3-version scale (patient version, family version, and clinician version), using deductive and inductive methods generated a pool of 64 items as an item bank. The first 30 items were deductively derived from literature review on focal constructs and target population as well as from systematic review of existing related scales. The next 34 items were inductively devised by conducting focus group discussions, in-depth interviews, and personal brainstorming across potential stakeholders. Thereafter, all the 64 items in the item bank were subjected to deliberate pretesting with the aid of expert reviews and cognitive testing interviews which refined them for more relevance and suitability. Only 28 items survived the process while 36 items were deleted at this stage for want of suitability. Likert-type scale has been the adopted scaling method alongside its corresponding scoring model for the resulting 28 items that survived the preceding rigorous pretesting reviews.

In implementing the test tryout, required research ethical clearances were earned based on certificate in human subjects' research course as well as the social and behavioral research curriculum completion for collaborative institutional training initiative (CITI) was appropriately fulfilled. All essential ethical considerations, such as voluntariness, confidentiality, autonomy, avoiding even minimal risk, ensuring individual privacy, and other required conditions for studies with human participants, were observed and consciously adhered to. Prior to the designed protocol implementation, research field assistants ($n = 5$) and research confederates ($n = 5$) were systematically recruited (one at each of the respective study

units/wards) among the health professionals. Individual informed consent(s) were duly obtained from each participating patient as well as their respective family members and clinical practitioners separately. The recruited field assistants and research confederates, who facilitated the pilot study protocol implementations alongside the researcher, lend credence to appropriate task of data collection. After administering the 28 items as a self-response scale, all the completed forms were retrieved while individual participants were being appreciated for their time and kind participation.

The completed and retrieved scale forms were systematically collated and then coded, preparatory for onward analysis as required. The analysis techniques then emphasized item reliability index, difficulty index, discrimination index and validity index which together ensured adequate soundness of the scale (Gandi 2019). Just as Gandi (2019) earlier noted, the item difficulty index and discrimination index were qualitatively determined (based on item pretesting process), while the statistical analysis (quantitative methods) emphasized ensuring reliability index and validity index of the retained items. The analyses conducted include content validity index (CVI), item-total statistics, Pearson's correlation analysis, and exploratory factor analysis (EFA). A befitting threshold of item minimum excellent significance level, set at 0.60, was adopted while all the pilot data analyses were respectively carried out at $p \leq 0.05$.

3 Results

Results of the study have shown significant reliability and validity for 21 items retained out of the pilot-tested 28 items that were subjected to analysis. Item correlation coefficients, based on Pearson, $r = 0.62\text{--}0.70$ ($p < 0.05$), had an overall average Cronbach's alpha as $\alpha = 0.66$. Thus, the raw Cronbach's and standardized Cronbach's alpha values were found to be 0.62 and 0.70, respectively.

The scale mean of the means (3.90) and mean of the variances (0.46) as well as the variance of means (0.19) and variance of variances (0.24) all corroborated its reliability, as shown in Table 1 (summary item statistics). Likewise, Table 2 shows that the mean (261.59), variance (574.96), and standard deviation (23.98) have constituted very good scale statistics.

Item-total statistics, which checks for any item(s) that might be inconsistent with average behavior of others, was analyzed in order to safely discard inconsistent item(s). As shown in Table 2, the item-total statistics analysis results reflected scale

Table 1 Summary item statistics

	Mean	Minimum	Maximum	Range	Maximum variance	Minimum variance
Item means	3.90	3.00	4.67	1.67	1.56	0.19
Item variances	0.46	0.00	2.88	2.87	888.79	0.24

Table 2 Scale version A for target participants (subjects of assessment) item-total statistics

Case	Item	Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (α)
1	I have basically done nothing to prevent my illness before now	22.99	22.02	0.68	0.70
2	I cannot change anything about the present situation of my illness	22.84	20.13	0.62	0.69
3	I am not able to sustain constructive thoughts because of my illness	22.96	21.52	0.70	0.75
4	My inactivity (or decreased activity level) affects me negatively	22.98	22.10	0.56	0.78
5	I cannot perform even my simplest regular tasks beyond a maximum of 30 min	22.91	21.49	0.50	0.79
6	Anything that requires physical strength is not for me	23.00	21.12	0.44	0.80
7	I can perform most daily tasks without any assistance	23.06	20.45	0.68	0.72
8	I feel I am the one to blame for not being able to overcome my illness situation	22.89	21.04	0.49	0.72
9	I feel that I have a responsibility to ensure improvement in my wellbeing	23.10	20.21	0.70	0.68
10	I think I can learn something from the illness	22.95	22.00	0.57	0.73
11	I think that I will recover and even be a stronger person than ever	22.79	20.10	0.62	0.72
12	I think it could have been much worse, but thanks for how it is now	22.89	21.62	0.56	0.77
13	I think that other people go through much worse experiences with their health	22.99	22.10	0.60	0.66
14	I continually think how horrible my health situation has been	22.98	20.38	0.58	0.75
15	It has been difficult for me to cope with my illness	23.20	21.22	0.43	0.70
16	It was a mild illness but has deteriorated overtime	23.02	20.52	0.62	0.68
17	I still go about my regular activities, even without assistance, despite the illness	22.88	21.00	0.56	0.70
18	My illness was worse than it is now	22.87	22.21	0.57	0.76
19	I now have more insight into my situation than before	22.90	21.53	0.58	0.88
20	The treatment(s) I have received (or am receiving) improve my health and quality of life	22.98	20.64	0.44	0.69
21	My relationship with others (family, clinicians, colleagues, friends etc.) enhances my wellbeing	22.89	20.12	0.67	0.70

mean if item deleted, correlated item-total correlation, and Cronbach's alpha if item deleted.

By investigating the item-total correlation, seven items with low correlations (below required alpha values) were dropped from the preceding 28 items to retain 21 items that correlated highly (0.60 and above). Table 2 shows the scale mean, if item deleted, for all the retained 21 cases with an average of 22.99 for the duly summated items. The scale variance, if item deleted, was summed up for all the 21 cases as 22.93, to be the variance of the summed items. By exploring alpha, and having deleted any or all of the low correlated items, the reliability of the scale would increase to 0.88 in either case (Table 2).

Table 2 shows that the corrected item-total correlation has provided empirical evidence to the extent that only few items correlated at low values which, undoubtedly, translated to the fact that just few items are construct-irrelevant in this case and have been deleted. The changes in Cronbach's alpha (for the retained 21 items) if any of the items were deleted have effectively supported and corroborated the corrected item-total correlations by indicating high correlations (as presented in Table 2) ab initio.

Table 3 presents scale mean, if item deleted, for all the retained 21 cases, averagely as 23.41, for the duly summated items. The scale variance, if item deleted, were summed up, for all the 21 cases as 23.02, to be the variance of the summed items. It presented the overall item-total statistics, which is known to check for any item(s) that might be inconsistent with average behavior of other items, as analyzed in order to ascertain the measure by discarding any inconsistent item(s). The item-total statistics result appropriately reflected scale mean if item deleted, correlated item-total correlation, and Cronbach's alpha if item deleted.

By investigating the item-total correlation (Table 3), seven items with low correlations (below required alpha values) were discarded from the preceding 28 items to retain 21 items that correlated highly (0.60 and above). By exploring alpha, and having deleted any or all of the low correlated items, the reliability of the scale would increase to 0.82 in either case (Table 3). The corrected item-total correlation provided empirical evidence that only few items correlated at low values, indicating that only few items are construct-irrelevant in this case. As seen in Table 3, the changes in Cronbach's alpha if any items were deleted have corroborated the corrected item-total correlations by indicating high correlations.

The mean scale shown in Table 4, if item deleted, for all the 21 cases has an average of 22.97 for the duly summated retained items. The scale variance (Table 4), if item deleted, was summed up for all the 21 cases as 22.93 which is the variance of the summed 21 retained items. By evaluating total correlations in Table 4, it could be noticed that the retained 21 items have respectively satisfied the desired psychometric requirements. There has been no significant variation among or between the respective patient family members on the presented results.

Likewise, by evaluating the total correlations in Table 4, it could be noticed that the 21 retained items have respectively satisfied the desired psychometric requirements. There has been no significant variation among or between clinical practitioners on the presented results.

Table 3 Scale version B for family members item-total statistics

Case	Item	Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (α)
1	Our sick person have basically done nothing to prevent his/her illness before now	23.42	22.12	0.68	0.74
2	He/she cannot change anything about the present situation of his/her illness	22.98	21.30	0.62	0.68
3	He/she reported inability to sustain constructive thoughts because of the illness	23.66	21.60	0.70	0.69
4	His/her inactivity (or decreased activity level) affects him/her negatively	23.50	22.76	0.56	0.78
5	He/she cannot perform even his/her simplest regular tasks beyond a maximum of 30 min	23.71	21.90	0.50	0.77
6	Anything that requires physical strength is not for him/her	23.22	21.32	0.44	0.82
7	He/she can perform most daily tasks without any assistance	23.20	20.50	0.68	0.80
8	He/she seem to feel that he/she is the one to blame for not being able to overcome the illness situation	22.89	21.41	0.49	0.72
9	He/she reported that he/she have a responsibility to ensure improvement in his/her wellbeing	23.15	22.31	0.70	0.68
10	He/she seem to believe that he/she can learn something from the illness	23.46	21.02	0.57	0.73
11	He/she have faith that he/she will recover and even be a stronger person than ever	23.57	20.18	0.62	0.72
12	He/she said it could have been much worse, but thanks for how it is now	23.00	21.72	0.56	0.79
13	He/she admits that other people go through much worse experiences with their health	23.02	22.30	0.60	0.66
14	He/she continually seem to think of how horrible his/her health situation has been	22.99	20.31	0.58	0.75
15	It has been (and still seems) difficult for him/her to cope with the illness	23.30	20.61	0.43	0.70
16	His/her illness initially seems mild but has deteriorated overtime	23.21	20.80	0.62	0.68
17	I observe that he/she still go about his/her regular activities, even without assistance, despite the illness	23.00	21.04	0.56	0.70

(continued)

Table 3 (continued)

Case	Item	Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (α)
18	To me, his/her illness was worse than it is now	22.97	22.31	0.57	0.75
19	He/she now seems to have more insight into his/her situation than before	22.98	21.45	0.58	0.80
20	The treatment(s) he/she have received (or am receiving) seems to improve his/her health and quality of life	23.32	21.00	0.44	0.68
21	It seems his/her relationship with us and with others (clinicians, friends etc.) enhances his/her wellbeing	22.98	20.64	0.67	0.72

Table 4 Scale version C for clinical practitioners item-total statistics

Case	Item	Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (α)
1	The patient reported that he/she did nothing to prevent the illness before now	22.99	22.02	0.68	0.70
2	He/she now seems helpless and cannot change his/her illness situation	22.84	20.13	0.62	0.69
3	Patient seems unable to sustain constructive thoughts because of his/her illness	22.96	21.52	0.70	0.75
4	His/her inactivity (or decreased activity level) seems to affect him/her negatively	22.98	22.10	0.56	0.78
5	Patient has not been able to perform even his/her simplest regular tasks beyond a maximum of 30 min	22.91	21.49	0.50	0.79
6	I think anything that requires physical strength is not for this patient	23.00	21.12	0.44	0.80
7	He/she can perform most daily tasks even without any assistance	23.06	20.45	0.68	0.72
8	Patient reported feeling he/she is the one to blame for not being able to overcome the illness situation	22.89	21.04	0.49	0.72
9	He/she reported that it is his/her responsibility to ensure improvement in his/her wellbeing	23.10	20.21	0.70	0.68
10	Patient seems to think he/she can learn something from the illness	22.95	22.00	0.57	0.73

(continued)

Table 4 (continued)

Case	Item	Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (α)
11	He/she reported thinking that he/she will recover and even be a stronger person than ever	22.79	20.10	0.62	0.72
12	Patient said it could have been much worse, but thanks for how it is now	22.89	21.62	0.56	0.77
13	He/she admits that other people go through much worse experiences with their health	22.99	22.10	0.60	0.66
14	He/she keep reporting thinking how horrible his/her health situation has been	22.98	20.38	0.58	0.75
15	It has been (and still seems) difficult for this patient to cope with the illness	23.20	21.22	0.43	0.70
16	The patient's illness initially seems mild but has deteriorated overtime	23.02	20.52	0.62	0.68
17	I observe that he/she still go about his/her regular activities, even without assistance, despite the illness	22.88	21.00	0.56	0.70
18	Patient reported that the illness was worse than it is now	22.87	22.21	0.57	0.76
19	He/she now reports having more insight into his/her situation than before	22.90	21.53	0.58	0.80
20	The treatment(s) patient have received (or am receiving) seems to improve his/her health and quality of life	22.98	20.64	0.44	0.69
21	It seems his/her relationship with us and with others (family, colleagues, friends) enhances his/her wellbeing	22.89	20.12	0.67	0.70

So far, it would be noticed that the respective Tables (i.e., Tables 2, 3, and 4) have reflected item-total statistics, showing all the retained 21 items with significant reliability. Items that were not consistent with how other items behaved were checked for and summarily deleted from the table for their nonsuitability. The reliability, after deleting inconsistent items from the tables, was based on scale mean, scale variance, corrected item-total correlation, and Cronbach's alpha.

Notwithstanding the findings from item-total statistics analysis, a purification based on exploratory factor analysis (EFA) was further conducted for the purification of all the scale items. This has addressed the respective items' skewness and kurtosis (Table 5). Thus, the retained 21 items were further explored and found to have significantly score values that adequately satisfied the skewness and kurtosis requirements.

Table 5 Exploratory analysis of mean, standard deviation, skewness, and kurtosis

	Maximum	Mean	Std. deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. error	Statistic	Std. error
1	5	3.36	0.96	-0.65	0.04	-1.39	0.08
2	5	3.68	0.95	0.64	0.04	-1.54	0.08
3	5	3.35	0.49	0.75	0.04	-1.15	0.08
4	4	3.00	0.81	-0.01	0.04	-1.47	0.08
5	5	4.65	0.50	-1.13	0.04	1.06	0.08
6	5	4.33	0.47	0.72	0.04	-1.49	0.08
7	5	3.82	0.59	-2.54	0.04	5.30	0.08
8	4	3.34	0.47	0.50	0.04	-1.51	0.08
9	4	3.67	0.47	-0.72	0.04	-1.49	0.08
10	4	3.33	0.47	0.72	0.04	-1.49	0.08
11	4	3.02	0.82	-0.04	0.04	-1.50	0.08
12	4	3.67	0.48	-0.79	0.04	-1.14	0.08
13	5	4.32	0.47	0.68	0.04	-1.29	0.08
14	5	4.33	0.47	0.74	0.04	-1.46	0.08
15	5	4.33	0.47	0.74	0.04	-1.46	0.08
16	5	4.33	0.47	0.74	0.04	-1.46	0.08
17	5	4.33	0.47	0.74	0.04	-1.46	0.08
18	5	4.33	0.47	0.73	0.04	-1.47	0.08
19	5	4.02	0.82	-0.04	0.04	-1.50	0.08
20	5	4.02	0.82	-0.03	0.04	-1.51	0.08
21	5	4.38	1.06	-1.51	0.04	1.34	0.08

The skewness and kurtosis (Table 5) consideration in the study requires that any items with absolute values of more than three for skewness and then less than eight for kurtosis are psychometrically inadequate, therefore not satisfactory for inclusion in the items to be retained. Consequent upon this, the exploratory purification retained all the 21 items on the basis of satisfying the criteria for inclusion.

4 Discussion

The study, which was designed to develop a suitable scale for assessing the mediating-moderating effects of cognitive coping strategies in the relationship between illness and quality of life, gave rise to psychoperiscope as a 21-item (3-versions) scale. Using the nine stages of test development, based on Gandhi psychometric model (2018), both deductive and inductive approaches as well as expert reviews and cognitive interviews were critically implemented. Gandhi psychometric model corroborates the perspectives of De Vaus (2001) and Yin (2014),

which hold that any design that uses more logical and comprehensive approach to investigating problem ensures that the evidence(s) obtained enables us to answer initial research questions as unambiguously as possible. Such designs would have to be adequately representative, by integrating different techniques, to accommodate various peculiarities toward attenuating/controlling extraneous (or confounding) variable effects. Combining different techniques in dramatic ways at such magnitude and intensity have been described in terms of multimethod and mixed methods designs. A multimethod research involves combining multiple elements of either a qualitative technique or quantitative technique while mixed methods research involves combining the elements of both qualitative and quantitative techniques in one study. The robust procedural nitty-gritty lends credence to psychoperiscope development.

Mixed methods embedded design was adopted in developing and pilot-testing the new scale because, according to Creswell (2014), it is the most rigorous procedure for collecting and analyzing data as well as interpreting and reporting the study findings that emanate from the data. The four major mixed methods designs (triangulation, embedded, explanatory, and exploratory designs) have been found psychometrically suitable in their respective rights. However, the embedded design which was specifically adopted for the study under review has best addressed the research problem toward achieving psychoperiscope development. This is because, in mixed methods embedded design, one data set provides a supportive secondary role where the study is also based on other data type or source. Creswell (2014) premised this on the empirical fact that single data set is not sufficient, that different questions need to be answered, and that each type of question requires different types of data. As a 3-version scale, the resulting psychoperiscope elicits three data sets on the same participant (i.e., the same subject of assessment) because one data set was considered insufficient.

Whatsoever may be the case, designs must be construct-relevant in order to help minimize or even avoid drawing incorrect causal inferences from data (De Vaus 2001). The mixed methods designs (triangulation, embedded, explanatory, and exploratory) have been found to appropriately approach the problem investigation in various logical ways that mostly lead to correct causal inferences. The overall goal of these designs has been to expand and strengthen a study's conclusions and make more empirical contributions to the published literature. This has been more significantly demonstrated by the way and manner any mixed methods design answers research questions more effectively in empirical ways than the other designs. Johnson and Christensen (2014) subscribe to this by corroborating that the mixed methods approach heightens knowledge by providing sufficient quality to achieve more legitimate multiple validities. Although exploratory design is said to be particularly useful in developing and testing new instruments, the embedded design is found to be more useful and optimally suitable in developing and testing new instruments (especially 3-version scales) for complex or multifaceted mixed methods assessments.

According to Schoonenboom and Johnson, the mixed methods embedded design has additional advantage of being implemented either sequentially or concurrently as the case may be. It is, therefore, more optimally suitable for empirical studies than other designs. To actualize the envisioned goal of developing psychoperiscope using mixed methods embedded design in practical terms, a suitable formula for sample size determination had to be formulated as part of the study. The systematically formulated computation formula helped in determining appropriate sample size for the study which adopted the mixed methods embedded design, using concurrent data collection procedure. The steps of this formula include: determining sample size for infinite population (n_i), determining attenuating adjusted sample size (n_a) to facilitate avoiding nonresponse effects (n_e), and converting the attenuating adjusted sample size (n_a) to a sample size for finite population (n_f).

The qualitative and quantitative approaches, which together answer research questions based on embedded design more adequately, were systematically implemented at every stage of the psychoperiscope development process. Thus, the embeddedness is twofold: (i) the quantitative components were embedded within qualitative design and (ii) the qualitative components were embedded within quantitative design. The mixed methods embedded design was specifically opted because of its advantages and more optimal suitability in designing performance-based, perception-based, and evaluation-based measures that the psychoperiscope represents in QOL assessment. For instance, embeddedness (by its very nature) helps check and minimize any possibility of social desirability that seems to define the weakness of some other scales.

Qualitative analysis, supported by Lawshe's (1975) content analysis, helped to ensure significant item content validity which translated to having a suitable construct-relevant scale. Quantitatively, item-total statistics helped to check for item(s) that might be inconsistent with the average behavior of other items on the scale as it were. This analysis (item-total statistics) was a huge contribution that suggested safe discarding of specific items with significant inconsistent characteristics in relation to the perceived relevant items. This was confirmed based on the correlated item-total correlation and the Cronbach's alpha if item deleted which provided empirical evidence suggesting that the items which correlated at low values are construct-irrelevant and deserved to be discarded. All the three sets of data that were analyzed have reflected a similar scenario in the presented findings thereof.

Notwithstanding the findings from item-total statistics analysis, a further purification by conducting exploratory factor analysis has added significant impetus to the scale items. Thus, exploratory factor analysis has effectively addressed the skewness and kurtosis outlook of respective items on the scale. This presented the retained items as only those having significant score values that have adequately satisfied the skewness and kurtosis requirements. Consequent upon this, therefore, the overall analyses have finally retained all the 21 items on the basis of satisfying the criteria for inclusion.

References

- Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks: Sage.
- Creswell, J. W. (2014). *A concise introduction to mixed methods research*. Thousand Oaks: Sage.
- Creswell, J. W., & Plano Clark, V. (2007). *Designing and conducting mixed methods research*. Thousand Oaks: Sage.
- De Vaus, D. A. (2001). *Research design in social research*. London: Sage.
- Epskamp, S., Borsboom, D., & Fried, E. I. (2017). Estimating psychological networks and their accuracy. *Behaviour Research Methods*, *50*(1), 195–212. <https://doi.org/10.3758/s13428-017-0862-1>.
- Fava, G. A., Tomba, E., & Sonino, N. (2011). Clinimetrics: The science of clinical measurement. *The International Journal of Clinical Practice*, *66*(1), 11–15. <https://doi.org/10.1111/j.1742-1241.2011.02825.x>.
- Feinstein, A. R. (1987). *Clinimetrics*. New Haven: Yale University Press.
- Gandi, J. C. (2018). *Development and validation of health personnel perceived quality of life scale* (Unpublished doctoral thesis). University of Ibadan, Nigeria.
- Gandi, J. C. (2019). Rumor scale development. In M. Wiberg, S. Culpepper, R. Jansen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative psychology. IMPS 2018 Springer proceedings in mathematics and statistics* (Vol. 265, pp. 429–447). Cham: Springer.
- Gandi, J. C., & Wai, P. S. (2010). Impact of partnership in coping in mental health recovery: An experimental study at the Federal Neuro-Psychiatric Hospital Kaduna. *International Journal of Mental Health Nursing*, *19*(5), 322–330.
- Garnefski, N., van den Kommer, T., Kraaij, V., Teerds, J., Legerstee, J., & Onstein, E. (2002). The relationship between cognitive emotion regulation strategies and emotional problems. *European Journal of Personality*, *16*, 403–420.
- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Towards a conceptual framework for mixed methods evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274.
- Johnson, R. B., & Christensen, L. (2014). *Educational research: Quantitative, qualitative, and mixed approaches*. Washington, DC: Sage.
- Lauritzen, S. L. (1996). *Graphical models*. Oxford: Clarendon Press.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575.
- Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly*, *12*, 245–265.
- Mayo, N. E. (2015). *Dictionary of quality of life and health outcomes measurement* (1st ed.). Milwaukee: International Society for Quality of Life Research (ISOQOL).
- Monat, A., & Lazarus, R. S. (1991). *Stress and coping: An anatomy*. New York: Columbia University Press.
- Morgan, D. L. (1998). *Focus group guidebook*. Thousand Oaks: Sage.
- Morse, J. M. (2003). Principles of mixed methods and multimethod research design. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods research* (pp. 189–208). Thousand Oaks: Sage.
- Schoonenboom, J., & Johnson, R. B. (2017). How to construct a mixed methods research design. *Kolner Z Soz Sozpsychology*, *69*(Suppl 2), 107–131.
- Schwartz, C. E., & Rapkin, B. D. (2004). Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. *Health and Quality of Life Outcomes*, *2*(16), 1–11.
- Soanes, C., & Stevenson, A. (2007). *Concise oxford English dictionary* (11th edition revised (3rd impression)). Oxford: Oxford University Press.
- Trochim, W. M. K., & Donnelly, J. P. (2008). *The research methods knowledge base* (3rd ed.). Cincinnati: Atomic Dog Publishing (a part of Cengage Learning).
- Upton, D., & Upton, P. (2007). A psychometric approach to health related quality of life measurement: A brief guide for users. In *Leading-edge psychological tests and testing research* (pp. 71–89). New York: Nova Science.
- Yin, R. K. (2014). *Case study research design and methods* (5th ed.). Thousand Oaks: Sage.

Modeling Household Food Insecurity with a Polytomous Rasch Model



Victoria T. Tanaka, George Engelhard Jr, and Matthew P. Rabbitt

Abstract The Household Food Security Survey Module (HFSSM) is an 18-item scale created and maintained by the US Department of Agriculture (USDA) that measures food insecurity in the United States. The HFSSM includes ten items that reference food hardships among adults in the household and eight items that reference food hardships among children. The scale was created and maintained using a dichotomous Rasch model (Engelhard et al., *Educ Psychol Meas* 78:1–19, 2017). However, the item responses that are collected for nine of the items are polytomous that are later dichotomized for creating the final scale. In 2006, the Committee on National Statistics (CNSTAT) reviewed the HFSSM and the USDA’s procedures for measuring food insecurity. They suggested modeling polytomous item responses with a polytomous model instead of dichotomizing item responses (Wunderlich and Norwood, *Food insecurity and hunger in the United States: an assessment of the measure*. The National Academies Press, Washington, DC, 2006). The purpose of this study is to explore modeling polytomous HFSSM items with a partial credit model, building on Nord’s (Assessing potential technical enhancements to the US household food security measures. US Department of Agriculture, Economic Research Service, 2012) work on the partial credit model and the HFSSM. The polytomous Rasch model is compared to the dichotomous Rasch model currently used by the USDA. The data suggest that the polytomous model provides better model-data fit, explaining 62% of the variation as opposed

The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or US government determination or policy. This research was supported in part by the US Department of Agriculture, Economic Research Service.

