Springer Proceedings in Mathematics & Statistics

Marie Wiberg · Steven Culpepper
Rianne Janssen · Jorge González
Dylan Molenaar *Editors*

Quantitative Psychology

The 82nd Annual Meeting of the Psychometric Society, Zurich, Switzerland, 2017

Springer

Springer Proceedings in Mathematics & Statistics

Volume 233

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 · Steven Culpepper
Rianne Janssen · Jorge González
Dylan Molenaar
Editors

Quantitative Psychology

The 82nd Annual Meeting of the Psychometric Society, Zurich, Switzerland, 2017

Springer

Editors

Marie Wiberg
Umeå School of Business,
Economics and Statistics
Umeå University
Umeå
Sweden

Jorge González
Faculty of Mathematics
Pontificia Universidad Católica
de Chile
Santiago
Chile

Steven Culpepper
Department of Statistics
University of Illinois
at Urbana-Champaign
Champaign, IL
USA

Dylan Molenaar
Department of Psychology
University of Amsterdam
Amsterdam
The Netherlands

Rianne Janssen
Faculty of Psychology
and Educational Sciences
KU Leuven
Leuven
Belgium

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-77248-6 ISBN 978-3-319-77249-3 (eBook)
https://doi.org/10.1007/978-3-319-77249-3

Library of Congress Control Number: 2018934883

Mathematics Subject Classification (2010): 62P15, 62-06, 62H12, 62F15, 62-07

Printed on acid-free paper

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

Preface

This volume represents presentations given at the 82nd annual meeting of the Psychometric Society, organized by the University of Zurich, and held in Zurich, Switzerland, during July 17–21, 2017. The meeting was one of the largest Psychometric Society meetings in the Society's history, both in terms of participants and number of presentations. It attracted 521 participants, with 295 papers being presented, of which 91 were part of a symposium. There were 105 poster presentations, 3 pre-conference workshops, 3 keynote presentations, 4 invited presentations, 2 career award presentations, 4 state-of-the-art presentations, 1 dissertation award winner, and 22 symposia.

Since the 77th meeting in Lincoln, Nebraska, Springer publishes the proceedings volume from the annual meeting of the Psychometric Society so as to allow presenters to quickly make their ideas available to the wider research community, while still undergoing a thorough review process. The first five volumes of the meetings in Lincoln, Arnhem, Madison, Beijing, and Asheville were received successfully, and we expect a successful reception of these proceedings too.

We asked authors to use their presentation at the meeting as the basis of their chapters, possibly extended with new ideas or additional information. The result is a selection of 34 state-of-the-art chapters addressing a diverse set of psychometric topics, including item response theory, factor analysis, causal inference, Bayesian statistics, test equating, cognitive diagnostic models, and multistage adaptive testing.

Umeå, Sweden	Marie Wiberg
Champaign, IL, USA	Steven Culpepper
Leuven, Belgium	Rianne Janssen
Santiago, Chile	Jorge González
Amsterdam, The Netherlands	Dylan Molenaar

Contents

Optimal Scores as an Alternative to Sum Scores

Marie Wiberg, James O. Ramsay and Juan Li

Abstract This paper discusses the use of optimal scores as an alternative to sum scores and expected sum scores when analyzing test data. Optimal scores are built on nonparametric methods and use the interaction between the test takers' responses on each item and the impact of the corresponding items on the estimate of their performance. Both theoretical arguments for optimal score as well as arguments built upon simulation results are given. The paper claims that in order to achieve the same accuracy in terms of mean squared error and root mean squared error, an optimally scored test needs substantially fewer items than a sum scored test. The top-performing test takers and the bottom 5% test takers are by far the groups that benefit most from using optimal scores.

Keywords Optimal scoring · Item impact · Sum scores · Expected sum scores

1 Introduction

Test scores must estimate the abilities of the test takers in a manner that is both accurate and unbiased, since they are used in many settings to make decisions about test takers. Sum scores (or number correct scores) have in the past been a common test score choice as they are easy for test takers to interpret and are easy to compute. Scores built on parametric item response theory (IRT; see Lord 1980; Birnbaum 1968) have also been used, although almost exclusively by test constructors,

M. Wiberg (✉)
Department of Statistics, USBE, Umeå University, 901 87 Umeå, Sweden
e-mail: marie.wiberg@umu.se

J. O. Ramsay
Department of Psychology, McGill University, Montreal, Canada
e-mail: james.ramsay@mcgill.ca; ramsay@psych.mcgill.ca

J. Li
Department of Mathematics and Statistics, McGill University, Montreal, Canada
e-mail: juan.li3@mcgill.ca

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_1

since test takers usually find it hard to understand the meaning of the parametric IRT scale scores, which may take any value on the real line. Test takers tend not to be convinced that a score of zero represents average performance. A further problem is that commonly not all items are satisfactorily modeled with parametric IRT models, even in large-scale tests that have been carefully developed.