V. T. Tanaka (✉) · G. Engelhard Jr
The University of Georgia, Athens, GA, USA
e-mail: vtanaka@uga.edu; gengelh@uga.edu

M. P. Rabbitt
U.S. Department of Agriculture, Economic Research Service, Washington, DC, USA
e-mail: matthew.rabbitt@usda.gov

to 58% with the dichotomous model. The use of a polytomous model increases the precision of the estimates of food insecurity.

Keywords Rasch model · Partial credit model · Household food insecurity

Food security exists when all members of a household have access to the food that they need for a healthy, active lifestyle (Coleman-Jensen et al. 2018). Every year since 1995, the US Department of Agriculture (USDA) measures household food insecurity at the national level with the Household Food Security Survey Module (HFSSM), an 18-item scale that is administered as a supplement to the Current Population Survey (CPS). The HFSSM was created and has been maintained with the Rasch measurement model (Engelhard et al. 2017). It is used to estimate national food insecurity prevalence rates, and influences food and nutrition policy-making decisions. Although the HFSSM includes polytomous items, item responses that are collected with the HFSSM are dichotomized for analysis. This study examines the use of a partial credit model for modeling the polytomous item responses, building on previous studies that examine the use of alternate models with the HFSSM (Nord 2012), and furthering the understanding of the psychometric properties of this scale. The challenges and implications of modeling polytomous item responses when measuring household food insecurity are also considered.

1 The Household Food Insecurity Survey Module

The full scale of the HFSSM consists of 18 items that describe the behaviors of households that face difficulty meeting their food needs. The first ten items (Items 1–10) reference food hardships among adults in the household, generally. The last eight items (Items 11–18) reference food hardships among children below the age of 18 in the household. Half of the HFSSM items elicit dichotomous responses (Yes/No) and half elicit polytomous, frequency of occurrence responses that are later dichotomized for analysis. Some of the polytomous items are follow-up questions for the dichotomous items that precede them. For example, item 5 is a follow-up item that asks

(If yes to question 4) How often did this happen—almost every month, some months but not every month, or in only 1 or 2 months?

where a response of “almost every month” or “some months but not every month” is coded as a Yes, and “in only 1 or 2 months” is coded No. These frequency of occurrence items are a source of local dependency that violates the assumptions of the Rasch model. However, Nord (2012) formally assessed the consequences of this local dependency for measurement of latent food insecurity and found that the effect of failing to adjust for local dependency has a negligible practical effect. To mitigate concerns about local dependency within our data, we combined the

base and frequency follow-up questions into trichotomous items. We found similar results to Nord's (2012) investigation, concluding that not accounting for the local dependency has negligible potential implications for our study.

In 2006, the Committee on National Statistics (CNSTAT) reviewed and assessed the HFSSM, noting "frequency and duration are . . . important elements for the USDA to consider in the . . . measurement of household food insecurity and individual hunger" (Wunderlich and Norwood 2006, p. 4). This statement underscores recommendation 5–1, which recommends exploring alternatives to the dichotomous Rasch model by modeling the polytomous item responses rather than dichotomized item responses (Wunderlich and Norwood 2006). Nord (2012) assessed this recommendation, finding that neither the polytomous nor dichotomous Rasch model were more strongly preferred, though the dichotomous model has advantages over the polytomous model including its "transparency and ease of explanation" (p. 25).

The purpose of this study is to explore the use of a polytomous Rasch model with the HFSSM data following the recommendations made by CNSTAT (Wunderlich and Norwood 2006) and building on previous research in this area (Nord 2012). The following research questions are addressed:

1. How does a polytomous Rasch model fit responses to the HFSSM compared to the dichotomous Rasch model that the USDA currently uses?
2. What benefits exist to adopting a polytomous Rasch model over the dichotomous Rasch model for food security measurement?
3. What are the implications for food security measurement and prevalence estimates when using a polytomous rather than dichotomous Rasch model?

Moving to a polytomous Rasch model has the potential to lead to gains in measurement precision and improved classification of household food security status. Therefore, this study adds to the existing body of literature on food security measurement and classification by considering CNSTAT's recommendation for enhancing measurements made with the HFSSM.

2 Methodology

2.1 Participants

The food security survey module the USDA designed is created in such a way that households that are unlikely to have indicators of food insecurity are not screened into the HFSSM. Therefore, our sample consists of low-income households more likely to experience food insecurity. We also pooled several years of cross-sectional data to ensure the statistical power of our analyses. This study included all households who provided valid responses to the HFSSM in 2014–2016, who had at least one child under the age of 18, and who were also below 185% of the federal poverty line ($N = 11, 511$). We use the income threshold of 185% of the federal

poverty line because it is the income screening threshold for a household to be administered the HFSSM. In the CPS, households with income above 185% of the federal poverty line that showed no signs of food stress are not administered the HFSSM to reduce respondent burden. Households with income above 185% of the federal poverty line that show signs of food stress are administered the HFSSM, but they represent a small proportion of the households administered the HFSSM (Engelhard et al. 2017; Nord 2012) and are omitted from our sample. For the Rasch analyses, all households that had extreme scores—either 0 or 18 for the dichotomous analyses and 0 or 26 for the partial credit analyses—were removed from the data set. In psychometric and economic analyses of food insecurity in the United States, reductions in sample size of this scale are not uncommon because of the screening of households into the HFSSM. We had a final sample size of $N = 6606$. All analyses were completed in the Rasch software, Facets (Linacre 2015).

2.2 Coding

Nord (2012) recommended several scales based on the structure and use of the scale:

1. Polytomous scale: Responses were coded 0 for “never,” 1 for “sometimes” or “yes, in only 1 or 2 months,” 2 for “often” or “yes, in some months but not every month,” and 3 for “yes, in almost every month.” Dichotomous items were coded 0 for “no” and 1 for “yes.”
2. Ever during the year scale: Responses were coded 0 for “never” or “no” and 1 for “sometimes,” “often,” or “yes.”

In the dichotomous analysis, responses are coded 0 for “no,” “never true,” or “yes, in only 1 or 2 months,” and 1 for “yes,” “often true,” “sometimes,” “yes, in almost every month,” “yes, in some months but not every month,” and “yes, in only 1 or 2 months.” In the partial credit analysis, responses are coded 0 for “no” or “never true,” 1 for “yes,” “often true,” or “yes, in almost every month,” 2 for “sometimes” or “yes, in some months but not every month,” and 3 for “yes, in only 1 or 2 months.”

2.3 The Dichotomous Rasch Model

The dichotomous Rasch model is used by the USDA to calibrate the HFSSM annually. This model describes the probability of a household endorsing an item of the HFSSM as a function of the household’s latent food insecurity and the difficulty of the item. In log-odds form, it is expressed as

$$\ln \left(\frac{P_{ni1}}{P_{ni0}} \right) = \theta_n - \delta_i \quad (1)$$

where P_{ni1} is the probability of household n endorsing item i , P_{ni0} is the probability of household n not endorsing item i , θ_n is the latent food insecurity measure (logit-scale location) of household n , and δ_i is the severity measure (logit-scale location) of item i (Rasch 1960/1980). Households and items are ordered along a line that represents household food insecurity: Households with greater food insecurity are expected to endorse more—and more difficult—items. Similarly, more difficult items are expected to be endorsed by fewer—and more food insecure—households. The Rasch model also meets the requirements for invariant measurement. These requirements are:

Household (person) measurement

1. The measurement of households must be independent of the particular items that happen to be used for the measuring: *Item-invariant measurement of households*.
2. A household with greater food insecurity must always have a better chance of affirming an item than a household with less severe food insecurity: *Non-crossing household response functions*.

Item calibration

3. The calibration of the items must be independent of the particular households used for calibration: *Household-invariant calibration of test items*.
4. Any household must have a better chance of affirming a less severe item than a more severe item: *Non-crossing item response functions*.

Unidimensionality

5. Items and households must be simultaneously located on a single underlying latent variable: *Wright map* (Engelhard 2013).

2.4 The Partial Credit Rasch Model

The partial credit Rasch model allows for the possibility of different numbers of response levels for different items on the same test—for example, the HFSSM, which has both dichotomous (Yes/No) items and polytomous (frequency of occurrence) items. Although the partial credit model allows for varying numbers of response levels, it is essential that these response levels are still ordered in such a way that an increase in score represents an increase in food insecurity (Engelhard and Wind 2018). The partial credit Rasch model also provides individual threshold estimates for each item. In log-odds form, it is expressed as:

$$\ln \left(\frac{P_{nik1}}{P_{nik0}} \right) = \theta_n - \delta_{ik} \quad (2)$$

where P_{nik1} is the probability of household n endorsing category k for item i , P_{nik0} is the probability of household n not endorsing category k for item i , θ_n is the latent food insecurity measure (logit-scale location) of household n , and δ_{ik} is the difficulty measure (logit-scale location) of category k for item i (Engelhard and Wind 2018). More response categories provide more information, which, in turn, provides greater measurement precision (Bond and Fox 2005).

3 Results

3.1 Dichotomous Results

The Wright map (Fig. 1) locates households and items on a logit-scale line that represents the latent construct, household food insecurity. Higher scores correspond to greater food insecurity, for households, and fewer endorsements, for items. The distribution of households is positively skewed, though the HFSSM items have good spread along the line. The child-referenced items tended to be more difficult for respondents to endorse. Summary results are presented in Table 1. The dichotomous Rasch model explained 58.48% of the variance in the data. As indicated by the Wright map, households had a low average measure of -2.63 , indicating low average food insecurity in this sample. Infit and outfit were acceptable for households and items. Reliability was fairly high for both (0.82 and greater than 0.99, respectively). Table 2 presents the household fit statistics summary. Infit was considered unproductive or distorting of measures for approximately 22.7% of households. Outfit was unproductive or distorting for about 9.6% of households.

The item summary is presented in Table 3. Items 1, 2, and 3 were the easiest, overall, for respondents to endorse. These items were household-level items that asked respondents about their anxiety about food running out, their access to the resources necessary to obtain more food, and their ability to afford balanced meals. Items 16, 17, and 18 were the most difficult for respondents to endorse. These items were child-referenced items that asked if children ever had to skip a meal and how often that occurred, and if children were ever unable to eat for an entire day. Overall, infit was good for all items. Outfit was poor for items 9, 10, 15, 16, 17, and 18, and bad for items 1, 2, 3, and 11. It should be noted that items 1, 2, and 3 are the first items of the adult-referenced items, and item 11 is the first of the child-referenced items.

Mear	+Household	-Items
6		
5	.	Child(ren) not eat for whole day
4	.	Child(ren) skipped meals frequency follow up Child(ren) skipped meals
3	.	Adult(s) not eat for whole day frequency follow up Adult(s) not eat for whole day Cut size of child(ren)'s meals
2	.	Respondent lost weight Child(ren) not eating enough Respondent hungry but did not eat
1	.	Adult(s) cut size or skipped meals frequency follow up
0	**	Adult(s) cut size or skipped meals Respondent ate less than should have
-1	**	Relied on low-cost foods for children Could not afford to eat balanced meals
-2	**	Food bought would not last
-3	****	Worried food would run out
-4	*****	
-5	*****	
-6	*****	
Mear	n = 138	-Items

Fig. 1 Wright map of the dichotomous analysis of the Household Food Security Survey Module (HFSSM) items. Households and items are located on a logit-scale line that represents the latent construct, household food insecurity. Higher scores correspond to greater food insecurity, for households, and fewer endorsements, for items

Table 1 Summary statistics for the dichotomous analysis

	Household	Items
Measure		
Mean	-2.63	0.00
<i>SD</i>	2.14	2.85
Outfit		
Mean	0.70	0.78
<i>SD</i>	1.10	0.52
Infit		
Mean	1.00	0.97
<i>SD</i>	0.55	0.17
Separation statistic	2.13	40.40
Reliability of separation	0.82	> 0.99
χ^2 (<i>df</i>)	38607.9* (6605)	44850.1* (17)
<i>N</i>	6606	18
Variance explained by the Rasch model	58.48%	

Table 2 Summary of household fit statistics

Label	Description	Range	Dichotomous Rasch		Polytomous Rasch	
			Infit MSE	Outfit MSE	Infit MSE	Outfit MSE
A	Productive for measurement	$0.50 \leq \text{MSE} < 1.50$	3117 47.2%	2154 32.6%	3290 49.8%	2864 43.4%
B	Less productive for measurement, but not distorting of measures	$\text{MSE} < 0.50$	1984 30.0%	3817 57.8%	1791 27.1%	3276 49.6%
C	Unproductive for measurement, but not distorting of measures	$1.50 \leq \text{MSE} < 2.00$	1152 17.4%	206 3.1%	1114 16.9%	183 2.8%
D	Unproductive for measurement, distorting of measures	$\text{MSE} \geq 2.00$	353 5.3%	429 6.5%	411 6.2%	283 4.3%

Note. MSE is the mean square error

3.2 Partial Credit Results

In the Wright map (Fig. 2), households and items are once again located on a logit-scale line that represents household food insecurity, where higher scores on the latent construct correspond to greater food insecurity for households and greater difficulty for items. The partial credit Wright map also includes the rating scale structure of the polytomous items of the HFSSM. The distribution of households

Table 3 Summary of the dichotomous and partial credit Rasch item analyses

Item	Dichotomous Rasch				Polytomous Rasch			
	Measure	SE	Infit	Outfit	Measure	SE	Infit	Outfit
1	-5.33	0.04	1.01	1.72	-3.17	0.03	1.11	1.12
2	-3.95	0.03	0.95	1.53	-2.08	0.03	1.02	1.01
3	-3.29	0.03	1.17	1.51	-1.77	0.03	1.24	1.26
4	-1.53	0.04	0.72	0.50	-1.84	0.03	0.66	0.44
5	-1.71	0.03	0.79	0.63	-2.01	0.03	0.77	0.59
6	-0.52	0.04	1.31	1.11	-1.12	0.02	0.96	0.60
7	-0.09	0.04	0.87	0.56	-0.53	0.04	0.83	0.53
8	1.07	0.05	1.02	0.55	0.53	0.05	0.94	0.48
9	1.63	0.06	0.89	0.33	1.04	0.06	0.82	0.31
10	2.25	0.07	1.02	0.39	0.87	0.03	1.23	0.55
11	-3.01	0.03	1.31	1.58	-1.60	0.03	1.34	1.47
12	-1.57	0.04	1.11	1.04	-0.46	0.03	1.11	1.02
13	0.28	0.04	1.01	1.04	0.75	0.04	1.05	1.03
14	1.59	0.06	0.93	0.53	1.01	0.05	0.95	0.53
15	2.36	0.07	0.79	0.33	1.71	0.07	0.87	0.36
16	3.10	0.09	0.81	0.25	2.41	0.09	0.84	0.29
17	3.52	0.10	0.87	0.31	1.86	0.05	1.12	0.62
18	5.20	0.20	0.92	0.11	4.39	0.19	1.02	0.17

is still positively skewed and the HFSSM items still have good spread along the line, with the child-referenced items tending to be more difficult for respondents to endorse. The rating structure did vary from item to item. A summary table of the analysis is presented in Table 4. The partial credit Rasch model explained 62.87% of the variance in the data. Households had an average measure of -2.91 (SD = 1.93). Infit and outfit were acceptable, and reliability was fairly high for both households and items (0.83 and greater than 0.99 respectively). Overall, the category statistics indicated the rating scales for the polytomous items functioned as expected. Household fit statistics summary information is presented in Table 2. Infit was considered unproductive or distorting of measures for approximately 23.1% of households. Outfit was unproductive or distorting for about 7.1% of households (Table 2).

The item summary for the partial credit Rasch analysis is also presented in Table 3. Items 1, 2, and 3 were, again, the easiest items for respondents to endorse. Items 16, 17, and 18 were the most difficult for respondents to endorse, with item 18 being almost twice as difficult as item 16, the second-most difficult item. Infit was good for all items overall. Outfit was poor for items 4, 8, 9, 15, 16, and 17. This represents an improvement in fit over the dichotomous Rasch model that had more instances of misfit, and several cases of serious misfit.

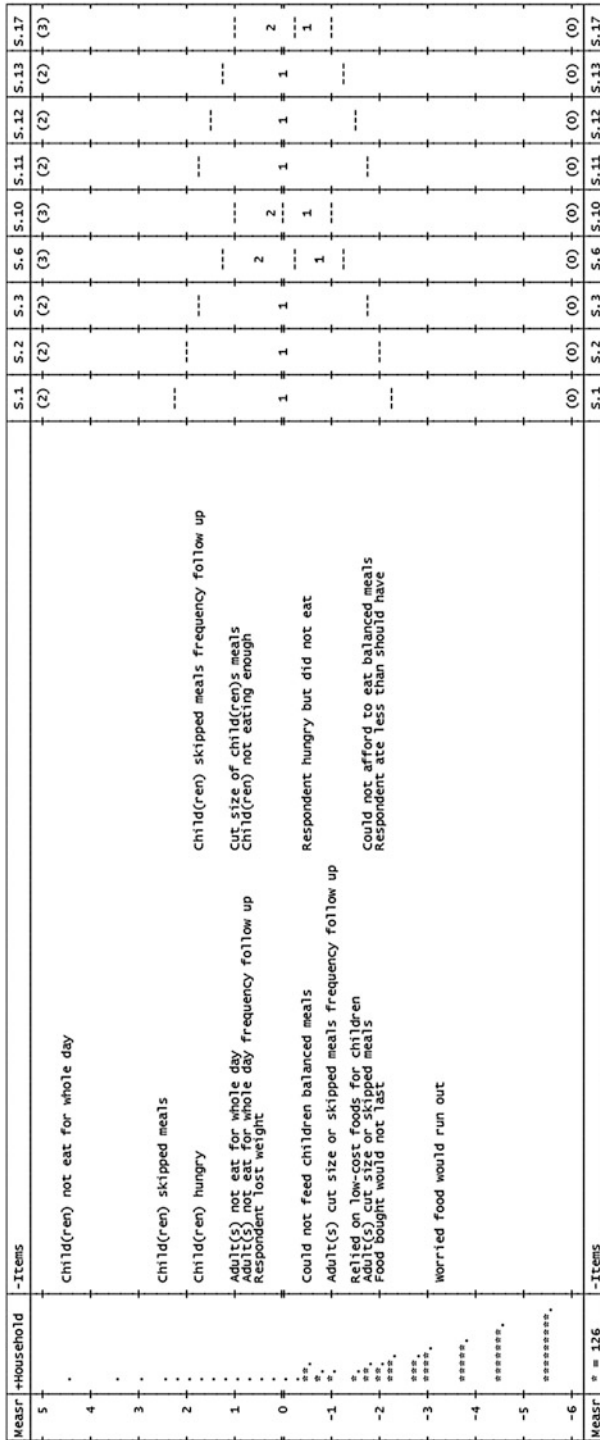


Fig. 2 Wright map of the partial credit analysis of the Household Food Security Survey Module (HFSSM) items. Households and items are located on a logit-scale line that represents the latent construct, household food insecurity. Higher scores correspond to greater food insecurity, for households, and fewer endorsements, for items. Rating scales for the polytomous items of the scale are presented on the right

Table 4 Partial credit category statistics

	Household	Items
Measure		
Mean	-2.91	0.00
<i>SD</i>	1.93	1.95
Outfit		
Mean	0.67	0.99
<i>SD</i>	0.84	0.18
Infit		
Mean	1.01	0.69
<i>SD</i>	0.62	0.37
Separation statistic	2.19	31.45
Reliability of separation	0.83	> 0.99
χ^2 (<i>df</i>)	49353.3* (6607)	28743.1* (17)
<i>N</i>	6608	18
Variance explained by the Rasch model	62.87%	

4 Discussion

The purpose of this study was to compare the dichotomous Rasch model currently in use by the USDA for food security measurement to a partial credit alternative that takes advantage of the polytomous structure of the HFSSM’s items. This is in response to suggestions made to improve the measure (Wunderlich and Norwood 2006) and is the first study since Nord (2012) to address these recommendations and concerns. We are also the first to examine household (person) fit using the polytomous Rasch model for food security; prior to this work, household fit had only been studied using the dichotomous Rasch model (Engelhard et al. 2017).

This paper expands our knowledge of household fit in the context of food security, which has a substantive effect on the food security monitoring used to evaluate the effectiveness of food assistance programs such as the Supplemental Nutrition Assistance Program (SNAP). The first research question was a comparison of the dichotomous and polytomous Rasch model. The results of this study demonstrate that, as Nord (2012) pointed out, neither model is clearly preferred. Both had adequate model-data fit, though the partial credit model explained more variance and had better outfit than the dichotomous model. The second research question asked is what benefits exist to adopting the partial credit model over the dichotomous model. As Nord (2012) noted, the dichotomous Rasch model is both easy to implement and to explain to policymakers; therefore the results of our research do not suggest moving away from USDA’s current practice of using this model.

The final research question addressed the implications for food insecurity measurement when selecting the polytomous over the dichotomous model. The use of a polytomous model has the benefit of gains in measurement precision and an improvement in household food security status classification decisions

Table 5 Partial credit category statistics

Item	Cat score	Count	Cum. %	Outfit MS	Rasch-Andrich thresholds	
					Measure	SE
1	0	1115	17	1.1		
	1	4015	78	1.1	-2.22	0.04
	2	1478	100	1.2	2.22	0.04
2	0	2241	34	1.1		
	1	3448	86	0.9	-1.97	0.03
	2	919	100	1.0	1.97	0.04
3	0	2840	43	1.3		
	1	2865	86	1.2	-1.63	0.03
	2	903	100	1.3	1.63	0.05
6	0	4383	66	1.1		
	1	510	74	0.3	0.24	0.04
	2	985	89	0.5	-1.15	0.05
	3	730	100	0.9	0.91	0.05
10	0	6137	93	1.6		
	1	108	95	0.2	1.39	0.07
	2	205	98	1.0	-1.59	0.08
	3	158	100	1.2	0.20	0.10
11	0	3097	47	1.4		
	1	2657	87	1.5	-1.52	0.03
	2	854	100	1.4	1.52	0.05
12	0	4345	66	1.2		
	1	1843	94	1.0	-1.34	0.03
	2	420	100	1.0	1.34	0.06
13	0	5589	85	1.1		
	1	834	97	1.0	-0.82	0.04
	2	185	100	1.4	0.82	0.09
17	0	6440	97	1.0		
	1	36	98	0.4	1.90	0.10
	2	86	99	1.1	-2.24	0.12
	3	46	100	0.5	0.34	0.19


(Nord 2012). Therefore, the utility of this model over the dichotomous model should be investigated further. This would include a closer inspection of the polytomous coding scheme. The HFSSM is used to estimate and summarize food insecurity prevalence rates in the United States and is targeted by policymakers for interventions intended to reduce food insecurity and improve nutrition across the country. This study is important because it considers an alternate method of modeling household food insecurity that could lead to an improved understanding of the food insecurity measure and of food insecurity prevalence in the United States.

References

- Bond, T. G., & Fox, C. M. (2005). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.
- Coleman-Jensen, A., Rabbitt, M. P., Gregory, C., & Singh, A. (2018). *Household food security in the United States in 2017*. Washington, DC: U.S. Department of Agriculture, Economic Research Service.
- Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.
- Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales*. New York: Routledge.
- Engelhard, G., Rabbitt, M. P., & Engelhard, E. M. (2017). Using household fit indices to examine the psychometric quality of food insecurity measure. *Educational and Psychological Measurement, 78*, 1–19.
- Linacre, J. M. (2015). Facets computer program for many-facet Rasch measurement, version 3.71.4 [Computer software]. Beaverton, OR: Winsteps.com.
- Nord, M. (2012). *Assessing potential technical enhancements to the US household food security measures*. Washington, DC: U.S. Department of Agriculture, Economic Research Service.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).
- Wunderlich, G. S., & Norwood, J. L. (Eds.). (2006). *Food insecurity and hunger in the United States: An assessment of the measure*. Washington, DC: The National Academies Press.

Classical Perspectives of Controlling Acquiescence with Balanced Scales



Ricardo Primi , Nelson Hauck-Filho , Felipe Valentini ,
and Daniel Santos 

Abstract Acquiescence, the tendency to agree regardless of the content of an item, is a commonly observed response style that may distort respondent scores. In the current study, we: (a) revised basic concepts of methods for measuring and controlling acquiescence, (b) describe some important properties of balanced scales, (c) examine if methods of controlling acquiescence provide ipsative scales, (d) explain the mechanism underlying the correction of acquiescence, and (e) compare the centering and standardizing correction methods. By using simulated data, we demonstrate that balanced scales are automatically controlled for acquiescence and that the scoring process does not yield ipsative scales. By contrast, the standardizing method of correction in fact undo the correction that takes place when using the centering method.

Keywords Acquiescence bias · Balanced scales · True-keyed items false keyed items

1 Introduction

1.1 What Is Acquiescence?

Acquiescence is the tendency to endorse the highest Likert categories regardless of the content of the item. One way to examine acquiescence is to include positively-

R. Primi (✉)
Universidade São Francisco, São Paulo, Brazil
EduLab21, Ayrton Senna Institute, São Paulo, Brazil
D. Santos
Universidade de São Paulo, São Paulo, Brazil
EduLab21, Ayrton Senna Institute, São Paulo, Brazil
N. Hauck-Filho · F. Valentini
Universidade de São Paulo, São Paulo, Brazil

keyed (PK) and negatively-keyed items (NK), that is, markers of opposite poles of a trait. For instance, suppose an item designed to measure negative emotional regulation, such as *i+*: “*I adapt easily to new situations without worrying too much,*” for which students must use the following scale to answer: “1” (not at all like me), “2” (little like me), “3” (moderately like me), “4” (a lot like me), and “5” (completely like me). Also, suppose an antonym paired item is included, such as *i-*: “*I have trouble controlling my anxiety in difficult situations.*” A student very high in acquiescence will tend to endorse categories “4” or “5” of both items, which is semantically inconsistent. By contrast, a student who scores high in emotional regulation and low in acquiescence, that is, a person whose item responses are primarily driven by the trait content, will tend to give opposite responses, for example, “4” to *i+* and “2” to *i-*.

Acquiescence represents a method factor, that is, a systematic source of variance unrelated to the target construct that researchers intend to measure (systematic error of measurement, McCrae 2015). It accounts for a sizable portion of the item’s variance in questionnaires, especially when assessing children. It can distort the inter-item covariance matrix of an instrument (internal structure validity), and bias correlations with external variables (criterion validity, see: Primi et al. 2019a, b, c).

Some researchers propose that acquiescence will manifest as an overall tendency to agree with positively keyed items from orthogonal trait factors. Acquiescence indeed will affect these scales (e.g., by increasing their correlation), so that it will be confounded to the true trait. People with high observed scores because of true elevations in all measured factors will be undistinguishable from people that have their scores inflated because of acquiescence. Therefore, in those scales composed of only positively keyed items, acquiescence is confounded with content trait and cannot be properly disentangled. However, it might be identified if we have a proper number of logical antonyms measuring both ends of a construct. This view agrees with Hofstee et al. (1998) who wrote “acquiescence may be defined as the discrepancy between the average over opposites and the scale midpoint. In this definition acquiescent responding is illogical and is therefore best treated as an artifact” (p. 898).

We also stress that negatively keyed items should be constructed avoiding the word “not” (‘I am not too talkative’), and rather using affirmative statements measuring the low end of a construct (‘I am a bit quiet’). This will avoid the burden of higher cognitive load on negatively keyed items, when contrasted to their positively keyed counterparts.

Research on acquiescence is moving to more advanced modeling approaches using Multidimensional Item Response Theory (MIRT, see Maydeu-Olivares and Coffman 2006; Primi et al. 2019b; Savalei and Falk 2014). Nevertheless, some misconceptions still appear in the literature. Our purpose in this chapter is (a) to revise basic concepts of methods for measuring and controlling acquiescence, (b) describe some important properties of balanced scales, (c) examine if methods of controlling acquiescence yields ipsative scales (d) explain the mechanism underlying the correction of acquiescence, and (e) compare the centering and standardizing correction methods. We propose a simple simulation providing R code to illustrate the concepts with visualizations from simulation.

1.2 *How to Measure and Control for Acquiescence?* *Re-centering Approach*

Consider a six-item scale composed by three pairs of antonym items scored in a five-point Likert scale using the previously described category labels of similarity to self. Let $i = 1, 2, 3$ be positive keyed items (PK), $i = a, b, c$ be the negative ones (NK), and x_{ij} be the original response of subject j on item i . The acquiescence index acq_j of a subject j is given by:

$$acq_j = \frac{1}{6} \left[\sum_{i=1}^3 x_{ij} + \sum_{i=a}^c x_{ij} \right] \tag{1}$$

Note that we are averaging items before reversing negatively phrased items to capture the overall tendency to agree with the scale categories. Since these items measure the same trait, but come from opposite ends of the trait continuum, agreement with positive items should co-occur with disagreement with negative items. Therefore, the expected score on this index will be $acq_j = 3$. If $acq_j > 3$ or $acq_j < 3$, this will indicate inconsistent responding in the form of high acquiescence or disacquiescence (i.e., disagree more than agree), respectively. Note that, in this example, all three items are logical opposites, the reason why the average over opposites is an acquiescence index acq_j .

When a subject answers “5” (completely like me) to an extraversion item as “I am often too talkative” and “3” (moderately like me) to its logical opposite “I am often too quiet,” his or her acquiescence index will be $acq_j = 4$, that is, 1 point away from the scale mid-point of 3. To this difference from the scale mid-point and the acquiescence index we call discrepancy. So, in order to re-center this subject response we can add the discrepancy $3-4 = -1$ of each item response, recoding “3” into “2” and “5” into “4.” In this way, recoded item means are settled back to the scale mid-point of 3. This method, proposed by Ten Berge (1999), is called re-center approach, and the recoded scores are controlled for acquiescence.

In brief, the re-centering procedure is done by: (a) calculating the acquiescence index acq_j over semantic opposite pairs for each individual j ; (b) recoding original item responses by subtracting the scale’s midpoint M_o of the acquiescence index: $M_o - acq_j$, and then adding this discrepancy to the original responses; (c) reversing negative items; and (d) calculating the total or average scores.

1.3 *Some Properties of Balanced Scales*

One interesting feature of balanced scales – those in which a positive keyed item is balanced with a negative opposite – is that their average/sum scores are automatically controlled for acquiescence. Let scr_j be the classical average/total score on the example of six items. If we do a little regrouping it will be:

$$\text{scr}_j = \frac{1}{6} \left[\sum_{i=1}^3 x_{ij} + \sum_{i=a}^c (6 - x_{jj}) \right] = \frac{1}{2} \left[\frac{\sum_{i=1}^3 x_{ij}}{3} - \frac{\sum_{i=a}^c x_{ij}}{3} \right] + 3 \quad (2)$$

Now let scr. rec_j be the individual j recoded score using the procedure outlined:

$$\begin{aligned} \text{scr.rec}_j &= \frac{1}{6} \left[\sum_{i=1}^3 (x_{ij} + (3 - \text{acq}_j)) + \sum_{i=a}^c 6 - (x_{ij} + (3 - \text{acq}_j)) \right] \\ \text{scr.rec}_j &= \frac{1}{6} \left[\sum_{i=1}^3 x_{ij} + 9 - \sum_{i=1}^3 \text{acq}_j + 18 - \sum_{i=a}^c x_{ij} - 9 + \sum_{i=a}^c \text{acq}_j \right] \\ \text{scr.rec}_j &= \frac{1}{6} \left[\sum_{i=1}^3 x_{ij} - \sum_{i=a}^c x_{ij} + 18 \right] \\ \text{scr.rec}_j &= \frac{1}{2} \left[\frac{\sum_{i=1}^3 x_{ij}}{3} - \frac{\sum_{i=a}^c x_{ij}}{3} \right] + 3 \end{aligned} \quad (3)$$

Therefore, $\text{scr}_j = \text{scr. rec}_j$. In balanced scales, there is no need for additional procedures, because the classical score is automatically controlled for acquiescence.

This is a very important characteristic to remember. Total scores are controlled for acquiescence variance. However, item scores are not. When researchers run item factor analysis on scales that contain true and false keyed items, raw item score variance is a mix of true variance, acquiescence, other systematic factors, and random error. The acquiescence often distorts the factor structure, producing two factors that separate positively from negatively phrased items, even when these items are supposed to measure a unidimensional construct (see Primi et al. 2019a). Based on these results from factor analysis, some might be tempted to conclude that items cannot be summed up because they measure two different factors.

However, ironically, when summing items from balanced scales, acquiescence – that is, the core reason that distort correlations and create a two-factor structure – is partialled out, yielding a cleaned total score close to a unidimensional solution. This can be verified by running item factor analysis on item scores controlled for acquiescence using the formula in step *b* outlined above (Primi et al. 2019a; Soto and John 2017, p. 2).

1.4 Noise Canceling Mechanism

The way balanced audio cables work offers a good analogy for understanding what happens on balanced scales. Balanced audio cables use two wires to carry two copies of the audio signal from a source, for instance a microphone. But the signal polarity is reversed in one of the wires. When external noise comes along the way and interferes with the signal while it travels to the receiver, it will affect both wires/signals. Since noise interfere on both wires, this results in two copies of the noise with positive sign. The receiver device flips back the signal from the reversed

wire. While reversing it will flip the voice signal from negative to positive, and it will flip noise in this wire from positive to negative. Now we end up with two copies of the noise: one negative and one positive. Finally, when the two signals are summed up in the receiver, noise cancels out, and the signal remains intact and amplified.

This is analogous to what happens with balanced personality scales: Person's true trait is the source signal we are interested in; acquiescence is the noise that comes along with the source signal. The inclusion of items that are logical bipolar opposites will make a copy of the signal with a positive sign – on true keyed items – and a negative sign – on false keyed items. Similarly, acquiescence will influence both items by introducing a positive sign (overall tendency to agree). When scoring the test, we reverse negative items and sum them up just as the audio device does with signal from two wires. Therefore, the source signal is intact and amplified, and acquiescence cancels out.

To make the analogy clearer, let us assume a simple model with no measurement errors and item effects, that is, no differences in item difficulties (see Primi et al. 2019c, for a MIRT formulation of this, conceiving acquiescence as differential person functioning). In this model, agreement with the item i by a person j , x_{ij} is a function of a person's true trait T_j and acquiescence A_j . In positively phrased items $x_{ij} = T_j + A_j$ whereas in negatively phrased items $x_{ij} = -T_j + A_j$. The core part of the formula describing the scoring procedure is the difference between positive and negative items $\sum_{i=1}^3 x_{ij} - \sum_{i=a}^c x_{ij}$. So, for a balanced test with two items $\text{scr. rec}_j = 1/2 (T_j + A_j - (-T_j + A_j)) = T_j$. It is interesting that this noise canceling mechanism was used intentionally by Mirowsky and Ross (1991) to create a measure of locus of control that have acquiescence and social desirability canceled.

1.5 Does Re-centering Produce Ipsative Scores?

Some researchers name the centering transformation as an ipsative transformation. Chan and Bentler (1998) explain what an ipsative scoring is: "Let $x = (x_1 \dots x_p)$ be a $p \times 1$ column vectors such that $\sum_{i=1}^p x_i = l'x = c$ where l is a $p \times 1$ unit vector and c is constant scalar. So x is a p -dimensional data vector with ipsative property" (p. 215).

The transformation outlined above that subtracts acq_j , the subject mean, from each item score x_{ij} producing a transformed score $x'_{ij} = x_{ij} - \text{acq}_j$ is indeed an ipsative transformation because if we sum these transformed scores x'_{ij} for all items they will sum to zero for each individual:

$$\sum_{i=1}^3 (x_{ij} - \text{acq}_j) + \sum_{i=a}^c (x_{ij} - \text{acq}_j) = 0 \quad (4)$$

Ten Berge (1999) examined properties of ipsative transformation of balanced personality scales containing NK and PK items. He noted that balanced scales are a special case with peculiar properties. The detail that needs to be remembered is that when calculating subjects' scores, we reverse half x'_{ij} item scores of NK items and then calculate average item scores. Because of this reversal, the scores are not further ipsative. There will be between-subject variance left. But this variance is disentangled from acquiescence (the variance related to acq_j). If scales are composed with only true keyed items or only false-keyed items, then we will have ipsative scores. This is the idea of Ten Berge (1999) paper's title: "A Legitimate Case of Component Analysis of Ipsative Measures, and Partialling the Mean as an Alternative to Ipsatization."

2 Simulation

2.1 What Does Correction for Acquiescence Really Do?

In order to understand the result of the acquiescence correction, we prepared a simple simulation in R.¹ We first simulated all response patterns of a balanced scale of six items: three positively keyed and three negatively keyed. All items are scored on a five-point Likert scale. This resulted in $5^6 = 15,625$ possible response patterns, which composed a database for the analyses that follows. Using this database, we computed scr_j , acq_j , $scr. rec_j$. Upper part of Figure 1 shows the scatter diagram of scr_j vs $scr. rec_j$ colored by acq_j . It can be observed (upper part) that the classical scores are the same as controlled for acquiescence scores.

Lower part shows scr_j vs acq_j , illustrating how acquiescence correction operates. When acquiescence is equal to the expected value of 3 under a consistent responding (x -axis), that is, negative item responses reflected from positive item responses, scores have a full amplitude of variation from 1 to 5 (y -axis). As responses deviate from the expected value either because individuals are acquiescent $acq_j > 3$ or disacquiescent $acq_j < 3$, scores amplitude is shrunk. So, when individuals agree with items in an inconsistent manner, their scores are regressed toward the mid-point of the scale. In extreme cases, when individual's acquiescence is 1 (pattern 11111) or 5 (pattern 55555), scores will equal the mid-point with no variance.

Figure 1 also illustrates how acquiescence variance is being partialled out. Imagine that a sample has many acquiescent individuals with $acq_j = 4$. The amplitude of their scores will have less variance. Hence, the total variance will be less than the maximum amplitude possible for a sample of consistent responders with $acq_j = 3$. This happens because part of the item response variance is due to acquiescence and, therefore, it is partialled out.

¹R code is available here: http://www.labape.com.br/acqu_mirt/methods_of_recoding.html
see also: https://github.com/rprimi/acqu_mirt

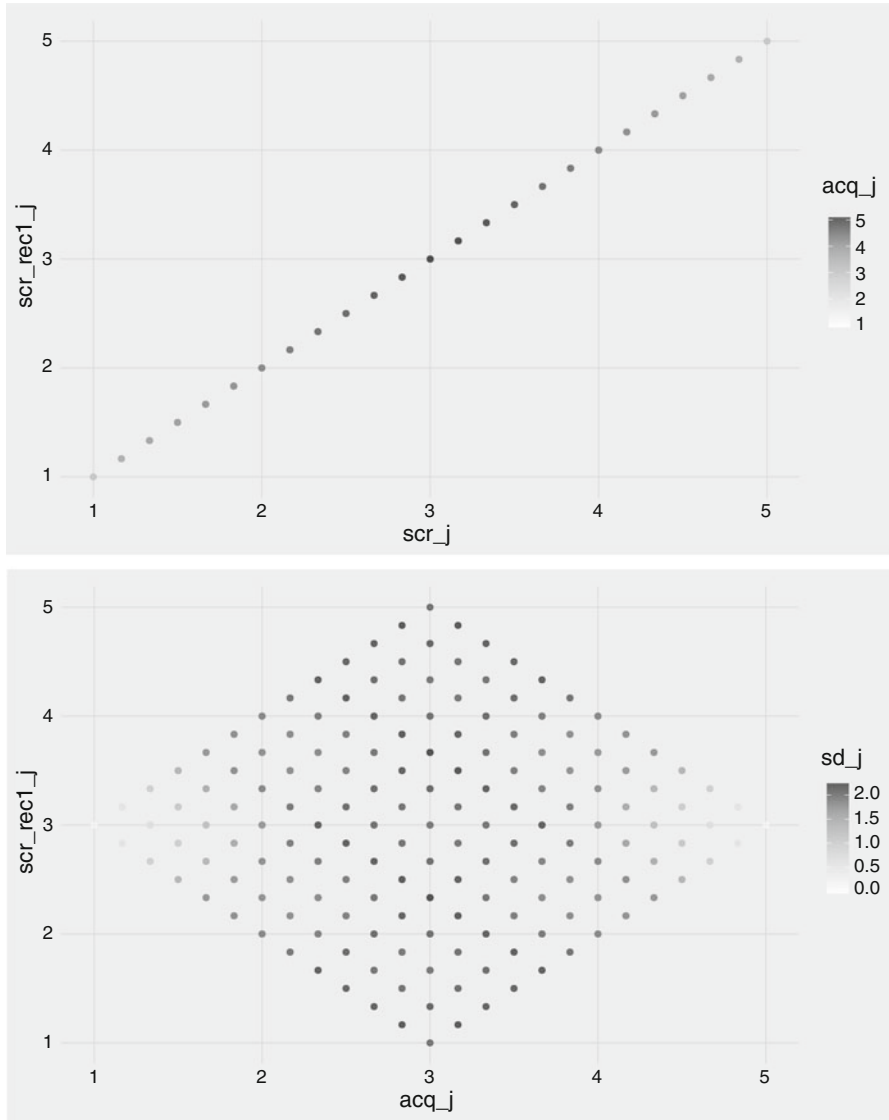


Fig. 1 Correlation of scr_j vs scr_rec_j colored by acq_j (upper panel) and scr_rec_j vs acq_j (lower panel)

2.2 What Happens When We Center and Standardize by an Individual's Spread?

Another method of acquiescence correction proposes that after subtracting the individual's mean on all items (acquiescence index acq_j) we divide by the individual's standard deviation. Hofstee et al. (1998) called this method "row standardization

that additionally corrects for individual differences in spread,” and they warn us that “contrary to the case of acquiescence a cogent rationale for this correction is lacking” (p. 901).

Using our simulated database, we computed the standardized score:

$$scr.z.rec_j = \frac{1}{6} \left[\sum_{i=1}^3 [(x_{ij} - acq_j) / sd_j] - 1 \sum_{i=a}^c [(x_{ij} - acq_j) / sd_j] \right] \tag{5}$$

The formula for the individual standard deviation can be written as:

$$sd_j = \sqrt{\left(\frac{\sum_{i=1}^{i=6} x_{ij}^2}{6} - acq_j^2 \right)} \tag{6}$$

Note that standard deviation is dependent on squared acquiescence. Figure 2 shows the relationship between the subject’s standard deviation (y-axis) and his acquiescence index (x-axis) making this dependency clear. Formula (6) has a quantity plus – 1 multiplying acquiescence squared so it has the shape of an inverted parabola. Note that, as acquiescence diverge from its expected value of 3 on both directions (to 1 or 5), the standard deviation decreases.

Note that the numerator of formula (5) is the re-centering approach, which decreases the item score – i.e., makes the item score less extreme, closer to the

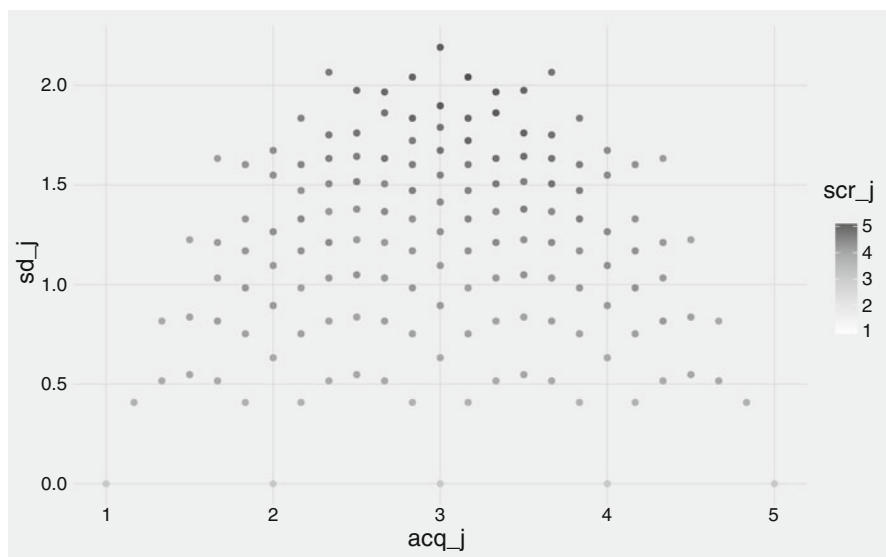


Fig. 2 Relationship between the individuals’ acquiescence index (x-axis *acq_j*) and their standard deviation (*sd_j*)

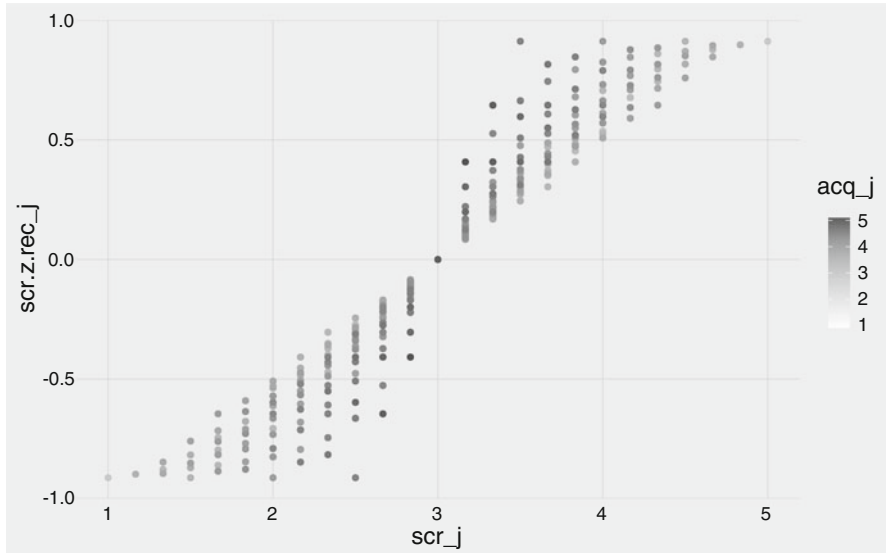


Fig. 3 Relationship between the individuals’ standardized score and their original score colored by individual’s standard deviation (spared of responses sd_j)

scale midpoint – proportionally to high acquiescence (or increases item score proportionally to high disacquiescence). Nevertheless, dividing this transformed score by sd_j will expand the amplitude of item score back, since high acquiescence is related to low sd_j .

Figures 3 and 4 illustrate these relationships. Figure 3 shows how the standardized score (y-axis) is related to the original score (x-axis) that is the result of re-centering approach with points colored by acquiescence. Note that individuals with high acquiescence had original scores with reduced amplitude (around 2–4) due to the automatic correction, but they are mapped onto the same amplitude –1 to 1 as individuals with expected acquiescence of 3.

Figure 4 shows standardized score on the y-axis versus acquiescence on the x-axis, similar to lower graph of Figure 1. Note that this figure does not have the diamond shape as before, meaning that the correction for acquiescence is not working properly. Even individuals with high acquiescence (or high disacquiescence) have the same amplitude on the recoded scores (y-axis). Another way to interpret this figure is that standardized scores are unrelated to acquiescence. In conclusion, standardizing misses the essence of controlling for acquiescence.

Imagine a subject A, that endorses “5” (completely like me) to an item “*I am too talkative*” and “1” (not at all like me) to “*I am too quiet*.” Now imagine a subject B, that also endorses “5” (completely like me) to the first item, but “4” (a lot like me) to the second item. Subject A will have a $scr_j = 5$ and $scr.z.rec_j = 1.5$ indicating high extraversion. Subject B will have a score $scr_j = 3.5$. The model will regress the subject B score toward the mid-point due the inconsistent pattern. Is person

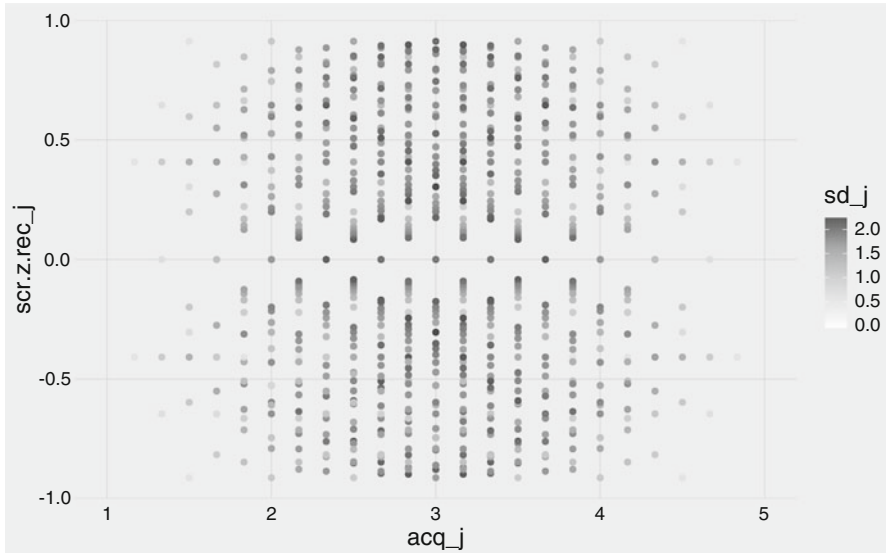


Fig. 4 Relationship between individual's standardized score and individual's acquiescence colored by the standard deviation (spared of responses sd_j)

B talkative or quiet? It is difficult to know considering his inconsistent response pattern.

Importantly, subject B standardized score will be $scr. z. rec_j = 1.5$ indicating a similar level of extraversion to subject A. The logic of standardized score is that person B uses only a tiny gradient of the full amplitude of the scale (4 and 5). Person B has a $sd_j = .5$ compared to $sd_j = 2$ of Person A. Therefore, this method considers this tiny deviation as indicative of a high level of extraversion in the context of a restricted use of the scale. This method standardizes the scale across subjects.

3 Discussion

We revised several basic properties of measuring and controlling for acquiescence. We emphasize four key conclusions that are important for research practice. First, we conclude that the mainstream unfavorable view about negatively phrased items is based on a questionable practice of internal structure analysis (Suarez-Alvarez et al. 2018). Researchers usually run item factor analysis on raw responses without controlling for acquiescence. As a consequence, results will likely show factors grouping PK items separate from NK items. This will lead to a conclusion advising that scales should be splitted into two factors or negative items should be avoided (Gehlbach and Artino 2017). Nevertheless, as we have shown, when we sum PK and NK items, a “noise canceling” mechanism operates, and part of the cause for the

separation of the two factors is removed from the total score. Our conclusion advises that researchers should run item factor analysis based on recoded item scores, as this removes acquiescence systematic error, producing unbiased item-correlation matrices that, in turn, will provide better evidence about whether items conform to a unidimensional bi-polar solution (Primi et al. 2019a).

Second key point refers to the method of measuring acquiescence. We concluded that good semantic opposite pairs requiring PK vs. NK items are needed to measure acquiescence as inconsistent responding. Other operationalizations do not require NK items and propose averaging agreement over several uncorrelated dimensions of scales composed of PK items (Wetzel et al. 2016). It can be inferred from our demonstration that scales composed of items with the same key (only PK or only NK) will not create the noise canceling mechanism.

The third key point refers to the methods of correction. This study shows graphically that the correction is essentially a “partialing out” action that removes acquiescence variance from the item scores (as was explained by Ten Berge 1999). By using this procedure, we can expect the scores validity to increase at the level of individuals. However, after correction, we shrink extreme scores that regress to the mid-point because we do not have much confidence on item scores to assert individual’s salience in one or another direction due to inconsistent responding. This is not exactly a more valid score. By contrast, when we study groups of individuals and correlations between measures, we do have more valid coefficients due to the clearance of a systematic error that might suppress or inflate correlations (see: Primi et al. 2019a).

Still one conclusion related to the method of correction is that centering is the method that should be used. We have shown that row standardization may undo the correction. Hofstee et al. (1998) has warned about the need of more studies for this method. We consider that studies on response process and cognitive laboratories will be important to shed light on the underlying processes of inconsistent responding or agreement behavior. If inconsistent responding over semantic opposites is due to general idiosyncratic restriction in the use of the full Likert categories, then row-standardization that equate scale use is justified. Alternatively, if inconsistent responding is more related to Messick’s (1966) “interpretative acquiescence” related to verbal comprehension skills, then only the centering method should be used.

Last, we highlight the mechanism of noise canceling as a clever method for identifying and disentangling systematic error from true trait variance. Interesting examples of using this mechanism in other types of bias (like defense bias) is tested by Mirowsky and Ross (1991). This is an example of experimental manipulation of item design features to create more pure measures as it is proposed for cognitive testing by Embretson (1994).

Finally, we point to some limitations of this study. An important one is the classical test theory assumptions of equal item difficulties of antonym pairs. Another limitation is the restriction of our simulations to balanced scales only. An interesting follow-up study would be to relax these assumptions with MIRT methods and investigate the scale properties when items are unbalanced as it is done in some examples in Ferrando and Lorenzo-Seva (2010), Primi et al. (2019b) and Savalei and Falk (2014).

Acknowledgments We acknowledge the support of the Ayrton Senna Foundation. The first, second, and third authors also received a scholarship from the National Council on Scientific and Technological Development (CNPq, 310909/2017-1), Coordination for the Improvement of Higher Education Personnel (CAPES, 88881.337381/2019-01) and São Paulo Research Foundation (FAPESP, 2018/10933-8).

References

- Chan, W., & Bentler, P. M. (1998). Covariance structure analysis of ordinal ipsative data. *Psychometrika*, *63*(4), 369–399. <https://doi.org/10.1007/BF02294861>.
- Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment. A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.
- Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 427–448. <https://doi.org/10.1348/000711009X470740>.
- Gehlbach, H., & Artino, A. R. (2017). The survey checklist (manifesto). *Academic Medicine*, *93*(3), 1. <https://doi.org/10.1097/ACM.0000000000002083>.
- Hofstee, W. K. B., Ten Berge, J. M. F. T., & Hendriks, A. A. J. (1998). How to score questionnaires. *Personality and Individual Differences*, *25*(5), 897–909. [https://doi.org/10.1016/S0191-8869\(98\)00086-5](https://doi.org/10.1016/S0191-8869(98)00086-5).
- Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344–362. <https://doi.org/10.1037/1082-989X.11.4.344>.
- McCrae, R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*(2), 97–112. <https://doi.org/10.1177/1088868314541857>.
- Messick, S. (1966). The psychology of acquiescence: An interpretation of research evidence. *ETS Research Bulletin Series*, i–44. <https://doi.org/10.1002/j.2333-8504.1966.tb00357.x>.
- Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 X 2 index. *Social Psychology Quarterly*, *54*(2), 127–145. <https://doi.org/10.2307/2786931>.
- Primi, R., De Fruyt, F., Santos, D., Antonoplis, S., & John, O. P. (2019a). True or false? Keying direction and acquiescence influence the validity of socio-emotional skills items in predicting high school achievement. *International Journal of Testing*. <https://doi.org/10.1080/15305058.2019.1673398>.
- Primi, R., Hauck-Filho, N., Valentini, F., Santos, D., & Falk, C. F. (2019b). Controlling acquiescence bias with multidimensional IRT modeling. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology* (pp. 39–52). Cham: Springer. https://doi.org/10.1007/978-3-030-01310-3_4.
- Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019c). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 447–465. <https://doi.org/10.1111/bmsp.12168>.
- Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*(5), 407–424. <https://doi.org/10.1080/00273171.2014.931800>.
- Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. <https://doi.org/10.1037/pspp0000096>.
- Suarez-Alvarez, J., Pedrosa, I., Lozano, L. M. B., Garcia-Cueto, E., Cuesta, M., & Muñiz, J. G. F. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, *30*(2), 149–158. <https://doi.org/10.7334/psicothema2018.33>.

- Ten Berge, J. M. (1999). A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization. *Multivariate Behavioral Research, 34*(1), 89–102. https://doi.org/10.1207/s15327906mbr3401_4.
- Wetzel, E., Ludtke, O., Zettler, I., & Bohnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment, 23*, 279–291. <https://doi.org/10.1177/1073191115583714>.

Testing Heterogeneity in Inter-Rater Reliability



František Bartoš , Patrícia Martinková , and Marek Brabec 

Abstract Estimating the inter-rater reliability (IRR) is important for assessing and improving the quality of ratings. In some cases, the IRR may differ between groups due to their features. To test heterogeneity in IRR, the second-order generalized estimating equations (GEE2) and linear mixed-effects models (LME) were already used. Another method capable of estimating the components for IRR is generalized additive models (GAM). This paper presents a simulation study evaluating the performance of these methods in estimating variance components and in testing heterogeneity in IRR. We consider a wide range of sample sizes and various scenarios leading to heterogenous IRR. The results show, that while the LME and GAM models perform similarly and yield reliable estimates, the GEE2 models may lead to incorrect results.

Keywords Inter-rater reliability · Mixed-effect models · Generalized estimating equations

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-43469-4_26) contains supplementary material, which is available to authorized users. The authors would also like to thank Professor Jee-Seon Kim for helpful comments on prior version of this manuscript. The authors take responsibility for any errors.

F. Bartoš (✉)
Faculty of Arts, Charles University, Prague, Czech Republic

P. Martinková
Faculty of Education, Charles University, Prague, Czech Republic

Department of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic
e-mail: martinkova@cs.cas.cz

M. Brabec
Department of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic
e-mail: mbrabec@cs.cas.cz

1 Introduction

Ratings by multiple raters are used in assessing quality of scientific articles, grant proposals or job candidates. The credibility of ratings is contingent upon its reliability, validity, and fairness (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). One type of reliability, the inter-rater reliability (IRR), measures the degree of consistency between raters. It can be defined as the ratio of true score variance to total score variance (Lord 1959; Novick 1966) which in the simplest case corresponds to the intra-class correlation coefficient in a random-effect model.

Furthermore, the IRR may differ between groups. For example, Martinková et al. (2018) proposed linear mixed-effect models (LME) to account for differences in variance terms between groups and confirmed significant differences in IRR when rating internal vs. external applicants. Whereas Mutz et al. (2012) utilized the second-order generalized estimating equations (GEE2) and confirmed differences in IRR in ratings of grant proposals from different disciplines. A question arises which of these two, or other possible approaches is superior for testing heterogeneity in IRR.

The aim of this study is to compare precision of different procedures in testing differences in IRR between two groups. We present a simulation study comparing LME, GEE and a newly considered approach based on generalized additive models (GAM). We designed a simulation study testing how do individual methods compare across scenarios in which the heterogeneity is introduced by differences in structural variances (variance of the random-effects in the LME framework), differences in residual variances, differences in means of the ratings, or their combination. Moreover, we varied the number of ratees and raters to explore how sample size influences precision of estimation of the individual model parameters and IRR itself.

2 Methods

2.1 *Inter-Rater Reliability*

In cases with nested measurements, such as the case of ratees rated by multiple raters, the IRR¹ might be estimated using a variance decomposition and calculating the intra-class correlation (ICC) (McGraw and Wong 1996; Shrout and Fleiss 1979). In the most trivial example, when assuming the only structural effect causing differences in ratings being the ratees themselves, the ICC can be described by a single random intercept mixed-effect model (Eq. 1) with the observed j th rating of

¹We are using IRR to refer to single-rater IRR in the paper.

i th ratee $Y_{i,j}$ modeled as the sum of overall mean μ , rate-specific intercept α_i and random error $\epsilon_{i,j}$

$$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j}. \quad (1)$$

The IRR is then defined as a proportion of the ratee variance σ_α^2 to the overall variance in ratings (Eq. 2), which corresponds to ICC

$$\text{IRR} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}. \quad (2)$$

Even in this trivial case, heterogeneity in the IRR might emerge due to differences in the structural variance σ_α^2 – the variability of the ratees’ true score, due to differences in the residual variance σ_ϵ^2 , or their combination. Although possible differences in the overall mean ratings μ do not affect the (2) directly, they might lead to a biased estimate if they are not accounted for in (1).

2.2 The Second-Order Generalized Estimating Equations

One way of estimating the IRR is by using the GEE2. Generalized estimating equations were originally introduced as a way of dealing with clustered data by using a “working correlation matrix” easing the computation in comparison to mixed-effect models (Lipsitz and Fitzmaurice 2008). GEE2 in comparison to the generalized estimating equations as introduced by Liang and Zeger (1986) allow not only to specify the model for the mean (Eq. 3) but also the residual variance (Eq. 4) and intra-class correlation (Eq. 5), to depend on a set of covariates (Yan and Fine 2004). We consider only a specific case in which the covariate is group membership, thus the Eqs. (3), (4), (5) consist only of group-specific (g) intercepts γ transformed to appropriate parameter (group specific mean μ_g , residual variance $\sigma_{\epsilon,g}^2$ and ICC_g) by a given link function (f_1, f_2, f_3 respectively).

$$f_1(\mu_g) = \gamma_{1,g}, \quad (3)$$

$$f_2(\sigma_{\epsilon,g}^2) = \gamma_{2,g}, \quad (4)$$

$$f_3(\text{ICC}_g) = \gamma_{3,g}. \quad (5)$$

2.3 Linear Mixed-Effect Models

LME models offer a different way of accounting for clustering in observed ratings $Y_{i,j,g}$ of j th rating of i th individual from group g by fully specifying joint distribution within clusters via latent variables (Eq. 6). In contrast to GEE2, they do not specify a model for ICC but random-effects directly, leading to estimates of the group-specific structural variances $\sigma_{\alpha,g}^2$, assuming normally distributed random effects with mean zero, group specific residual variance $\sigma_{\epsilon,g}^2$, assuming normally distributed residuals with mean zero, and group specific mean μ_g

$$Y_{i,j,g} = \mu_g + \alpha_{i,g} + \epsilon_{i,j,g}. \quad (6)$$

The mixed-effect models can be estimated either in a classical frequentist framework using maximum or restricted maximum likelihood (ML, REML) or by Markov Chain Monte Carlo (MCMC) in a Bayesian framework (Browne and Draper 2006).

2.4 Generalized Additive Models

The GAM models (Wood 2017) are generally fitted using penalized likelihood with quadratic penalties and generalized cross-validation (GCV) estimates of unknown penalization constants. LME models can be viewed as a special case of this general formulation, where the penalty matrix has (somewhat unusually in the context of GAMs motivated by smoothing) full rank, leading to more convenient computations. Penalty coefficients are related to the variance of random effects. In our context, we view (part of the) GAM framework just as a tool for alternative and flexible estimation of LME models.

3 Simulation Study

3.1 Data Generation

The data generating mechanism corresponds to the LME model specification in Eq. 6 with a given number of ratees (I) from two groups (g) who are being rated J times. Equation 6 implies the average rating of ratees (μ_g), structural variance ($\sigma_{\alpha,g}^2$), and residual variance ($\sigma_{\epsilon,g}^2$) resulting in group-specific IRR

$$\text{IRR}_g = \frac{\sigma_{\alpha,g}^2}{\sigma_{\alpha,g}^2 + \sigma_{\epsilon,g}^2}. \quad (7)$$

Table 1 The simulation scenarios setting

Scenario	μ_1	μ_2	$\sigma_{\alpha, 1}$	$\sigma_{\alpha, 2}$	$\sigma_{\epsilon, 1}$	$\sigma_{\epsilon, 2}$	IRR ₁	IRR ₂
1	0.00	0.00	0.67	0.67	0.74	0.74	0.45	0.45
2	0.00	0.00	0.67	0.67	0.67	0.82	0.50	0.40
3	0.00	0.00	0.60	0.74	0.74	0.74	0.40	0.50
4.1	0.00	0.00	0.60	0.73	0.66	0.81	0.45	0.45
4.2	0.00	0.00	0.73	0.60	0.66	0.81	0.55	0.35
5	-0.20	0.20	0.67	0.67	0.74	0.74	0.45	0.45
6	-0.20	0.20	0.67	0.67	0.67	0.82	0.50	0.40
7	-0.20	0.20	0.60	0.74	0.74	0.74	0.40	0.50
8.1	-0.20	0.20	0.60	0.73	0.66	0.81	0.45	0.45
8.2	-0.20	0.20	0.73	0.60	0.66	0.81	0.55	0.35

With values inspired by results of Martinková et al. (2018) we manipulated standardized mean differences between the groups ($\mu_2 - \mu_1 = 0, 0.4$), structural variance ratios ($\sigma_{\alpha, 1}^2 / \sigma_{\alpha, 2}^2 = 1, 1.5$), and residual variance ratios ($\sigma_{\epsilon, 1}^2 / \sigma_{\epsilon, 2}^2 = 1, 1.5$), while constraining the overall variance to 1 and the mean IRR across groups to 0.45. This led to eight simulation scenarios, with scenarios 4 and 8 split into two subscenarios depending on whether the structural and residual variance ratios differed in the same or the opposite direction (Table 1). Moreover, we manipulated the number of times the rates were rated ($J = 3, 5$) and the number of ratees per group ($I = 25, 50, 100, 200$) in each scenario. In total, 10 (scenarios including subscenarios) \times 2 (number of ratings) \times 4 (number of ratees) = 80 conditions were simulated, 1000 times each, implying 80,000 randomly generated datasets. Code for the data generating process is provided in the Appendix.

3.2 Model Implementation

The GEE2 models were estimated in R (R Core Team 2019) using `geepack` package (Halekoh et al. 2006) with fully iterated jackknife variance estimator, exchangeable covariance matrix, identity link to mean, exponential link to the residual variance, and modified Fisher-z transformation restricting the ICC to $[-1, 1]$ interval as in Mutz et al. (2012). All the remaining settings of `geese()` function were left at their default values, with the maximal number of iterations being increased to 500 just for the case of slower convergence.

To fit LME models, we used three types of implementation as specified below. First, the `lme4` package (Bates et al. 2015) was used as in Martinková et al. (2018). Because `lme4` does not allow for the specification of the residual variance in LME, we also used the `nlme` package (Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2019). Finally, Bayesian estimates through MCMC were implemented

through customized models written in Stan (Carpenter et al. 2017) and fitted via the `rstan` package (Stan Development Team 2019).

The LME with `lme4` package were fitted using REML with default settings. The results were bootstrapped 1000 times, as in Martinková et al. (2018), in order to obtain the standard errors for the structural and residual variance and the IRR estimates with confidence intervals (CI). In the case of nonconvergence, the nonconvergent fit was updated using 20,000 additional function evaluations.

In models fitted with the `nlme` package the structural variance was specified using the “pdDiag” argument in `random`, and the residual variance by the “varIdent” argument in `weights`, in order to allow variances to differ by group. The models were fitted by REML with all other settings kept at default values. All the transformations and computation required for obtaining the final estimates and their standard errors, including the one IRR, were done by a delta method implemented in the `car` package (Fox and Weisberg 2019), which is a method for error propagation that allows to obtain an approximate distribution for a function of an asymptotically normal statistical estimator (Doob 1935). In the case of nonconvergence, the number of iterations, optimizations steps, and objective function evaluations were increased to 500 and initial estimate refinements to 50.

The Stan models were written with identity link to the dummy coded group means, structural variances, and residual variances. For all parameters estimated using Stan, 95% CI were computed using the samples from the posterior distribution. After a preliminary check of computations, we used a noncentral parametrization for the models with $i = 25$, and central parametrization otherwise (Betancourt and Girolami 2015). We used a normal prior distribution with mean 0 and standard deviation 1 for the means and a half-normal prior distribution with mean 0 and standard deviation 1 for the structural and residual variances. The models were fitted using Hamiltonian Monte Carlo with two chains and iterations set to 2000 out of which 1000 was set aside for warm-up. All the remaining settings were kept at the default values. In the case of nonconvergence, the average proposal acceptance probability was increased to 0.95 and the maximum tree-depth to 15.

The GAM models were fitted using the `mgcv` R package (Wood 2011) which allows estimating the random-effects using a “smoother function” – which leads to penalized likelihood estimation. The model specification for residual variance requires `gauss` family function, however, we did not manage to set the optimizer in a way that would not produce underestimated estimates for the structural variance across all our simulation scenarios. We, therefore, used model facilitating `Gaussian` family function which does not allow to specify a model for the residual variance. The random effects were specified using a “re” type spline and the models were fitted using REML, with all the remaining settings at the default values. The IRR and its standard errors were computed using the delta method. In the case of nonconvergence, the maximum number of iterations was increased to 500. Code for the models corresponding to the most complex scenarios is provided in the Appendix; codes for all models is accessible in the online supplementary material available at <https://osf.io/9sajx/>.

3.3 *Simulation Evaluation*

Models were deemed as nonconvergent if there was a warning or error message returned while the model was fit or accessed. In the case of Stan models, we deemed the model as nonconvergent in following cases: any divergent iteration, an effective sample size below 100, Bayesian fraction of missing information below 0.20, or R-hat above 1.10, where R-hat is an indicator of chains not converging to stationary distribution, defined as the average variance of draws from one chain to the variance of draws from all chains (Gelman et al. 1992). Otherwise, the model was refitted using the settings specified above. The fitting times include the refitting process and the bootstrap, in case of models fitted using $1m \times 4$, with both chains in Stan models running in parallel.

The bias of group means, structural variances, residual variances, and IRR estimates were computed as an average difference between the estimated and true values. Root mean square error (RMSE) was calculated by taking the root of the mean of squared differences between true and estimated values. Both of these indices were computed from all models corresponding to the data generating mechanism and also for only those marked as converged.

In addition, the coverage of 95% CI was computed as a proportion of CI containing the true value.

Furthermore, the methods were also compared in terms of power and error rate. In cases where the models corresponding to the data generating mechanism allowed for a particular estimate to differ between the groups, the power was computed as a proportion of significant z -tests at the $\alpha = 0.05$ level. In the remaining cases, the error rate for a particular estimate was computed as a proportion of significant z -tests at the $\alpha = 0.05$ level in a model identical to the one corresponding to the data generating process but allowing the particular parameter to vary by group.² In Stan and lme4 models, the difference in IRR was tested by an overlap of 95% CI for the difference between the by group IRR with zero.

Because there were minimal differences in all evaluation metrics between all models and only models that converged, we present only results for all models (unconditional on convergence). The differences between $J = 3$ and $J = 5$ were also rather minimal (apart from up to 10% increase in power for structural variances); therefore, only results from simulations with $J = 3$ are presented. Results conditional on convergence and $J = 5$ can be found in the online supplemental material.

²That is, in scenario 1, the error rate for mean parameter in mixed-effect model was computed using estimates from a model corresponding to the scenario 5. Furthermore, the error rate for IRR in scenario 1 was computed only for GEE2, because only GEE2 offers the possibility to let the ICC parameter vary by group with the remaining parameters being equal across groups.

4 Results

4.1 Convergence and Fitting Times

All the GEE2 and GAM models converged and only three GEE2 models needed to be refitted. The nlme and Stan models also converged in almost all simulations. The lowest convergence for nlme models occurred in scenarios 3 and 7 for nlme models with 97.9% of models converging before and 99.7% after a refit, and Stan models not dropping below 98.8% in scenario 4.2 before and 99.4% after refitting. However, while the lme4 models almost always converged in scenario 1 and 5, with 100.0% and 99.8% convergence before refit and 100.0% after refitting, in scenario 3 and 7 the convergence dropped to 23.1% and 22.4% before and only to 39.4% and 39.3% after refitting (Supplementary Table 1). A closer look into the lme4 convergence issues revealed that a higher sample size led to worse convergence, 51.0% and 50.0% convergence after refit in scenarios 3 and 7 with $I = 25$ and $J = 3$, and 23.9% and 25.2% convergence after refit with $I = 200$ and $J = 5$, for more details see supplementary Table 2.

In regard to fitting times, the GEE2 models were fitted the fastest ($Mdn = 0.02$ s), followed by LME models implemented in nlme ($Mdn = 0.09$ s), the GAM models ($Mdn = 1.07$ s), LME models implemented in Stan ($Mdn = 5.28$ s), with the slowest models being the LME models implemented in lme4 due to bootstrapping ($Mdn = 16.42$ s). All of the fitting times increased with the sample size (see Fig. 1).

4.2 Bias and RMSE

The bias of the mean estimates was very low across all scenarios and methods ($< |0.009|$). However, GEE2 produced a significant amount of positive bias for residual standard deviations (> 0.217) and for IRR estimates (> 0.238) across all sample

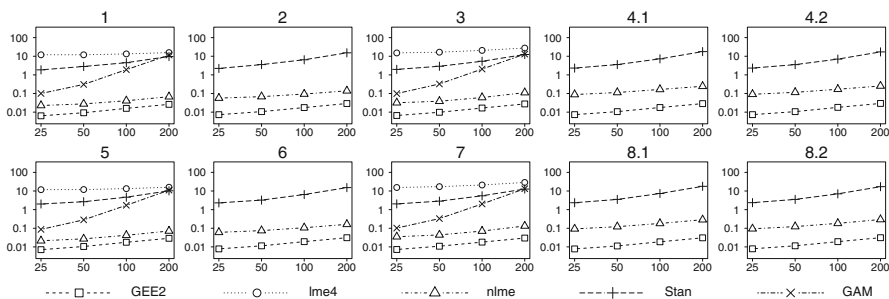


Fig. 1 Median fitting times (in seconds) of different algorithms for the number of ratings $J = 3$, across scenarios (plots) and number of rates per group ($I = 25, 50, 100, 200$, see x-axis)

sizes and scenarios. The remaining methods provided reasonable estimates with a considerably lower bias for structural standard deviations ($< |0.014|$), residual standard deviations ($< |0.020|$), and IRR ($< |0.034|$) (Fig. 2).

The RMSE tells a very similar story, with very low values for mean estimates across all models with a visible improvement with the sample size. As in the bias case, the RMSE for residual standard deviation (> 0.239) and IRR (> 0.277) estimates from GEE2 models were rather high and mostly the result of bias. The remaining methods produced RMSE considerably smaller in structural standard deviation (< 0.167), residual standard deviation (< 0.089), and IRR (< 0.124), with all RMSE decreasing with the sample size (Fig. 3).

4.3 CI Coverage

The CI coverage of the mean estimates was close to the nominal bound across all models (94.3%). However, the mean CI coverage across all scenarios of GEE2 models was low not only for residual variances ($M = 2.9\%$) but also for IRR estimates ($M = 26.2\%$). The other methods performed much better, with the only CI coverage lower than 90% for one of the structural variances in scenario 4.2 and 8.2 in nlme models, with mean CI coverage of 94.7% for structural variances, 94.9% for residual variances and 94.0% for IRR across all models and sample sizes (Fig. 4).

4.4 Power and Error-Rate

The error rate for testing differences between group means was around the nominal level for all scenarios and on average 5.4% and power was swiftly improving with the sample size across all models (Fig. 5). The models fitted with GEE2 showed a high error rate (up to 48.3%) in testing the group differences in the residual standard deviations in scenarios when the structural standard deviations differed as well, which increased with the sample size (scenarios 3 and 7, Fig. 5). Moreover, GEE2 models exhibited much lower power (as low as 12.6%) to detect group differences in IRR estimates in comparison to the other methods ($> 21.7\%$) when the residual variances were different (scenarios 2, 4.2, 6, and 8.2). The occasional nonmonotonic patterns in GEE2 models resulted from high bias and RMSE of the estimated parameters. The remaining methods retained adequate error rate when testing group differences in structural standard deviations ($M = 4.3\%$), residual standard deviations ($M = 5.1\%$), and IRR estimates ($M = 5.0\%$). The power of LME and GAM models increased with the sample size; however, while it approached 100% for both the tests of differences between residual standard deviations (max = 99.0%) and IRR (max = 99.1%), it was rather low for tests of differences between structural standard deviations (max = 53.5%).

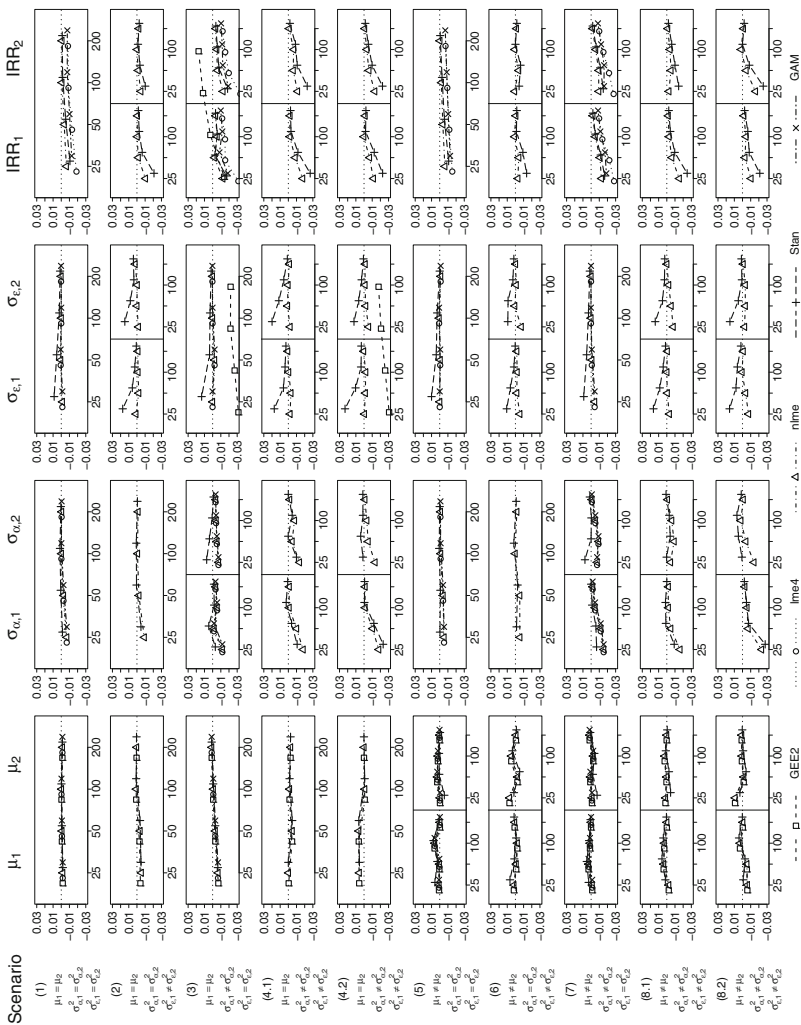


Fig. 2 The bias of parameters (columns) from all models generating mechanism (rows) and number of rating $J = 3$ across number of rates per group ($I = 25, 50, 100, 200$, see x-axis) (bias of residual variance and IRR in GEE2 models is out of the plotting range)

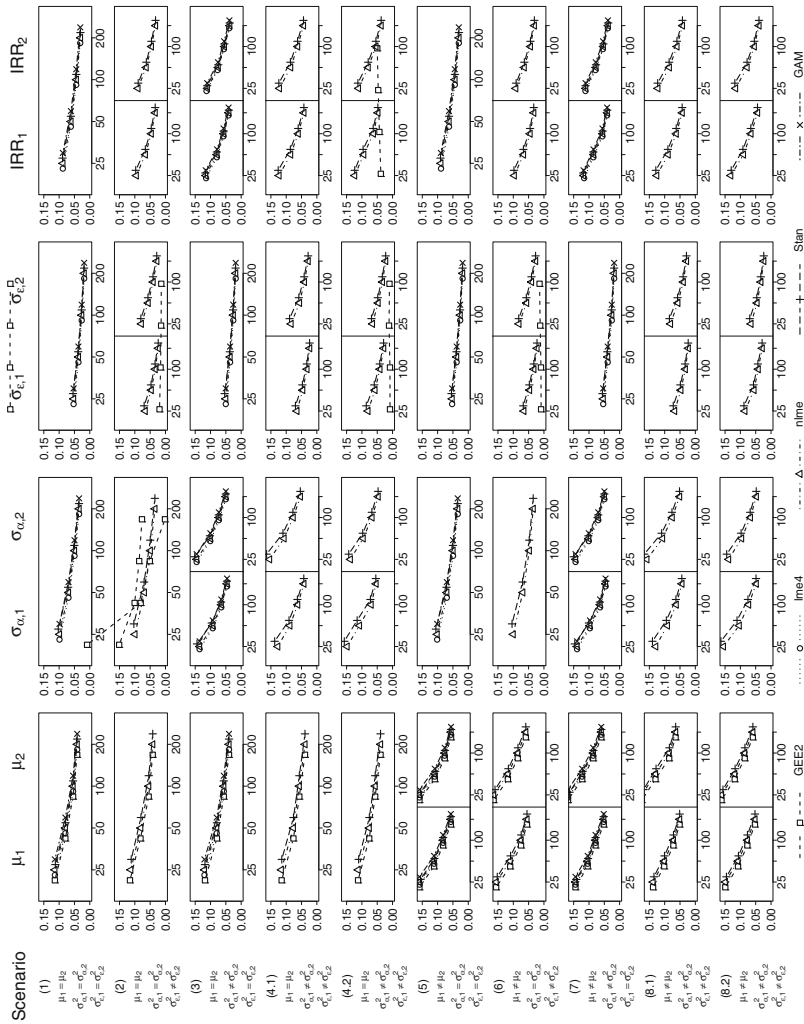


Fig. 3 The RMSE of parameters (columns) from all models corresponding to the data generating mechanism (rows) and number of rating $J = 3$ across number of rates per group ($l = 25, 50, 100, 200$, see x-axis) (RMSE of residual variance and IRR in GEE2 models is out of the plotting range)

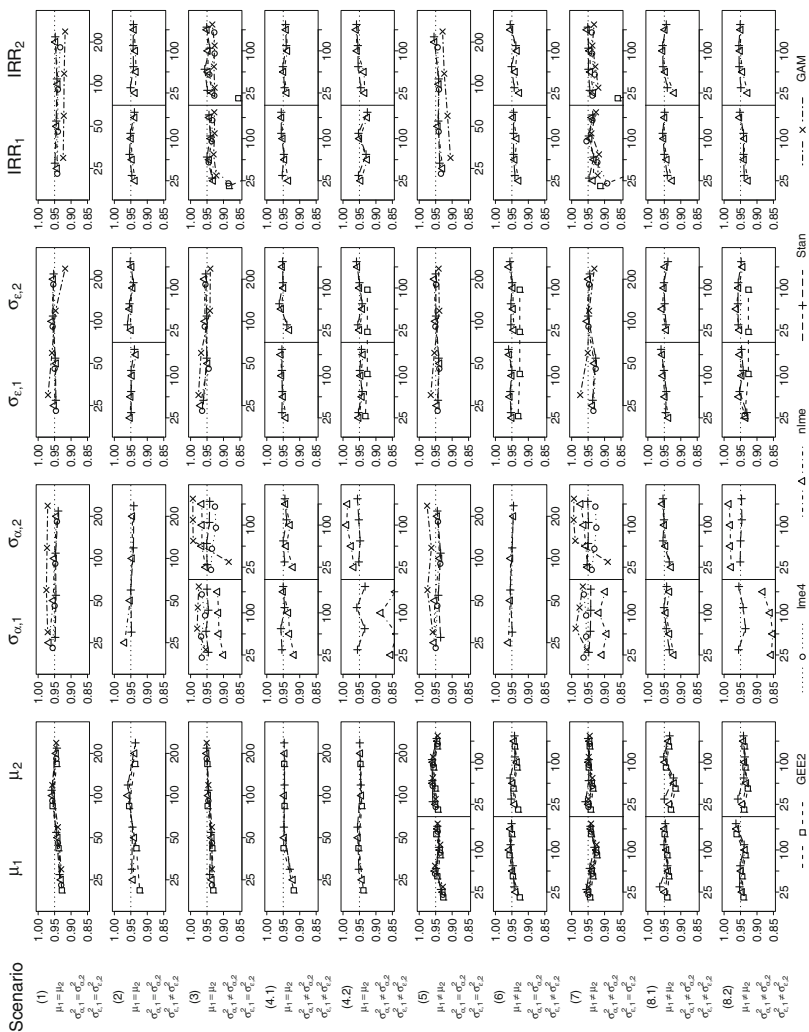


Fig. 4 The CI coverage of parameters (columns) from all models corresponding to the data generating mechanism (rows) and number of rating $J = 3$ across number of rates per group ($l = 25, 50, 100, 200$, see x-axis) (the CI coverage of residual variance and IRR in GEE2 models is out of the plotting range)

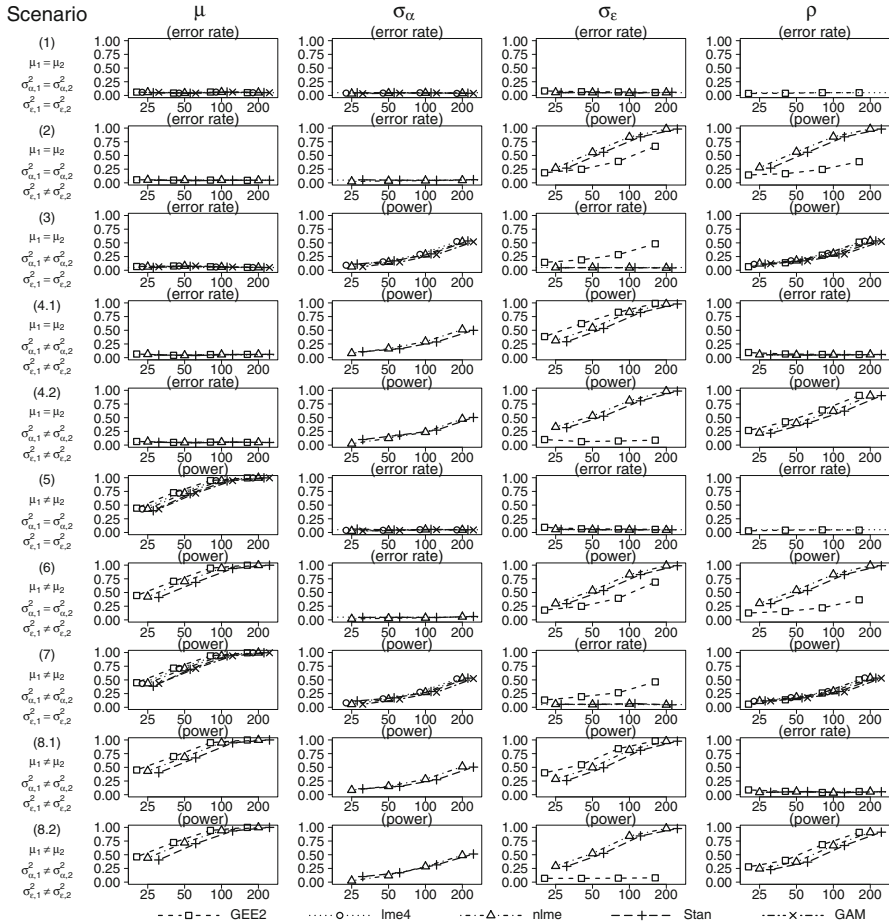


Fig. 5 The error-rate and power for a z -test of difference between parameters (columns) from all models corresponding to the data generating mechanism (rows) and number of rating $J = 3$ across number of rates per group ($I = 25, 50, 100, 200$, see x -axis)

5 Discussion

This study compared the GEE2, LME and GAM models in estimating IRR and other parameters. As a main result, we uncovered an unsatisfactory performance of GEE2 implemented in the geepack package in estimating group residual variances and IRR as measured by bias, RMSE, and CI coverage, and in testing differences in group residual variances and IRR as measured by error-rate, and power. This may be due to less precise model specification in GEE2 than in LME and GAM models. We might only speculate that some fine-tuning of geese.control tolerances could lead to improvements in the quality of fit (while possibly also worsening convergence rates), but in this study we insisted on using default values as a proxy for the

behavior of a typical user. Nevertheless, the estimates of group means produced by GEE2 and also by all other methods were precise and with both standard errors improving with the sample size.

The LME implemented in the `lme4` package exhibited the worst convergence behavior, especially in cases with different structural variances by groups and higher sample sizes. The results from the nonconverged models did not seem to affect the results,³ hinting on a possible difference in the strictness of the convergence checks. Finally, the disadvantage of the mixed-effect models implemented in `lme4` over implementation in `nlme` was in longer fitting times, due to bootstrapping when calculating CI for variance components and IRR, while the biggest drawback was being their inability to estimate models with different parameters for residual variances across groups.

On the other hand, models implemented in the `nlme` package showed almost perfect convergence, allowed to fit a model with different parameters for residual variances across groups and displayed the second fastest fitting times next to GEE2. On top of that, the estimated parameters were satisfactory and comparable to the remaining LME methods and GAM models. Nonetheless, one of the drawbacks of `nlme` might be the difficulty of obtaining CI for the IRR estimates, requiring numerous transformations and the delta method.

In comparison to models implemented in the `nlme` package, the models implemented in Stan required a longer time to fit and their coding was more challenging. However, this burden might be overcome, for example, by the `brms` package (Bürkner 2017) and when using the Bayesian approach, the subsequent manipulation with the estimates was much simpler.

The time required to fit the GAM models was on the higher end, especially with the increasing number size. But there were no problems with convergence and models yielded estimates with similar qualities, power and error rate as those obtained by LME models. Moreover, the GAM models allow formulating a wider range of models than LME models, thus creating further possibilities for future applications.

Furthermore, it is important to note that a small number of ratings for each ratee implied a low power to detect differences in the structural variances between the groups no matter which method was used. An increasing number of ratings per ratee led to a noticeable improvement; however, more than five ratings per ratee would be still needed to achieve adequate power.

There are several limitations worth mentioning. In order to make the simulations feasible, the study relied on the most simplified data generating processes possible. This way, we were able to assess the difficulty of estimating the individual components and their combinations; however, the generalizability of the findings to more complex data scenarios is unsure. Further studies will be needed to explore more complicated scenarios. Furthermore, throughout the study, it was assumed that the data generating process is known and the only issue to solve

³The authors do NOT advise to ignore convergence warnings. These results might be solely due to the specific simulation scenarios settings and models fitted.

is a proper estimation of parameters of the given model or testing differences in given parameters. That is scarcely the case in the most real-life problems, where a substantial theory and/or model selection mechanisms are needed. While from ordinary user perspective testing significance of group effects seems straightforward in GEE2 even in case of more effects or their interactions, model-building and testing may seem more complicated in LME and GAM framework. All that said, the presented simulations may serve as a helpful guideline for estimating and testing differences in IRR between groups.

6 Conclusions

This simulation study has shown that while both LME and GAM are reliable methods for estimating differences in IRR between groups, GEE2 can lead to biased results, inadequate coverage of CI and tests with high error-rates. Therefore, we advise that either LME or GAM is used for testing heterogeneity in IRR.

Acknowledgments The work was partially supported by grants PRIMUS/17/HUM/11 and SVV 2019 – 260482 realized at the Charles University, Faculty of Education and Faculty of Arts and by the long-term strategic development financing of the Institute of Computer Science (Czech Republic RVO 67985807). Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the program “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042), is greatly appreciated.

A.1 Appendix

Function for data simulation takes vectors of length two as an argument for group-specific means (μ), structural standard deviation (σ_{α}), residual standard deviation (σ_{ϵ}), and two integers (I , J) defining the number of ratees in each group I and number of ratings per ratee J .

```
simulate_data <- function(mu, sigma_alpha, sigma_epsilon,
                          I, J){
  # create group membership
  g <- c(rep(1,I), rep(2,I))
  # random effects / ratees' true scores
  alpha <- rnorm(2*I, mu[g], sigma_alpha[g])
  # ratees' indices
  id <- rep(1:(2*I), J)
  # observed values Y
  Y <- rnorm(length(id), alpha[id], sigma_epsilon[g[id]])
  data <- cbind.data.frame("id"=id, "group"= g, "y"=Y)
  data <- data[order(data$id),]
  return(data)
}
```

Thus, the data for the first simulation scenario with 25 rates rated 3 times could be generated as follows:

```
data <- simulate_data(mu = c(0.00, 0.00),
                     sigma_alpha = c(0.67, 0.67),
                     sigma_epsilon = c(0.74, 0.74),
                     I = 25, J = 3)
```

Functions for fitting models corresponding to the most complex scenario 8:

```
### GEE2 model corresponding to scenario 8
# design matrix for correlation parameters
z <- with(data[!duplicated(data$id)],
model.matrix(~ as.factor(group) - 1))
gee_model <- geepack::geese(y ~ as.factor(group) - 1,
                          zcor = z, corstr = "exchangeable",
                          sformula = ~ as.factor(group) - 1,
                          id = id, data = data,
                          mean.link = "identity",
                          sca.link = "log",
                          cor.link = "fisherz",
                          family = gaussian)

### nlme model corresponding to scenario 8
nlme_model <- nlme::nlme(y ~ as.factor(group) - 1,
                        random = list(id = nlme::pdDiag(
                          form = ~ as.factor(group) - 1)),
                        weights = nlme::varIdent(
                          form = ~ 1 | as.factor(group)),
                        data = data, method = "REML")

### lme4 model corresponding to scenario 4
lmer_model <- lme4::lmer(y ~ as.factor(group) - 1
                       + (0 + as.factor(group)|id),
                       data = data, REML = TRUE)

### Stan model corresponding to scenario 8
# stan code with non-central parametrization
stan_code_111_nc <- c('
data {
  int<lower=0> N;
  int<lower=0> N_id;
  int id[N];
  int group[N];
  int id_group[N_id];
  vector[N] y;
}
parameters {
  vector[2] mu_group;
  vector[N_id] alpha_z;
  vector<lower=0>[2] sigma_alpha;
  vector<lower=0>[2] sigma_epsilon;
}
```

```

transformed parameters {
  vector[N_id] alpha;

  for(j in 1:N_id){
    alpha[j] = alpha_z[j]*sigma_alpha[id_group[j]];
  }
}
model {
  vector[N] mu;
  vector[N] sigma;

  target += normal_lpdf(mu_group | 0, 1);
  target += normal_lpdf(sigma_alpha| 0, 1)
           - normal_lccdf(0 | 0, 1);

  target += normal_lpdf(sigma_epsilon| 0, 1)
           - normal_lccdf(0 | 0, 1);
  target += normal_lpdf(alpha_z | 0, 1);

  for(n in 1:N){
    mu[n] = mu_group[group[n]] + alpha[id[n]];
    sigma[n] = sigma_epsilon[group[n]];
  }

  target += normal_lpdf(y | mu, sigma);
}
')
# formatting data for stan
data_stan <- list(
  "id"      = data$id,
  "group"   = as.integer(data$group),
  "y"       = data$y,
  "id_group" = data$group[!duplicated(data$id)],
  "N"       = nrow(data),
  "N_id"    = length(unique(data$id))
)
# compiling model
stan_111_nc <- rstan::stan_model(model_code =
                                stan_code_111_nc)
# sampling the model
model_stan <- rstan::sampling(stan_111_nc,
                              data = data_stan,
                              iter = 2000, warmup = 1000,
                              chains = 2, cores = 2)

# GAM model corresponding to data generating scenario 8
# remapping the ids for more efficient fitting
data$id_new <- as.factor(ifelse(data$group == 1,data$id,
                               data$id - length(unique(data$id))))
data$group <- as.factor(data$group)
mgcv_model <- mgcv::gam(y ~ group - 1
                       + s(id_new,bs = "re", by = group),
                       data = data, method = "REML")

```


References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current Trends in Bayesian Methodology with Applications*, 79, 30.
- Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3), 473–514. <https://doi.org/10.1214/06-BA117>.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>.
- Doob, J. L. (1935). The limiting distributions of certain statistics. *The Annals of Mathematical Statistics*, 6(3), 160–169. <https://doi.org/10.1214/aoms/1177732594>.
- Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Retrieved from <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- Gelman, A., Rubin, D. B., & others. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.
- Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geePack for generalized estimating equations. *Journal of Statistical Software*, 15(2), 1–11.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22. <https://doi.org/10.1093/biomet/73.1.13>.
- Lipsitz, S., & Fitzmaurice, G. (2008). Generalized estimation equations for longitudinal data analysis. In *Longitudinal data analysis* (pp. 43–78). New York, Chapman & Hall/CRC.
- Lord, F. M. (1959). Statistical inferences about true scores. *Psychometrika*, 24(1), 1–17. <https://doi.org/10.1007/BF02289759>.
- Martinková, P., Goldhaber, D., & Erosheva, E. (2018). Disparities in ratings of internal and external applicants: A case for model-based inter-rater reliability. *PLoS One*, 13(10), e0203002. <https://doi.org/10.1371/journal.pone.0203002>.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>.
- Mutz, R., Bornmann, L., & Daniel, H.-D. (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLoS One*, 7(10), e48509. <https://doi.org/10.1371/journal.pone.0048509>.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1), 1–18. [https://doi.org/10.1016/0022-2496\(66\)90002-2](https://doi.org/10.1016/0022-2496(66)90002-2).
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2019). *nlme: Linear and nonlinear mixed effects models*. Retrieved from <https://CRAN.R-project.org/package=nlme>
- R Core Team. (2019). *R: A language and environment for statistical computing*. Retrieved from <https://www.R-project.org/>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.
- Stan Development Team. (2019). *RStan: The R interface to Stan*. Retrieved from <http://mc-stan.org/>
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1), 3–36.
- Wood, S. N. (2017). *Generalized additive models: An introduction with R*. New York: Chapman and Hall/CRC.
- Yan, J., & Fine, J. (2004). Estimating equations for association structures. *Statistics in Medicine*, 23(6), 859–874. <https://doi.org/10.1002/sim.1650>.

An Application of Regularized Extended Redundancy Analysis via Generalized Estimating Equations to the Study of Co-occurring Substance Use Among US Adults



Sunmee Kim, Sungyoung Lee, Ramsey L. Cardwell, Yongkang Kim, Taesung Park, and Heungsun Hwang

Abstract According to the National Survey on Drug Use and Health (NSDUH), the co-use of recreational substances is prevalent in the US population and engenders serious public health consequences. Additionally, substance use is an example of a complex social phenomenon that involves a large number of potentially correlated predictors. Considering the interdependence in the use of cigarettes, alcohol, and marijuana among US adults, the purpose of this study is to investigate simultaneously the effects of multiple sets of predictors (regarding substance initiation age, mental health status, and socioeconomic status) on the use of these three substances. For this, we applied a recently proposed extension of extended redundancy analysis (ERA), named GEE-ERA, to the 2012 NSDUH data. ERA performs data reduction and linear regression simultaneously, producing a simpler description of directional

Sunmee Kim and Sungyoung Lee are co-first authors.

S. Kim (✉)

Department of Psychology, McGill University, Montreal, QC, Canada
e-mail: sunmee.kim@mail.mcgill.ca

S. Lee

Center for Precision Medicine, Seoul National University Hospital, Seoul, South Korea

R. L. Cardwell

Department of Educational Research Methodology, University of North Carolina at Greensboro, Typo, NC, USA

Y. Kim

Department of Statistics, Seoul National University, Seoul, South Korea

T. Park

Department of Statistics, Seoul National University, Seoul, South Korea

Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea

H. Hwang

Department of Psychology, McGill University, Montreal, QC, Canada

relationships between multiple sets of predictors and response variables. The new extension, GEE-ERA, combines ERA with generalized estimating equations (GEE) to enable fitting a regression on a set of correlated responses with unknown correlation structure. This method also adopts ridge-type regularization to address any potential overfitting, while the strength of the regularization is determined automatically through cross-validation. The major findings obtained by applying GEE-ERA to the 2012 NSDUH data are (1) earlier substance use was associated with greater current use of both cigarettes and alcohol; (2) worse mental health status influenced greater marijuana use, only; and (3) a lower level of SES was associated with higher levels of both cigarette and marijuana use.

Keywords Co-occurring substance use · Substance initiation age · Mental health · Socioeconomic status · Component-based dimension reduction · Extended redundancy analysis · Generalized estimating equations · Regularization

1 Background

The current substance use epidemic in the US leads to adverse public health consequences, such as drugged driving (National Institute on Drug Abuse 2019) and smoking- or alcohol-related cancers (US Department of Health and Human Services (DHHS) 2004). According to the 2012 National Survey on Drug Use and Health (NSDUH), an estimated 62% of Americans aged 12 and older used at least one recreational psychoactive substance (i.e., tobacco, alcohol, or illicit drug) within the past year, including 9% who met the criteria for substance abuse disorder (US DHHS, Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality (CBHSQ) 2013). Moreover, the same 2012 NSDUH data show a positive association between cigarette and alcohol use, as well as a correlation between degree of alcohol use and rate of illicit drug use (of which marijuana use accounts for the vast majority) (US DHHS, SAMHSA, CBHSQ, 2013). Considering that the vast majority of substance users (91% in 2012) use more than one substance, either concurrently or sequentially, a statistical model that simultaneously analyzes use of multiple substances would provide a more complete representation of the phenomenon of substance co-use among US adults.

Further complicating the study of substance use among US adults is the large number of predictors that have been demonstrated in previous studies to explain the use of one or more substances (e.g., Daza et al. 2006; Hu et al. 2006; Kandel et al. 2004; Robinson et al. 2006). Categories of such predictors include (1) substance initiation age (i.e., age of first cigarette, alcohol, and/or marijuana use), (2) indicators of mental health (e.g., major depressive episode during past year, daily functional impairment level, etc.), and (3) indicators of socioeconomic status (SES; education level, health insurance coverage, family income, employment status). Considering such a high-dimensional set of predictors, the major difficulty in investigating the effect of numerous predictors on the concurrent use of substances

is the lack of statistical methods capable of providing a comprehensible description of directional relationships among many sets of variables, without suffering from potential multicollinearity issues.

Thus, in the present work, we use extended analysis (ERA; Takane and Hwang 2005) combined with generalized estimating equations (GEE; Liang and Zeger 1986) to investigate associations between the aforementioned predictor sets and correlated use of multiple substances. ERA is a statistical method that relates multiple sets of predictors to response variables. In ERA, a component is extracted from each set of predictor variables such that it accounts for the maximum variation of response variables. In this regard, ERA performs data reduction and linear regression simultaneously, producing a simpler description of directional relationships between multiple sets of predictors and response variables. Recently, a new extension of ERA was proposed for the analysis of clustered or correlated response variables (Lee et al. 2019). In this extension, GEE is combined with ERA to model response variables with an unknown correlation structure. This new method, called GEE-ERA hereinafter, can handle different types of response variables (e.g., continuous, binary, or count) that are assumed to follow an exponential family distribution. The method also incorporates ridge-type regularization to address potential overfitting when many predictors per component are considered or when many components influence the response variables. The regularization strength is determined automatically using cross-validation (CV).

The remainder of the paper is organized as follows. We begin by briefly reviewing GEE-ERA focusing especially on its advantages for the analysis of co-occurring substance use in the US. We then apply the method to data from the 2012 National Survey on Drug Use and Health (NSDUH), an annual survey that provides extensive statistical information on the use of recreational psychoactive substances and various associated sociopsychological variables. This application shows that GEE-ERA can identify meaningful predictors while taking into account the correlation structure of nicotine, alcohol, and marijuana use and preventing overfitting by the regularization strategy. We conclude by discussing the implications of the method and topics for future research.

2 Method

2.1 Model Specification

In GEE-ERA (Lee et al. 2019), we assume that there are Q response variables and K different sets of predictors, each of which consists of P_k predictors ($k = 1, \dots, K$). Let y_{iq} denote the value of the q th response variable measured on the i th respondent ($i = 1, \dots, N$; $q = 1, \dots, Q$). We assume that y_{iq} follows an exponential family distribution with a mean μ_{iq} and variance $\phi\sigma_{iq}^2$, where ϕ is a dispersion parameter which may or may not be of substantive interest. Let w_{kp} denote the component weight assigned to x_{ikp} . Let $f_{ik} = \sum_{p=1}^{P_k} x_{ikp}w_{kp}$ denote the i th component score of

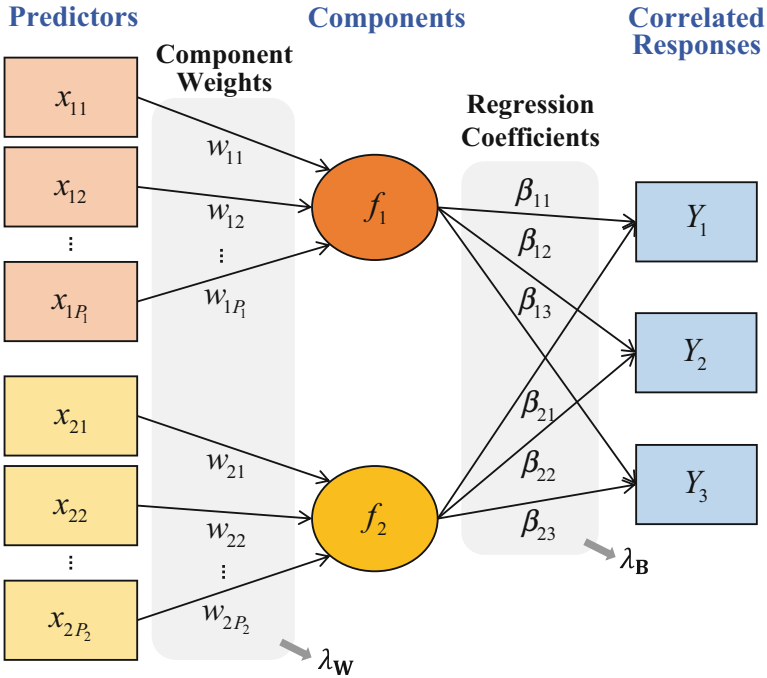


Fig. 1 An example of GEE-ERA model. Square boxes indicate observed predictor and response variables. Circles represent predictor components. Two regularization parameters, λ_W and λ_B , determine the strength of the regularization on component weights and regression coefficients, respectively

the k th component, which is the sum of weighted predictors for the i th observation in the k th predictor set. Let β_{kq} denote the regression coefficient relating the k th component to the q th response variable. Let η_{iq} and $g(\cdot)$ denote the i th linear predictors of the q th response and a link function, respectively. We assume that all the predictors and response variables are standardized with zero means and unit variances (Takane and Hwang 2005). The GEE-ERA model is then expressed as

$$g(\mu_{iq}) = \eta_{iq} = \sum_{k=1}^K \left[\sum_{p=1}^{P_k} x_{ikp} w_{kp} \right] \beta_{kq} = \sum_{k=1}^K f_{ik} \beta_{kq}, \tag{1}$$

where the marginal expectation of the responses μ_{iq} is related to a linear predictor through a known link function. Figure 1 displays an example of the GEE-ERA model, where three response variables are assumed to be affected by each of the two components.

Let $\tilde{y}_i = [y_{i1}, \dots, y_{iQ}]'$ be a Q by 1 vector of the responses of the i th respondent. Let Σ_i be the Q by Q within-respondent covariance matrix of \tilde{y}_i . When respondents are measured on multiple response variables simultaneously, the assumption of independence of response variables in ordinary ERA can be violated.

Moreover, the true covariance structure is often unknown in practice. To resolve these issues in ERA, the method of GEE (Liang and Zeger 1986) was applied to specify the unknown covariance structure using the so-called “working” correlation matrix. The working covariance matrix has the form

$$cov(\tilde{y}_i) = \Sigma_i = \phi A_i^{1/2} R_i(\mathbf{a}) A_i^{1/2}, \tag{2}$$

where $R_i(\mathbf{a})$ is a Q by Q working correlation matrix that is assumed to be fully specified by the vector of unknown nuisance parameters \mathbf{a} , and $A_i^{1/2}$ is a Q by Q diagonal matrix of marginal variances with $var(\mu_{iq})$ as the q th diagonal element (Liang and Zeger 1986). Liang and Zeger (1986) suggested various choices for $R_i(\mathbf{a})$ (see Sect. 3.2), which is constant across all respondents. In this way, we can treat the covariance structure as a nuisance instead of attempting to model it accurately when estimating ERA parameters. This method also can provide asymptotically unbiased parameter estimates and their robust standard errors regardless of the covariance structure specified (Lee et al. 2019).

2.2 Parameter Estimation and Significance Testing

GEE-ERA aims to estimate both ERA parameters (i.e., w_{kp} and β_{kq}) and nuisance correlation parameters (i.e., \mathbf{a} and ϕ) in an iterative manner. Specifically, it seeks to minimize the following penalized least squares criterion for estimating parameters:

$$\begin{aligned} \phi_{(\alpha, \mathbf{W}, \mathbf{B})} = \sum_{i=1}^N & \left[(\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)' \Sigma_i^{-1} (\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i) \right] \\ & + \lambda_{\mathbf{W}} \text{trace}(\mathbf{W}'\mathbf{W}) + \lambda_{\mathbf{B}} \text{trace}(\mathbf{B}'\mathbf{B}), \end{aligned} \tag{3}$$

where \tilde{z}_i is a Q by 1 vector of the so-called adjusted response variable (McCullagh and Nelder 1989, Ch. 2), \mathbf{W} denotes a $\sum_{k=1}^K P_k$ by K matrix of component weights, \mathbf{B} denotes a K by Q matrix of regression coefficients, \tilde{x}_i denotes a vector of predictors for the i th respondent, and $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{B}}$ denote tuning parameters for component weights and regression coefficients, respectively. The tuning parameters control the influence of the ridge penalty terms, $\text{trace}(\mathbf{W}'\mathbf{W})$ and $\text{trace}(\mathbf{B}'\mathbf{B})$. We apply G -fold CV to determine the values of $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{B}}$ automatically. To minimize (3), GEE-ERA uses a regularized alternating least squares algorithm, in which each of \mathbf{W} , \mathbf{B} , and Σ_i is updated, with the other two parameter sets held constant, until convergence. Refer to Lee et al. (2019) for a detailed description of the algorithm.

To test statistical significance of parameter estimates, GEE-ERA can use resampling methods, such as permutation tests for obtaining exact p -values (as described in Lee et al. 2019) and bootstrapping (Efron and Tibshirani 1986) for constructing confidence intervals. In the present analysis, we used bootstrap percentile confidence intervals, i.e., the 5th and 95th percentiles of bootstrap distribution of parameter estimates based on 1,000 bootstrapped replications of the data.

3 An Empirical Application

3.1 Data and Model Specification

The data used here is a subset of the 2012 National Survey on Drug Use and Health (NSDUH) dataset (US DHHS, SAMHSA, CBHSQ, 2013). NSDUH has been conducted every year in all 50 states and the District of Columbia since 1971. The objective of this survey is to serve as a major source of information on tobacco, alcohol, and illicit drug use, and on mental health and other health-related issues in the US. The 2012 NSDUH was conducted from January through December 2012 and interviewed US residents aged 12 and older. Among 51 states, eight of them had a sample designed to yield 3,600 respondents per state, and the remaining 43 states had a sample designed to yield 900 respondents per state. The respondents were asked to answer various questions regarding their use of substances, as well as mental and physical health issues. Each respondent's sociodemographic characteristics (e.g., age, race, marital status, education, financial circumstances, etc.) were also measured. A description of the data set is provided on GitHub at <https://github.com/QuantMM/2012NSDUH>. On the page, we also explained in detail where readers can download the original dataset.

In the present analysis, we examined the effects of predictors related to substance initiation age, mental health, and SES on cigarette, alcohol, and marijuana use. To do so, we utilized the subset of data with valid responses to three substance use variables. This did not result in any missing values for predictor variables of interest. Table 1 presents summary statistics of all variables included in the analysis using data from $N = 881$ respondents. The three response variables, all referring to monthly use on average, are the number of cigarettes smoked (Y_1), the number of alcoholic beverages consumed (Y_2), and the number of days of marijuana or hashish use (Y_3). We identified a total of 11 predictors that were available in the 2012 NSDUH data based on previous studies concerning the predictors of substance use on samples of US adults. Then, the predictors were grouped into the three categories—substance initiation age (F_1), mental health (F_2), and SES (F_3)—which were represented as components in the ERA model. Table 1 also shows which component is associated with which predictors. Figure 2 displays the specified GEE-ERA model, where three sets of predictors related to F_1 , F_2 , and F_3 were to influence each of three response variables.

3.2 Working Correlation Structure of Substance Use Variables

As noted above, previous studies suggested the co-occurrence of the three response variables. In the present data, there was a significant positive association between Y_1 and Y_2 , $r = .18$, $p < .01$. Also, Y_1 and Y_3 were positively correlated, $r = .16$, $p < .01$, whereas Y_2 and Y_3 were not, $r = -.02$, $p = .58$.

Table 1 A description of variables and summary statistics for the 2012 NSDUH data

Variable names	Measures (range or categories)	Mean (Q1, Q3)
Response variables		
Y ₁ : Cigarettes	Number of cigarettes smoked per response in past month	200 (14, 315)
Y ₂ : Alcohol	Number of alcohol beverage drank in past month	55 (12, 64)
Y ₃ : Marijuana	On average, number of days used marijuana or hashish during the past 12 months	8.7 (2, 13)
Predictors		
F ₁ : Age of first use		
Cigarette onset	Age of first use	15.81 (14, 18)
Alcohol onset	Age of first use	16.82 (15, 18)
Marijuana onset	Age of first use	16.94 (15, 18)
F ₂ : Mental health		
Distress level	Nonspecific psychological distress scale (K6) score	2.01 (0, 2)
Impairment	Daily functional impairment due to problems with emotions, nerves, or mental health	1.09 (0, 3)
Suicidal thought	Serious thoughts of suicide in the past year (Yes=1/No=0)	% Yes: 9.58
Depression	Major depressive episode in the past year (Y=1/N=0)	% Yes: 12.5
F ₃ : SES		
Education	5th grade or less (=5), 6th grade (=6), . . . , sophomore/junior (=14), senior/grad or more (=15)	12.41 (12, 14)
Insurance	Having any health insurance (Y/N)	% Yes: 71.75
Family income	Less than \$10,000 (=1), \$19,999 (=2), . . . , \$74,999 (=6), \$75,000 or more (=7)	4 (2, 6)
Employment status	Employed (Y=1/N=0)	% Yes: 67.02

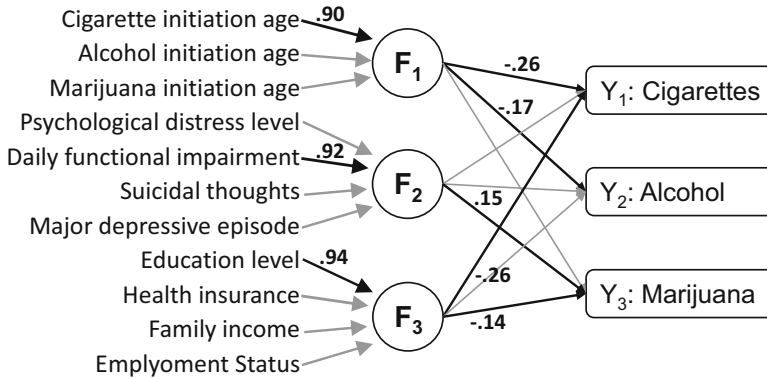


Fig. 2 The specified ERA model for the 2012 NSDUH dataset. Black and bolded arrows represent statistically significant component weights and regression coefficients using bootstrapped confidence intervals with $\lambda_W = 0.12$ and $\lambda_B = 0$

Table 2 The estimated working correlation and dispersion parameters across four different working correlation structures using the 2012 NSDUH data

	Independent	Exchangeable	AR-1	Unstructured
Working correlation structures	$\begin{pmatrix} - & 0 & 0 \\ 0 & - & 0 \\ 0 & 0 & - \end{pmatrix}$	$\begin{pmatrix} - & \rho & \rho \\ \rho & - & \rho \\ \rho & \rho & - \end{pmatrix}$	$\begin{pmatrix} - & \rho & \rho^2 \\ \rho & - & \rho \\ \rho^2 & \rho & - \end{pmatrix}$	$\begin{pmatrix} - & \rho_1 & \rho_2 \\ \rho_3 & - & \rho_4 \\ \rho_5 & \rho_6 & - \end{pmatrix}$
Working correlation estimates	—	$\hat{\rho} = -.003$	$\hat{\rho} = -.005$	$\hat{\rho}_1 = -.015$ $\hat{\rho}_2 = .001$ $\hat{\rho}_3 = .036$
$\hat{\phi}$.002	.002	.002	.002
QIC	2.853	2.853	2.860	2.869

The top row of Table 2 illustrates the four different working correlation structures considered in GEE-ERA to model the relationships in their co-occurrence: independent (all pairwise correlations fixed to zero), exchangeable (all correlations assumed to be equivalent), autoregressive or AR-1 (all first-order correlations assumed to be equivalent and higher-order correlations a function of the first-order correlation parameter), and unstructured (all correlations assumed to be different and not systematically related). Table 2 also summarizes the working correlation and estimated dispersion parameters for each type of correlation structure from the present analysis, as well as the values of QIC, a modified Akaike information criterion for GEE models (Pan 2001). All results in the table were obtained without any regularization, i.e., $\lambda_W = \lambda_B = 0$.

As shown in the table, the estimated correlation parameters changed noticeably in both sign and magnitude across the chosen correlation structures. However, the GEE-ERA parameter estimates in Tables 3 and 4 were robust across different working correlation specifications. The final working correlation was chosen based

Table 3 The estimated component weights for the GEE-ERA model in Fig. 2 with different working correlation structures using the 2012 NSDUH data. Bolded numbers indicate statistically significant estimates using bootstrapped confidence intervals

Components	Predictors	Working correlation			
		Independent	Exchangeable	AR-1	Unstructured
F ₁ : Age of first use	Cigarette onset	.90	.90	.90	.90
	Alcohol onset	.34	.34	.34	.33
	Marijuana onset	.02	.02	.02	-.01
F ₂ : Mental Health	Distress level	-.16	-.16	-.16	-.16
	Impairment	.92	.92	.92	.92
	Suicidal thought	.45	.45	.44	.44
	Depression	-.19	-.17	-.17	-.17
F ₃ : SES	Education	.94	.94	.94	.94
	Insurance	.28	.28	.27	.27
	Family income	-.09	-.09	-.08	-.08
	Employment status	-.29	-.29	-.29	-.29

Table 4 The estimated regression coefficients for the GEE-ERA model in Fig. 2 with four different working correlation structures using the 2012 NSDUH data. Bolded numbers indicate statistically significant estimates using bootstrap confidence intervals

Components	Responses	Working correlation			
		Independent	Exchangeable	AR-1	Unstructured
F ₁ : Age of first use →	Y ₁ : Cigarettes	-.26	-.26	-.26	-.26
	Y ₂ : Alcohol	-.17	-.17	-.17	-.17
	Y ₃ : Marijuana	-.09	-.09	-.08	-.08
F ₂ : Mental health →	Y ₁ : Cigarettes	.12	.12	.12	.12
	Y ₂ : Alcohol	-.02	-.02	-.02	-.02
	Y ₃ : Marijuana	.15	.15	.15	.15
F ₃ : SES →	Y ₁ : Cigarettes	-.26	-.26	-.26	-.26
	Y ₂ : Alcohol	-.05	-.05	-.05	-.05
	Y ₃ : Marijuana	-.14	-.14	-.14	-.14

on the value of QIC: Since independent and exchangeable structures resulted in equal QIC values, the more parsimonious of the two, i.e., independent, was chosen.

3.3 Regularization and Empirical Results

After choosing the final correlation structure, we applied regularization on both component weights and regression coefficients. As the values of the regularization strengths, i.e., λ_W and λ_B , are dependent on the data, they can be determined using data-driven methods, such as CV. We used 10-fold CV for different possible values of λ_W and λ_B . The optimum values were chosen by comparing the average mean-

squared errors, where the values ranged from 0 to 10 with a step size of .05. The lowest error was obtained with $\lambda_{\mathbf{W}} = .15$ and $\lambda_{\mathbf{B}} = 0$. The statistically significant estimates of component weights and regression coefficients with these final values are given in Fig. 2.

As depicted in Fig. 2, the component weight estimate for cigarette initiation age was positive and statistically significant, indicating that cigarette initiation age contributed to forming F_1 , initiation of substance use, in explaining substance uses. Neither alcohol nor marijuana initiation age was statistically significant. For F_2 , mental health status, only the level of daily functional impairment showed a statistically significant contribution. Finally, for F_3 , SES, only education level made a significant contribution to explaining the use of the three substances.

Figure 2 also shows the statistically significant regression coefficient estimates. First, the negative association between F_1 and both Y_1 and Y_2 indicated that a younger age of substance initiation was associated with an increased number of cigarettes smoked and alcoholic beverages consumed, with the effect appearing larger for cigarette use. Additionally, worse mental health status was associated with more days of marijuana use among American adults. There was no influence of mental health status either on cigarette or on alcohol use. Finally, American adults with lower levels of SES were found to report greater levels of both cigarettes smoked and days of marijuana use, where cigarette use was more strongly associated with SES level than marijuana use.

4 Conclusion

The present analysis applied GEE-ERA, a recently proposed extension to ERA, to data from the 2012 NSDUH survey on substance use. Substance use, including use of multiple substances, is prevalent in the American population and the source of numerous public health concerns. Additionally, substance use is known to involve multiple categories of predictors, including the predictor sets considered in the present analysis—initiation of substance use, mental health status, and SES. We investigated the relationship of these predictors with cigarette, alcohol, and marijuana use. GEE-ERA permits the simultaneous analysis of the numerous predictors and multiple, correlated response variables by simultaneously conducting data reduction and multivariate multiple regression while also modeling the correlation structure of the response variables. This method also employs ridge-type regularization to address potential overfitting, determining the strength of the regularization automatically through CV, and conducts significance tests on ERA parameters (i.e., component weights and regression coefficients) using bootstrapping. The method thus protects against the common problems of multicollinearity among predictors, overfitting, and improper use of asymptotic statistical inference while producing easy-to-interpret parameter estimates.

The present analysis has demonstrated the utility of GEE-ERA while also providing insight on the phenomenon of substance use in the US. Nevertheless,

there are several ways to expand upon the present analysis. First, given that the NSDUH is an annual survey, the present analysis can be replicated with data from subsequent years. Also, future studies should include additional predictors that have been found to significantly relate to substance use, such as personality characteristics (Hittner et al. 2020) or sexual orientation discrimination (Evans-Polce et al. 2019). Unfortunately, the 2012 NSDUH data did not include variables relevant to these factors. Additionally, as noted earlier, we used a subset of the 2012 NSDUH data with valid responses to the substance use variables. However, in practice, missing data occur frequently, and listwise deletion—a simple and oft-used approach to missing value imputation—may lead to substantial loss of information. Future work is needed to explore an alternative way of dealing with missing data for the proposed method. For example, it may be worthy of incorporating doubly robust estimation (Carpenter et al. 2006) into the method, which has been a popular approach for handling missing data within GEE, seeing that the results obtained from GEE are only valid under the strong assumption of missing completely at random. And finally, considering previous research that uncovered heterogeneous subgroups characterized by demographic covariates (e.g., gender or ethnicity), each of which yielded different effects of predictors on substance use, it will be worthwhile to further extend GEE-ERA to identify potentially heterogeneous subgroups of observations based on such covariates.

References

- Carpenter, J. R., Kenward, M. G., & Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 571–584. <https://doi.org/10.1111/j.1467-985X.2006.00407.x>.
- Daza, P., Cofta-Woerpel, L., Mazas, C., Fouladi, R. T., Cinciripini, P. M., Gritz, E. R., & Wetter, D. W. (2006). Racial and ethnic differences in predictors of smoking cessation. *Substance Use & Misuse*, *41*(3), 317–339. <https://doi.org/10.1080/10826080500410884>.
- Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*, 54–75.
- Evans-Polce, R. J., Veliz, P. T., Boyd, C. J., Hughes, T. L., & McCabe, S. E. (2019). Associations between sexual orientation discrimination and substance use disorders: Differences by age in us adults. *Social Psychiatry and Psychiatric Epidemiology*, 1–10. <https://doi.org/10.1007/s00127-019-01694-x>.
- Hittner, J. B., Penmetza, N., Bianculli, V., & Swickert, R. (2020). Personality and substance use correlates of e-cigarette use in college students. *Personality and Individual Differences*, *152*, 109605. <https://doi.org/10.1016/j.paid.2019.109605>.
- Hu, M. C., Davies, M., & Kandel, D. B. (2006). Epidemiology and correlates of daily smoking and nicotine dependence among young adults in the united states. *American Journal of Public Health*, *96*(2), 299–308. <https://doi.org/10.2105/AJPH.2004.057232>.
- Kandel, D. B., Kiros, G. E., Schaffran, C., & Hu, M. C. (2004). Racial/ethnic differences in cigarette smoking initiation and progression to daily smoking: A multilevel analysis. *American Journal of Public Health*, *94*(1), 128–135. <https://doi.org/10.2105/AJPH.94.1.128>.

- Lee, S., Kim, S., Kim, Y., Oh, B., Hwang, H., & Park, T. (2019). Pathway analysis of rare variants for the clustered phenotypes by using hierarchical structured components analysis. *BMC Medical Genomics*, *12*(5), 100. <https://doi.org/10.1186/s12920-019-0517-4>.
- Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. <https://doi.org/10.1093/biomet/73.1.13>.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*, 2nd edn. New York: Chapman and Hall.
- National Institute on Drug Abuse. (2019). Drugged driving. <https://www.drugabuse.gov/publications/drugfacts/drugged-driving>.
- Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120–125. <https://doi.org/10.1111/j.0006-341X.2001.00120.x>.
- Robinson, L. A., Murray, D. M., Alfano, C. M., Zbikowski, S. M., Blitstein, J. L., & Klesges, R. C. (2006). Ethnic differences in predictors of adolescent smoking onset and escalation: A longitudinal study from 7th to 12th grade. *Nicotine and Tobacco Research*, *8*(2), 297–307. <https://doi.org/10.1080/14622200500490250>.
- Takane, Y., & Hwang, H. (2005). An extended redundancy analysis and its applications to two practical examples. *Computational Statistics and Data Analysis*, *49*(3), 785–808. <https://doi.org/10.1016/j.csda.2004.06.004>.
- US Department of Health and Human Services (DHHS). (2004). The health consequences of smoking: A report of the surgeon general. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta.
- US DHHS, Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality (CBHSQ). (2013). National survey on drug use and health database. Inter-university Consortium for Political and Social Research (ICPSR) [distributor].

Permutation Test of Regression Coefficients in Social Network Data Analysis



Wen Qu, Haiyan Liu, and Zhiyong Zhang

Abstract In social and behavioral sciences, researchers are interested in the relationships between individuals' attributes and the formation of social relations within a social network. Logistic modeling is a popular approach to address those research interests (Wasserman and Pattison (1996) *Psychometrika* 61:401–425. <https://doi.org/10.1007/BF02294547>). However, the nature of network data (e.g., small size, non-normality, and dependence) violates the assumptions of logistic regression, which can lead to an unreliable inference. To remedy the consequences of these violations with a normal-based hypothesis test, we present the permutation test procedure within the social network framework. The permutation test, on the significance of a regression parameter, can improve the accuracy of the hypothesis decision. In this study, we conducted a simulation to compare the performance of the permutation test and the asymptotic likelihood ratio test under various conditions. The simulation results confirm the advantages of the permutation test as expected.

Keywords Permutation test · Social network · Logistic regression

1 Introduction

1.1 Social Network

Social network analysis (SNA) has been increasingly implemented by researchers in social and behavioral sciences in recent decades. Its application ranges from academic research in fields, such as anthropology, sociology, and psychology, to

W. Qu (✉) · Z. Zhang
University of Notre Dame, Notre Dame, IN, USA
e-mail: wqu@nd.edu; zzhang4@nd.edu

H. Liu
University of California, Merced, CA, USA
e-mail: hliu62@ucmerced.edu

practical usage in areas like politics, communication, and marketing (Hoff et al. 2002; Maya-Jariego and Holgado 2015; Wasserman and Pattison 1996). A social network typically encompasses a set of actors and social relations among them (Wasserman and Pattison 1996). The actors could be individuals or other social units but not limited to human beings. In a network graph, a dyad comprises a pair of actors and the relationship between them. When researchers use SNA to design studies with dyads, they engage in dyadic level analysis. SNA can also be conducted with different levels of actors in the structure. However, in this study, we focus on the dyadic level.

In psychology area, SNA is ideally suited for many subfields including social, developmental, clinical, and educational psychology because it naturally links individuals' attributes and the relationships among them (Clifton and Webster 2017; Saqr et al. 2018; Wasserman and Faust 1994; Westaby 2014). Different techniques of evaluating social network information have been used in the literature, such as the exponential random graph model (ERGM) (Anderson et al. 1999; Robins et al. 2007), latent space model (Hoff et al. 2002; Liu et al. 2018b), and structural equation model (Liu et al. 2018a). The underlying model among the methods mentioned is related to logistic regression. Therefore, in this study, we limit our scope to the multiple logistic regression model with network structured data.

1.2 Permutation Test

Normal-based statistics are often used in statistical hypothesis testing, which, however, requires either normally distributed data or a large sample size. Moreover, independent observations are assumed. In other words, given the level of predictors, the observations on the outcome variable should be independent of each other. However, with social network data, these assumptions can be easily violated because the networks are often of small sizes and dyads in a network are not independent of each other. For example, in a friendship network, for one actor to have three friends in the network, it requires three different actors to have at least one friend. Because of the dependency of social ties, the standard error estimate based on Fisher information is no longer valid, leading to incorrect empirical type I error rates and power rates.

One of the alternative approaches is the permutation test, which can be conducted with social network data due to its flexibility (Farine and Whitehead 2015; Farine 2017; LaFleur and Greevy 2009). Specifically, the permutation test does not require independent observations and normal population distributions for reliable statistical analysis. The only assumption is exchangeability under the null hypothesis. Additionally, the permutation test is more robust to outliers and missing data. In the existing literature, the permutation test is adopted by researchers to address the inaccurate standard error estimates with a small sample size (Potter 2005) or dependence among data (Ke and Zhang 2018). The community of social network analysis has also paid attention to the application of the idea of permutation to

social network data. For instance, Farine (2017) used a permutation procedure to compare the subnetwork structures in different communities. For example, they used the permutation test to compare the degree (i.e., number of social connections an actor has) differences between females and males. In this study, we would generalize or adapt the idea of permutation to a more general modeling framework of social network data. We thus introduce a permutation procedure under the logistic modeling framework to test the effect of a covariate on the presence/absence of a social connection while controlling other potential confounders. The proposed approach is more flexible in at least two aspects. First, it can model the effects of multiple covariates and allows us to control the impact of potential confounders. Second, the covariate of interests is not necessary to be categorical but could also be continuous.

To conduct such a permutation test, we first need to build a distribution of parameters of interest under the null hypothesis using a permutation procedure. This procedure is accomplished by creating a randomized relationship between the outcome variable (social network structure) and the covariate(s) at each permutation step. The critical point of the permutation procedure involves creating a random dataset where only the part of interest is randomly permuted while all the other parts remain unchanged. Finally, the statistical decision of a hypothesis is made based on the achieved significance level (ASL), which is computed as

$$ASL = \#\{|\hat{\beta}_i^*| \geq |\hat{\beta}|\}/N_p, \quad (1)$$

where $\hat{\beta}$ is the logistic coefficient of the original dataset, $\hat{\beta}_i^*$ is the logistic coefficient of the i th permuted dataset, and N_p is the number of permutations.

The subsequent sections of the paper are arranged as follows. Firstly, we use an empirical study, as an example, to illustrate the procedure. Secondly, we present a simulation study to compare the performance of the permutation approach against that of the normal-based test. Finally, we conclude with a discussion.

2 Network Data Permutation

In this section, we illustrate the permutation procedure using an empirical example. The data were collected by the Lab for Big Data Methodology at the University of Notre Dame. The dataset contains the information of 165 junior college students. Each student was asked to report the friendship status (0 = not friends; 1 = friends) with the other 164 students. In addition, information related to gender, smoking status (0 = not smoking cigarettes; 1 = smoking cigarettes), and academic performance (continuous scores from 18 to 87) were also collected and used in this study.

We adopted a logistic regression model to test whether predictors like gender, smoking status, and academic performance were related to the existence of the

friendship. We were particularly interested in the relationship between friendship and smoking status.

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}(p_{ij}) = \alpha + \beta' H_{ij} \tag{2}$$

where y_{ij} is the friendship indicator between students i and j ($0 =$ not friend, $1 =$ friend). The effects of the predictors were tested in the form of nodal covariates H_{ij} which is a vector of three covariates $(HGender_{ij}, HSmoke_{ij}, HGPA_{ij})'$. Further, each of the nodal covariates was defined as follows:

$$HGender_{ij} = \begin{cases} 1, & \text{if students } i \text{ and } j \text{ are of the same gender} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

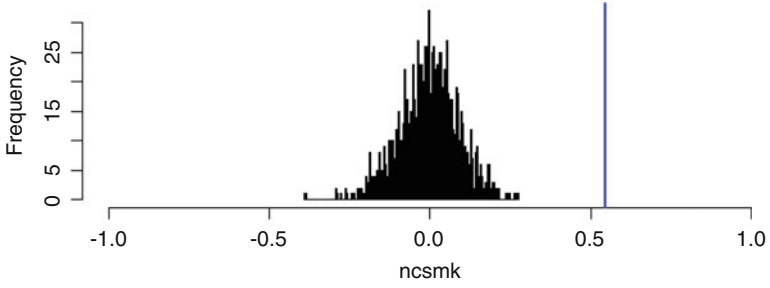
$$HSmoke_{ij} = \begin{cases} 1, & \text{if both students } i \text{ and } j \text{ smoke cigarette} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$HGPA_{ij} = |gpa_i - gpa_j| \tag{5}$$

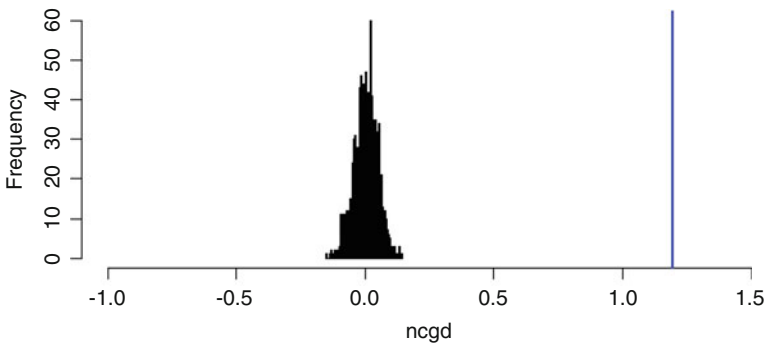
We used a three-step procedure to conduct the permutation test.

1. First, we obtain model parameter estimates through a maximum likelihood approach, and the model parameters are named as $\hat{\beta}_{gender}$, $\hat{\beta}_{smoking}$, and $\hat{\beta}_{gpa}$, which are the coefficient estimates of the three nodal covariates defined by Equation (3), (4), and (5), respectively.
2. Second, we construct a null distribution of parameters. For instance, to construct a null distribution of the coefficient of $Hsmoking$, we randomly permute the data of $Hsmoke$ without replacement, and the resulting data are denoted as $Hsmoke^*$. Meanwhile, data on the other two nodal covariates, $Hgender$ and $Hgpa$, remain the same. We then fit logistic regression model in Equation (2) using data on $Hgender$, $Hsmoke^*$, and $Hgpa$. The parameter estimates of $Hsmoke^*$ is recorded (i.e., $\hat{\beta}_{smoke}^*$). By repeating this step 10,000 times, we could obtain $\hat{\beta}_{smoke,1}^*, \hat{\beta}_{smoke,2}^*, \dots, \hat{\beta}_{smoke,10,000}^*$, which form an empirical distribution of $\hat{\beta}_{smoke}$ under the null model because the link between friendship network data and the smoking covariate is broken with the permuted data.
3. Third, we evaluate the position of the parameter estimates, i.e., $\hat{\beta}_{smoke}$, in the null distribution by computing the ASL using Equation (1).

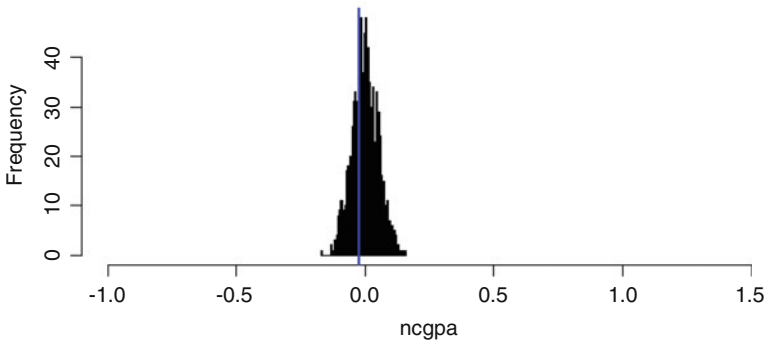
The result for the smoking covariate is shown in Fig. 1a, where the blue line indicates the original parameter estimate $\hat{\beta}_{smoke}$. The ASL value equals to 0, which means that $\hat{\beta}_{smoke}$ is above (or below) all the values of its empirical distribution under the null model. We can then conclude that the observed association (i.e., $\hat{\beta}_{smoke}$) are unlikely to happen by chance. We therefore reject the null hypothesis and conclude that in this dataset, students who both smoke are more likely to be friends.



(a)



(b)



(c)

Fig. 1 Permutation distribution of (a) smoking, (b) gender, (c) GPA effects with friendship data under the null

Similarly, we find that the $ASL_{gender} = 0$ (Fig. 1b), which indicates that with this data, students of the same gender are more likely to be friends. On the other hand, the

GPA plot in Fig. 1c suggests that academic performance is not a significant predictor in the context of existence of friendship in this dataset with the $ASL = 0.635$.

3 Simulation Study

Our goal at this stage is to compare the performance of the permutation approach against that of the normal-based test. To do so, we conduct a simulation study by manipulating three factors: parameter values, sample size, and covariance among dyads.

3.1 Study Design

We use the model in the empirical data analysis as the data generation model. The three covariates are generated in the following way:

$$Gender \sim Binomial(N, p = 0.5),$$

$$Smoking \sim Binomial(N, p = 0.25),$$

$$GPA \sim N(3.3, 0.5).$$

Using the generated data for smoking, gender, and GPA, we then construct the nodal covariate H_{gender} , H_{smoke} , and H_{gpa} using the Equation (3), (4), and (5).

In the data generation procedure, we specified two sets of slope parameters (β), and for each set, three intercept parameter values (α) were used to quantify the network density, which is the friendship percentage in the network. The settings are shown in Table 1. The two sets of slopes indicate whether or not the effects of the covariates exist on the friendship network data. The values of the second set were obtained from the empirical study.

Table 1 Slope and intercept parameter settings

β	α	Network density
(0, 0, 0)	-1.30	10%
	-0.60	30%
	0	50%
(0.50, 0.55, -0.02)	-1.76	10%
	-0.80	30%
	-0.27	50%

Since we focus on undirected binary networks, we only need to generate either the upper triangle or lower triangle part of a network. In a network with N actors,

there are $N(N - 1)/2$ dyads in the upper triangle of the adjacency matrix of a network. Considering the inherently dependent structure of the social network, we incorporate the potential dependence among dyads. We first generated the underlying continuous latent variable \mathbf{Y}^* following a multivariate normal distribution with a specified covariance matrix which ensured the correlated data structure. We then dichotomized \mathbf{Y}^* by setting a threshold of 0, to form a binary \mathbf{Y} variable. Even though the outcome \mathbf{Y} was a binary variable, it still inherited dependence from the underlying continuous variable.

To evaluate the impact of the degrees of dependence among dyads, we considered correlation among them from low to high using values 0, 0.1, 0.3, and 0.5. These correlation values allowed us to examine if the friendship network structure could be fully explained by the covariates or not.

For each combination of parameters, the sample sizes were set to be 20, 35, and 50.¹ The corresponding dyads sizes were 190, 595, and 1225.

In total, the number of conditions was 72 per parameter set, and 1,000 replications of the data were generated under each condition. With each generated replication, we compared the normal-based hypothesis test and the permutation test. The significance level of both tests was recorded (p-value and ASL).

3.2 Evaluation

The evaluation of the normal-based and the permutation tests was based on the comparison between type I error and the achieved statistical power for both tests. The simulation conditions involving the true zero effect slope set were used for investigating the type I error rate, while the other simulation conditions were used for evaluating the power.

3.3 Results

3.3.1 Without Dependence

These conditions were obtained with zero covariance in the friendship data. In Fig. 1a, we plotted the type I error rate against the network density. The permutation test was closer to 0.05 than the normal-based test across almost all network density conditions. Figure 2a shows that the patterns of powers were very similar between

¹In logistic regression, the rule of sample size is $NP \geq 10K/p$ (NP denotes dyads sizes which represents the sample size in logistic regression in this context, k is the number of predictors, p is the proportion of outcome variable equal to 1). Therefore, given 20 people in the friendship network, the NP would be equal to 190. Given three predictors with $p=0.1$, the size would be at least 300. Therefore, $NP=190$, would still be considered a small sample.

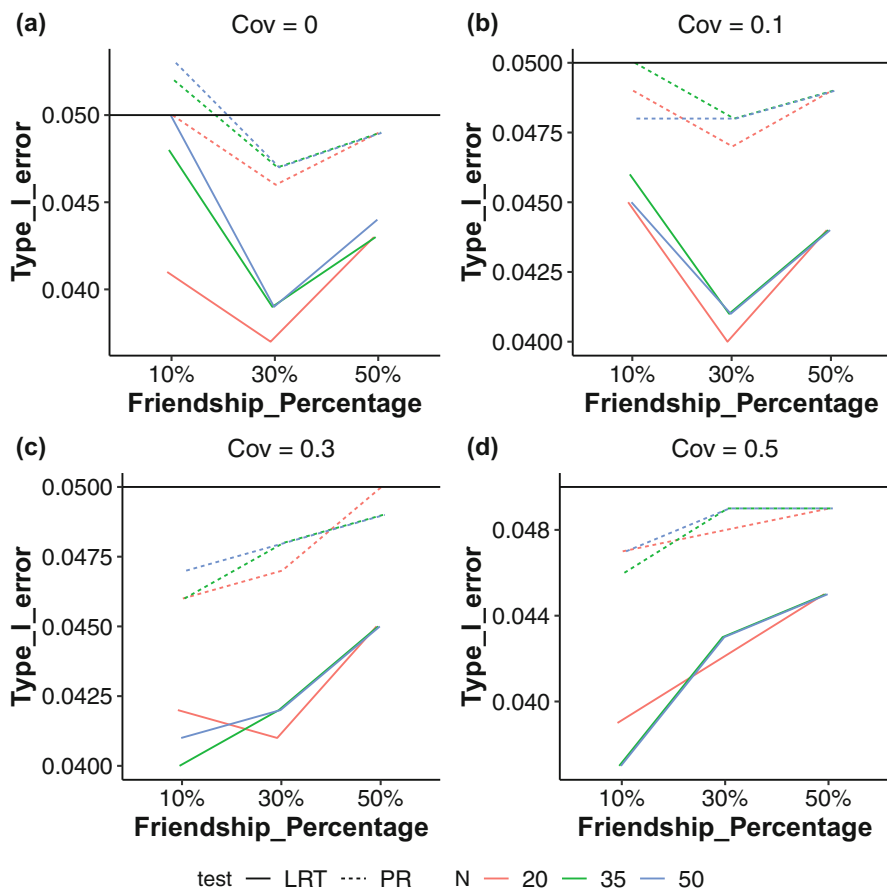


Fig. 2 Type I error rates of likelihood ratio test (LRT) and permutation test (PR)

these two tests. That is, when sample size and network density increased, the power increased. Although the difference was trivial, the permutation test still had a slightly larger power rate across all conditions.

3.3.2 With Dependence

When we introduced dependence into the network data, the results showed a pattern similar to previous outcomes. In Fig. 2b, c, d, the permutation test shows more stable results in terms of type I error rate, compared to the non-dependence conditions, especially when network density is low. In Fig. 3b, c, d, the power shows similar patterns. Specifically, powers increased with a higher network density and a larger

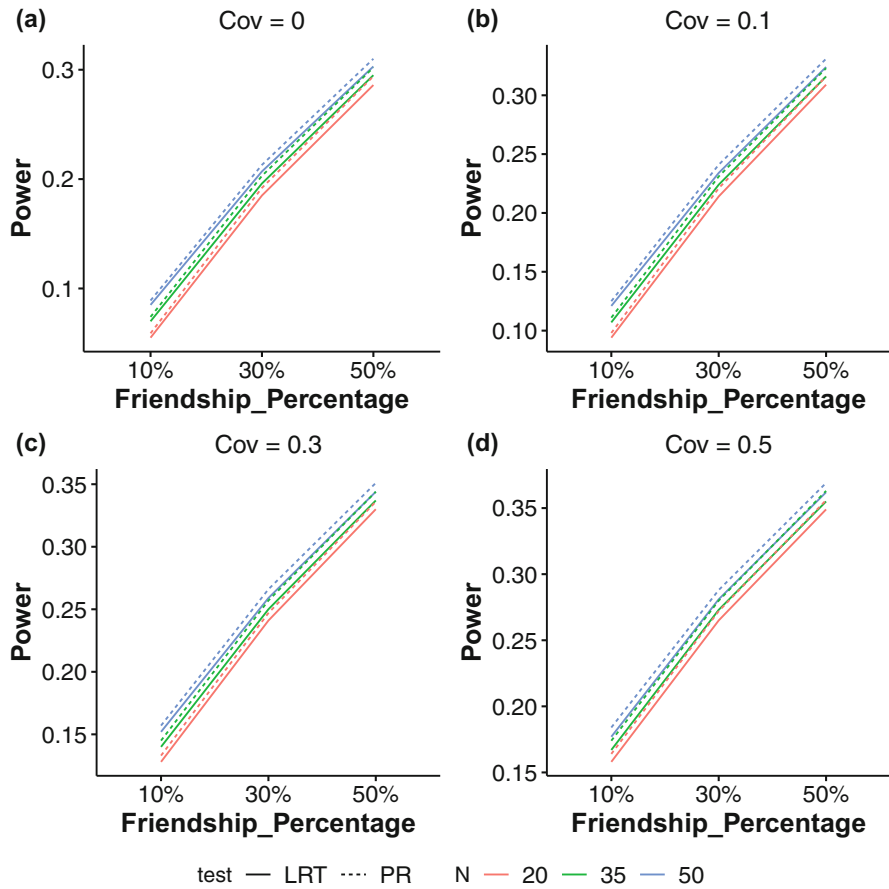


Fig. 3 Statistical power of likelihood ratio test (LRT) and permutation test (PR)

sample size. On the other hand, larger dependence (increasing covariance) increased the power of both tests.

4 Conclusion and Discussion

In this paper, we proposed a permutation test of logistic regression coefficients in social network data analysis. The main advantages of the permutation approach are its capacity to handle relatively small samples and the fact that it does not require the independence assumption to be met. Compared to the traditional normal distribution-based hypothesis test approach, the permutation test can better control type I error and yield slightly larger power across all conditions.

One concern, however, is that using the logistic regression approach with transitivity in the network may result in a biased estimate of the covariate (van Duijn et al. 2009). Future study should consider a transitivity term in the logistic model to address this problem. Furthermore, in the current study, we permuted the covariate. It is also possible to permute the network, which can also be investigated in the future.

References

- Anderson, C. J., Wasserman, S., & Crouch, B. (1999). A p^* primer: Logit models for social networks. *Social Networks*, 21(1), 37–66. [https://doi.org/10.1016/S0378-8733\(98\)00012-4](https://doi.org/10.1016/S0378-8733(98)00012-4).
- Clifton, A., & Webster, G. D. (2017). An introduction to social network analysis for personality and social psychologists. *Social Psychological and Personality Science*, 8(4), 442–453. <https://doi.org/10.1177/1948550617709114>.
- Farine, D. R. (2017). A guide to null models for animal social network analysis. *Methods in Ecology and Evolution*, 8(10), 1309–1320. <https://doi.org/10.1111/2041-210X.12772>.
- Farine, D. R., & Whitehead, H. (2015). Constructing, conducting and interpreting animal social network analysis. *Journal of Animal Ecology*, 84(5), 1144–1163. <https://doi.org/10.1111/1365-2656.12418>.
- Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460), 1090–1098. <https://doi.org/10.1198/016214502388618906>.
- Ke, Z., & Zhang, Z. J. (2018). Testing autocorrelation and partial autocorrelation: Asymptotic methods versus resampling techniques. *The British Journal of Mathematical and Statistical Psychology*, 71(1), 96–116.
- LaFleur, B. J. & Greevy, R. A. (2009). Introduction to permutation and resampling-based hypothesis tests. *Journal of Clinical Child and Adolescent Psychology*, 38(2), 286–294 (PMID: 19283606). <https://doi.org/10.1080/15374410902740411>.
- Liu, H., Jin, I. H., & Zhang, Z. (2018a). Structural equation modeling of social networks: Specification, estimation, and application. *Multivariate Behavioral Research*, 53(5), 714–730. <https://doi.org/10.1080/00273171.2018.1479629>.
- Liu, H., Jin, I. H., Zhang, Z., & Yuan, Y. (2018b). Social network mediation analysis: A latent space approach. arXiv preprint arXiv:1810.03751.
- Maya-Jariego, I., & Holgado, D. (2015). Network analysis for social and community interventions. *Psychosocial Intervention*, 24(3), 121–124. <https://doi.org/10.1016/j.psi.2015.10.001>.
- Potter, D. M. (2005). A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Statistics in Medicine*, 24(5), 693–708. <https://doi.org/10.1002/sim.1931>.
- Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks*, 29(2), 173–191. (Special Section: Advances in Exponential Random Graph (p^*) Models). <https://doi.org/10.1016/j.socnet.2006.08.002>.
- Saqr, M., Fors, U., Tedre, M., & Nouri, J. (2018). How social network analysis can be used to monitor online collaborative learning and guide an informed intervention. *PLOS ONE*, 13(3), 1–22. <https://doi.org/10.1371/journal.pone.0194777>.

- van Duijn, M. A., Gile, K. J., & Handcock, M. S. (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, *31*(1), 52–62. <https://doi.org/10.1016/j.socnet.2008.10.003>.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.
- Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. *Psychometrika*, *61*, 401–425. <https://doi.org/10.1007/BF02294547>.
- Westaby, J. D., Pfaff, D. L., & Redding, N. (2014). Psychology and social networks: A dynamic network theory perspective. *American Psychologist*, *69*(3), 269–284. <https://doi.org/10.1037/a0036106>.

Index

A

- Achieved significance level (ASL), 379–383
- Acquiescence bias
 - balanced scales, 335–336
 - definition, 333
 - internal structure analysis, 342
 - ipsative transformation, 337–338
 - item factor analysis, 343
 - limitations, 343
 - method factor, 334
 - methods of correction, 343
 - negative emotional regulation, 334
 - noise canceling, 336–337, 343
 - re-centering approach, 335
- Analysis of covariance (ANCOVA), 148
- Analysis of variance (ANOVA), 79–80
- Autism spectrum disorder, 278
- Automated test assembly (ATA), 5
- Autoregressive (AR) model, 257
- Average treatment effect (ATE), 148
- Average treatment effect for the treated (ATT), 148

B

- Balanced scales, 335–336
- Bayesian credible intervals (BCIs), 82, 85–87, 89–91
- Bayesian estimation, 100–101
- Bayesian hierarchical modeling
 - MCMC estimation, 80, 82–83, 86
 - R software environment, 87
- Bi-factor IRTree (BF-IRTtree)
 - advantages, 49
 - bi-factor model, 48

- DICs, 50
 - estimated correlation, 50–51
 - expression, 48–49
 - five-point response scale, 50, 52
 - intended-to-be-measured latent factor, 48
 - wording effect, 48, 49, 52
- Binary pseudo items (BPI), 46, 47
- Bounded response time (BRT) model, 97–98, 101–102

C

- CAI+exposure method, 208
- Career paths, academia
 - Anderson, C.J., 6–7
 - von Davier, A., 11–12
 - Embretson, S., 7–8
 - Meulman, J., 9–10
 - Moustaki, I., 10–11
 - Wiberg, M., 12
 - Yan, D., 13–14
- Child Behavior Checklist (CBCL), 275
- Chi square of Mantel-Haenszel (χ^2 MH), 71, 72, 74, 76
- Classical test theory (CTT), 213
- Cognitive coping strategies, 300, 303, 305, 306, 315
- Collaborative institutional training initiative (CITI), 308
- Composite likelihood
 - Bayesian estimation, 40–41
 - maximum likelihood, 38–39
 - person parameter estimation, 32, 42
 - weighted likelihood estimation, 41–42

- Computerized adaptive testing (CAT)
 CAI method, 202, 203
 discrimination parameter, 202
 drawbacks, 202
 educational assessments, 201
 evaluation criteria, 206–207
 item selection methods, 207–209
 research and development, 5
 simulation design, 205–206
 SMB method, 202
a-stratification, 202
- Conditional standard error of measurement (CSEM)
 conceptual comparison, 215–216
 CTT/IRT, 213
 empirical comparison, 216–220
 fixed precision CAT, 215
 GVS scale transformation, 223
 score scale, 213
 stabilizing transformation, CAT, 220–222
 transitioning linear tests, 222–223
- Confirmatory factor analysis (CFA), 174, 183, 291
- Constructed-response (CR) items
 Bayesian algorithm, 267
 data cleaning, 266
 deviance information criterion, 266–267
 DIC values, 267
 economic test, 269
 instructional effects, 273
 instruments, 265
 LDA, 264
 model selection, DIC, 267
 participants, 265
 qualitative evidence, 263
 rubric-based words, 268
 sLDA, 264–265
 topic models, 264
- Continuous *a*-stratification index (CAI), 202–204
- Co-occurring substance use
 adverse public health consequences, 366
 categories, 366
 CV, 367
 data and model specification, 370
 ERA, 367
 model specification, 367–369
 parameter estimation, 369
 predictors, 366
 regularization and empirical results, 373–374
 significance testing, 369
 substance use variables, 370–373
- Counterbalanced (CB) design, 122
- Covariance matrix
 factor analysis, 187
 FA model, 186
 graphical lasso, 189–190, 199
 high-dimensional settings, 186
 ML method, 196
 MSEs, 191
 multivariate analysis, 186
 off-diagonal elements, 197
 probabilistic PCA, 199
 regularized methods, 186
 ridge maximum likelihood method, 189
 ridge-type estimation, 186
 simulation, 191–192
 sparse precision matrix, 186
 ULS, 187, 188
 Woodbury identity, 187–188
- Cross-validation (CV), 130, 189, 228, 256, 366
- Cumulative distribution functions (CDF), 121
- D**
- Differential evolution (DE), 230–232, 243
- Differential item functioning (DIF), 276
 and ASD, 282
 bias analysis, 73
 chi square, MH, 72, 74
 ETS, 71
 item analysis, 71–72, 74
 measurement bias across groups, 276
 simulation-based procedure, 280
 uniform bias, 71
- Duble-smoothing (DS), 130
- Dynamic time warping (DTW), 250–253
- E**
- Educational Testing Service (ETS), 11, 13, 71, 72, 75, 78
- Embedded design, 301–303, 305, 306, 316, 317
- Equipercntile transformation, 121
- Equivalent groups (EG) design, 122
- Euclidean distance (EucD), 257
- Expected a posteriori (EAP), 86, 173, 176, 183
- Exploratory factor analysis (EFA), 291
- Exponential random graph model (ERGM), 377–378
- F**
- Factor analysis (FA) models, 172, 187, 188, 196
- Fingerprint method, 229, 230, 243

- Fisher information matrix, 32, 36, 38, 39, 202, 203
- Forced-choice data
 - binary coding, 33–34
 - classical scoring methods, 32
 - efficiency gains, 39
 - IRT approach, 32
 - likelihood function, 34–36
 - Thurstonian MIRT model, 33, 41, 42
- Four-parameter logistic (4PL) model, 56, 57
- Four-parameter normal ogive (4PNO) model
 - cognitive and non-cognitive tests, 56
 - developments, 64
 - estimation methods, 57
 - hierarchical framework of response and RT, 56, 57, 64
 - high-stakes tests, 56
 - low-stakes tests, 56
 - MH and MML methods, 57
 - posterior distribution, 58–59
 - psychopathology measurement, 56
 - with RT, 56, 60
 - simulation, 59–60
 - 3PL model, 56
- Freeze control, 206
- G**
- Gandi psychometric model, 305, 308, 315
- Gender equality, 5
- Gender gap, in STEM, 2, 3
- Generalized additive models (GAM), 348, 350, 359–361
- Generalized least squares (GLS) methods, 186
- Gibbs sampling approach, 57, 59, 64–66
- Graded response model (GRM), 278
- Grade point average (GPA), 136
- Graphical lasso, 189–190
- Graphical representation, 71, 74, 75, 178
- Guessing parameter, 70, 114, 206
- H**
- Hierarchical bounded response time (HBRT) model, 99, 105, 107
- Hierarchical framework of response and RT, 56, 57, 64
- Hierarchical joint model
 - Bayesian estimation, 100–101
 - BRT model, 97–98, 101–102
 - HBRT model, 99, 105, 107
 - parameter recovery, 102–105
 - simplex distribution, 96–97
- Higher education, 285, 287–289
- Household Food Security Survey Module (HFSSM)
 - behaviors, 320
 - coding, 322
 - dichotomous analysis, 326
 - dichotomous Rasch model, 322–323
 - food insecurity measurement, 329
 - household fit statistics, 326
 - item summary, 324
 - partial credit category statistics, 330
 - partial credit Rasch model, 323–324, 326–329
 - participants, 321–322
 - policymakers, 330
 - polytomous items, 320
 - psychometric properties, 320
 - Wright map, 324, 325
- Hyperprior distributions
 - description, 80
 - half-*t* and half-Cauchy distributions, 84–85
 - inverse-gamma distribution, 84
 - levels, 85
 - MCMC approach, 82
 - and prior distributions, 83
 - random-effect variances, 83
 - relative bias, 88
 - relative efficiency, 90
 - specification, 80
 - uniform distribution, 83–84
 - variance parameters, 83
- I**
- Identification bounds, 137, 139–141
- Illness, 300, 303, 305, 306
- Intensive longitudinal data, 247
 - big data, 249
 - limitation, 248
- Inter-rater reliability (IRR)
 - bias, 354–356
 - CI coverage, 355, 358
 - convergence and fitting times, 354
 - credibility of ratings, 348
 - data generation, 350–351
 - data simulation, 361–363
 - definition, 81
 - GAM models, 350
 - heterogeneity, 349
 - ICCs, 81, 90, 91, 348
 - inverse-gamma distribution, 84
 - LME models, 348, 350
 - MCMC estimation, 82–83
 - model implementation, 351–352
 - power and error-rate, 355, 359

- Inter-rater reliability (IRR) (*cont.*)
 random-effect model, 348
 RMSE, 354, 355, 357
 simulation evaluation, 353
 variances components, 81
- Intra-class correlation coefficients (ICCs)
 Bayesian hierarchical models, 87
 BCI coverage rates, 89
 interrater reliability, 81
 MCMC estimation, 82–83
 posterior means, 86
 posterior modes, 86
 random-effect variances, 85
 relative bias, 88
 relative efficiency, 90, 91
- Invariant measurement
 item calibration, 25
 item-invariant measurement of persons,
 24–25
 person-invariant item calibration, 24
 person measurement, 25
 requirements, 25
 variable map, 25
- Inverse weights of propensity scores (IWPS),
 148
- IRT observed-score equating (IRTOSE), 122
 IRT true-score equating (IRTTSE), 122
 Item-invariant measurement of persons, 24–25
- Item response theory (IRT)
 Bayesian approach, 73
 factor score estimation, 171
 graphical representation, 74
 guessing parameter, 70
 MAP/EAP estimation, 183
 models, 112, 114, 122
 IPL-G, 71
 parameter estimation, 70–71
 for polytomous data, 4
 properties, 172
 Rasch measurement theory, 20, 26
 regression, 172
 student's ability, 70
 Thurstonian IRT model, 33–34, 41, 42 (*see also* Forced-choice data)
- Item response theory tree (IRTtree)
 BPI, 46, 47
 five-point response scale, 47
 NW items, 46
 PW items, 46
 response processes, 45
 response styles, 45–46
 three response processes, 46
 tree-like structure, 46
- K**
 Kernel equating (KE) framework
 continuous score distribution, 113
 equated scores, 117, 118
 error, 117
 evaluation measure and study design, 116
 IRT modeling, 112, 144
 with Kequate, 128–130
 log-linear modeling, 112–114
 model classes, 117
 model selection, 112, 116–119
 presmoothing model selection criteria, 115
 SEE, 117, 118
 SNSequate, 131–132
 steps, 113
 test score equating, 111
- L**
 Latent Dirichlet allocation (LDA), 264
 Linear mixed-effect (LME) models, 348, 350,
 359–361
 Logistic regression, 378–380, 385
- M**
 Machine learning community, 250
 Machine learning problems
 minimization procedure, 164
 PARAFAC, 164
 Penrose regression, 165
 SCoTLASS, 164
 SPCA procedure, 164
 Tucker3's loss function, 165
- Mantel Haenszel (MH)
 chi square (χ^2 MH), 71, 72, 74, 76
 MH Delta statistic, 72
- Marginal maximum likelihood (MML), 57
- Markov chain Monte Carlo (MCMC)
 ANOVA, 79–80
 BCIs, 82
 estimation of ICCs, 80, 82
 hyperparameters, 82
 MLE, 80
- Maximum a posteriori (MAP), 173
- Maximum likelihood (ML)
 Bayes posterior mean estimator, 175
 log likelihood, 176
 modal value, 175
 Newton-Raphson algorithm, 175
- Maximum likelihood estimation (MLE),
 36–38, 80, 86, 206
- Mean maximum absolute (MMA), 149

Mean squared errors (MSEs), 151–152, 191–192, 202

Measurement bias

- AQ-10, 281
- autism spectrum disorder, 278
- bivariate normal distribution, 277
- classification agreement, 277, 282
- diagnostic measures/screeners, 275
- DIF, 276
- graphical approaches, 277
- groups, 276
- implementation, 278
- invariance, 282
- item parameters, 282
- limitation, 278
- measurement invariance, 276
- mental health disorder, 275
- method
 - empirical example, 278–279
 - general procedure, 280–281
- mixed distribution, 282
- Monte Carlo simulations, 278, 282
- practical implications, 276
- probability, 276
- screening performance, 277, 278
- sensitivity, 281, 282
- significance testing procedures, 276
- simulation-based procedure, 278, 282, 283
- specificity, 281
- unidimensional linear factor structure, 277
- variance, 281

Mental health, 366, 370

Metropolis-Hastings (MH) algorithms, 57

Mixed-effect models, 349

Mixed method, 301–303

Modal estimation, 179–180

Monte Carlo simulation, 278

Multidimensional Item Response Theory (MIRT), 334

Multinomial logistic regression (MLR) model, 149

Multiple factor model

- log likelihood, 178
- normal likelihood, 178–179
- observation missing, random, 182
- regression coefficients, 177

Multi-process IRTree, 45

Multivariate regression

- LASSO solution, 164
- penalized loss function, 163
- Procrustes penalty function, 162

N

Nearest neighbor (NN) classification, 256

Negatively worded (NW) items, 46, 48–52

Neural networks, 152, 153

Non-equivalent groups with anchor test (NEAT) design, 112, 122

Non-equivalent groups with covariates (NEC) design, 122

Non-parametric bootstrap

- collection information, 233–234
- differential evolution, 231–232
- finding estimates, optimization, 231
- finite mixtures
 - data generation, 239
 - generating modal parameters, 238–239
 - min-log-likelihood function, 237
 - probability density function, 238
 - univariate normal density functions, 237
- generating new datasets, 230–231
- multiple cost functions, 232–233

Nonverbal synchrony, 249

O

Objectivity in science, 23–24

One-factor model

- Gaussian likelihood, 175
- mean and variance, 182–183
- regression parameter, 174
- unidimensional IRT, 173

One parameter logistic (1PL) model, 73, 114

One parameter logistic with guessing (1PL-G) model, 73

Optimization procedure, 240

P

Partial identification approach

- conditional expectation, 137–138
- marginal effect, 138–140

Pedagogical competencies

- definition, 286
- descriptive analyses, 292–293
- exploratory factor analyses, 294–296
- higher education, 285, 287
- institutions, 286
- measurement, 287–288
- method
 - data analysis, 291
 - instrument, 288–289
 - participants, 289–291

- Pedagogical competencies (*cont.*)
 national policies, 286
 post-secondary education, 285
 psychometric evaluation, 286
 reliability analyses, 293
- Penalized estimation
 estimated core arrays, 169
 interpretable solutions, 161
 machine learning problems, 164–165
 multivariate analysis procedures, 167
 penalty function, 162
 SPCA (*see* Sparse principal component analysis (SPCA))
 target matrix, 168
See also Multivariate Regression
- Penalty function, 162
- Permutation test
 with dependence, 384–385
 flexibility, 378
 independent observations, 378
 logistic modeling framework, 379
 network data permutation, 379–382
 normal-based statistics, 378
 null hypothesis, 378, 379
 simulation study
 evaluation, 383
 study design, 382–383
 SNA, 377–378
 without dependence, 383–384
- Person-invariant item calibration, 24
- Person parameter estimation
 Bayesian estimation
 composite likelihood, 40–41
 genuine likelihood, 39–40
 independence likelihood, 40
 maximum likelihood
 composite likelihood, 38–39
 efficiency, 39
 genuine likelihood, 36
 independence likelihood, 37–38
 precision, 32
 Thurstonian IRT models, 42
 weighted likelihood estimation
 composite likelihood, 41–42
 genuine likelihood, 41
- Philosophy of measurement, 27
- Positively worded (PW) items, 46, 48–52
- Percent relative error (PRE), 130
- Predictive validity
 behavior of interest, 136
 conditional distribution, 137
 definition, 136
 educational measurement literature, 137
 estimation, identification bounds, 141
 evolution, 141
 explicit expression, 143–144
 higher education, 136
 identification bounds, 137
 language and communication test, 142
 marginal effects, 142
 mathematics test, 142
 methodological approach, 137
 non-observed group, 143
 observed group, 143
 partial identification approach, 137
 problem of learning, 136
 quality of selection, 136
 selection process, 141
 selection test, 136, 137
 statistical procedures, 136
- Principal component analysis (PCA), 291
- Procrustes penalty function, 162
- Professional successes, 2
- Programme for International Student Assessment (PISA), 95, 105–107
- Propensity scores (PS)
 average absolute maximum covariate balance, 153
 average parameter recovery statistics, 153–155
 condition and estimation method, 151–152
 covariates, 154
 data mining models, 154
 definition, 148
 difference-in-differences designs, 149
 estimation of multiple-group propensity scores, 157
 intervention and comparison groups, 147–148
 machine learning models, 156
 MG IWPS and trimming, 156
 multiple-group (MG), 148, 149
 non-randomized designs, 156
 observational and quasi-experimental designs, 148
 regression models, 153
 simulation study, 149–151
 social and health sciences, 147
 treatment effect estimation, 148, 151–152, 156
 treatment groups, 156
- PSU standardized test, 72, 74, 75, 78, 141
- Psychometric models
 linear models, 25
 modern measurement theories, 26
 scaling tradition, 26
 test-score tradition, 25
- Psychometrics, 300, 301, 303–306

Psychometrics history

- bifactor model, 5
- CAT research and development, 5
- eugenics research, 4
- factor analytic techniques, 4
- foundation, 3
- IMPS19 session, 2
- instructional materials, 4
- IRT models, 4
- journal, 3
- LOGIST software, 4
- missionary, 4
- Nightingale Rose Diagram, 3
- “a personal equation”, 3
- Polar Area Diagram, 3
- Rule-Space model, 5
- “test guru” psychometrician, 4

Psychoperiscope

- behavioral vs. cognitive coping, 300
- clinimetrics, 301
- cognitions, 300
- conceptual framework, 303–305
- coping, 300
- deductive and inductive approaches, 315
- exploratory analysis, 315
- Gaussian graphical model, 300
- item correlation coefficients, 309
- item statistics, 309
- item-total correlation, 311
- item-total statistics, 310, 312–314, 317
- logical and comprehensive approach, 301, 316
- materials, 308
- mediating-moderating effects, 315
- methods
 - research design and study setting, 305–306
 - target population and sample participants, 306–307
- mixed method, 301–303, 316, 317
- multimethod method, 301, 316
- network structure, 300
- procedure, 308–309
- psychometrics, 301
- QOL measures, 300, 301
- qualitative and quantitative approaches, 317
- statement of problem and purpose, 303

Q

- Quality of life (QOL), 300, 301, 303, 305, 306, 315

R

- Random effects, 80, 82–87
- Rasch measurement theory
 - bibliometric search, 20
 - complementary approach, 22
 - concepts, 23
 - description, 20
 - dichotomous and polytomous responses, 21
 - frequency of articles, 21
 - guiding research questions, 27
 - invariant measurement (*see* Invariant measurement)
 - IRT, 20
 - models for measurement, 26, 27
 - objectivity in science, 23–24
 - psychometric models, 25–26
 - research, 27
- Rasch models, 320
 - binomial trials, 21, 26
 - dichotomous, 21, 26
 - extensions, 22
 - facets, 26
 - partial credit, 21, 26
 - philosophy of measurement, 27
 - Poisson count, 21, 26
 - rating scale, 21, 26
- Regression solutions, 180–182
- Regularization, 367, 368, 373, 374
- Reliability, 286, 291–293
- Reliability and structure validity, *see* Pedagogical competencies
- Re-sampling methods
 - bootstrap, 228
 - computational burden, 228
 - cross-validation methods, 228
 - data characteristics, 229
 - degree of overlap, 228
 - fingerprint method, 229
 - flexibility, 228
 - optimization procedures, 230
 - prepaid method, 229
 - statistical inference tasks, 228
 - synergized bootstrap method, 230
- Response time (RT)
 - and accuracy, 96
 - application, 105
 - BRT model, 97–98
 - with 4PNO (*see* Four-parameter normal ogive (4PNO) model)
 - hierarchical framework of response and RT, 56, 57, 64
 - lognormal distribution, 96

- Response time (RT) (*cont.*)
 measurement accuracy, 56
 mental activity, 96
 simplex distribution, 96–97
- Root mean square error (RMSE), 60–62, 64, 104, 105, 206, 353
- R packages for test equating
 artificial intelligence, 132
 CDF, 121
 data collection designs, 121–122
 equating methods, 122–123
 equipercntile transformation, 121
 IRT parameter linking, 124–127
 IRT true-score and observed-score equating, 127–128
 machine learning, 132
 NEAT design, 123
 test equating methods, 122
 traditional methods, 123–124
- S**
- Standard error of equating (SEE), 130
 α -Stratification method, 202
 Second-order generalized estimating equations (GEE2), 348, 349, 359–361
 Sensitivity, 281
 Simplex distribution, 96–98, 101, 105, 107
 Simulation
 acquiescence correction, 338–339
 automatic correction, 341
 database, 340
 re-centering approach, 340
 standard deviation, 340–342
 Simulation-based procedure, 280
 Single group (SG) design, 122
 Slipping parameters, 56, 59–61, 63, 64
 Social network analysis (SNA), 377–378
 Societal structures, 5–6
 Socioeconomic status, 307
 Sparse principal component analysis (SPCA)
 interpretability of solutions, 161
 minimization procedure, 164
 wine data
 components, 166, 167
 PARAFAC, 167
 Tucker3, 167
 UCI Machine Learning Repository, 166
 varimax-rotated loading matrix, 166
 Specificity, 277, 280–282
 Squared correlation (sqrC), 250, 251, 257
 Standard DE method, 241
 Standard error of measurements (SEM), 208
 STEM domains, 2
 Substance initiation age, 366, 370
 Supervised LDA (sLDA), 264–265
 economic tests, 272
 model prediction and score proportion, 271
 model selection, 270
 regression, 269
 topic proportion matrix, 270
 Supplemental Nutrition Assistance Program (SNAP), 329
 Synergized bootstrap
 comparing methods, 241
 DE populations, 235
 D*M optimization, 243
 fingerprint method, 243
 initial updating scheme, 234
 standard DE method, 241–242
 traditional DE, 236–237
- T**
- Three-parameter logistic (3PL) model, 114, 125, 216
 bias analysis, 73
 with Cox PH model, 57
 4PL and 4PNO, 56
 interpretation, 70
 measurement theories, 26
 and 1PL-G, 71
 Thurstonian IRT model
 forced-choice data
 binary coding, 33–34
 model equations, 34
 notation, 33
 normal-ogive Thurstonian models, 41
 person parameter estimation, 32, 42 (*see also* Person parameter estimation)
 Time-invariant membership, 248
 Time series classification, 248
 alignment/synchrony, 249
 balanced/imbalance condition, 258, 259
 clustering problem, 249
 DTW, 250
 Euclidean distances, 250
 intraindividual time profiles, 249
 nearest neighbor classification, 256–257
 phase alignment, DTW, 251–253
 similarity design/data generation, 257–258
 similarity measures, 251
 WCDmin, 258
 windowed cross lagging, 253–255
 Transparency, 6, 321
 True-/false keyed items, 337, 338
 Two parameter logistic (2PL) model, 73, 114, 278

U

- Unidimensional likelihood-based score estimates
 - IRT scoring, 172–173
 - ML estimation, 175–176
 - one-factor model, 174–176
 - quick closed form computation, 176–177
- Unweighted least squares (ULS) method, 187
- US Department of Agriculture (USDA), 320

V

- Variance components, 80–82, 360

W

- “Watterson estimator”, 5
- Windowed cross lagging
 - breaking time-dependent data, 253
 - data sequences, 254
 - DTW, 254
 - procedure, 255
 - R scripts, 255
- Wording effects
 - BF-IRT model, 48–49, 52
 - empirical data, 50
 - IRTree models, 46
 - NW items, 48, 49, 51
 - PW items, 48, 49