A choice other than using parametric IRT models is to use nonparametric methods to estimate test takers' ability and the item characteristic curves (ICC). Nonparametric IRT has been used in several studies in the past. Mokken (1997) examined nonparametric estimation and how it worked in connection to monotonicity. Ramsay (1991, 1997) proposed ICC estimation using kernel smoothing over quantiles of the Gaussian distribution. This technique gave fast and reasonably accurate ICC estimation, and was implemented in the computer program TestGraf. Rossi et al. (2002) and Ramsay and Silverman (2002) used the expectation-maximization (EM) algorithm to optimize the penalized marginal likelihood, and the estimates came close to the three-parameter logistic IRT model as the smoothing penalty was increased. Ramsay and Silverman (2005) proposed a nonparametric method for not strictly monotonic curve estimates. Woods and Thissen (2006) and Woods (2006) proposed a method for simultaneously estimating item parameters using a spline-based approximation to the ability distribution. Lee (2007) made a comparison of a number of nonparametric approaches.

As yet another alternative approach to test scoring, this paper will focus on optimal scoring. This method was proposed by Ramsay and Wiberg (2017a) and practical concerns were discussed in Ramsay and Wiberg (2017b). The basic idea behind optimal scoring is to use the interaction between the test takers' responses on each item and the impact of the corresponding items on the estimate of their performance by letting high-slope items be more influential than low-slope items when calculating the test scores. Optimal scoring differs substantially from previous nonparametric approaches in several important ways. First, it uses a faster and more sophisticated approach than the EM algorithm. Second, it uses spline basis expansions over non-negative closed intervals to facilitate the interpretation of the test scores for the test takers. A featured shared with the other nonparametric methods is that it succeeds to get well-fitting ICC's when parametric IRT models fail to give a good fit. The overall aim of this paper is to discuss the nonparametric IRT based optimal scores as a good alternative to sum scores and expected sum scores and to illustrate this with real and simulated test data. This paper also differs from Ramsay and Wiberg (2017b) by extending the comparison to include expected sum scores.

The next section describes the quantitative skill test used as an illustration, followed by a third section where three different test scores are defined. The fourth section contains a description on how to estimate the ICC's with optimal scoring. In the fifth section a comparison between sum scores, expected sum scores and optimal scores are given. The paper ends with a short discussion, which includes some concluding remarks.

2 A College Admission Test and Its Empirical Test Distribution

The data used in this paper come from an administration of the Swedish Scholastic Assessment Test (SweSAT), which is a binary scored multiple-choice college admissions test. The SweSAT contains a verbal and a quantitative parts, each containing 80 items. Sum scores are routinely used in the SweSAT, although the obtained scores are equated to scaled scores, which are comparable over test administrations and these scaled scores are used by test takers in their college applications. A sample of 30,000 test takers who took the quantitative part of the SweSAT is used throughout the paper and the empirical distribution of the sum scores is displayed in Fig. 1. From this figure, we can draw the conclusion that a majority of the test takers found the SweSAT difficult, with a median score of 35, a lowest score of 4 and no test taker with a perfect score. In Fig. 1 we have added a smooth function of the distribution, which was constructed from a B-spline expansion of the log density (Ramsay et al. 2009), since the empirical distribution of the sum scores did not resemble any of the common parametric densities.

Note that in general the distribution of the θ estimates can be transformed whether or not a parametric or nonparametric IRT is used. Suppose we have a one-to-one increasing and smooth transformation $\varphi = h(\theta)$, then there exists an alternative item response function $P_i^*(\varphi)$, so that $P_i^*(\varphi) = P_i(\theta)$. Thus, we can transform any specified distribution of θ into an alternative distribution of φ. For example, we transform from the whole real line into a closed interval such as $[0, n]$ by defining $\varphi = n/(1 + e^{-\theta})$.

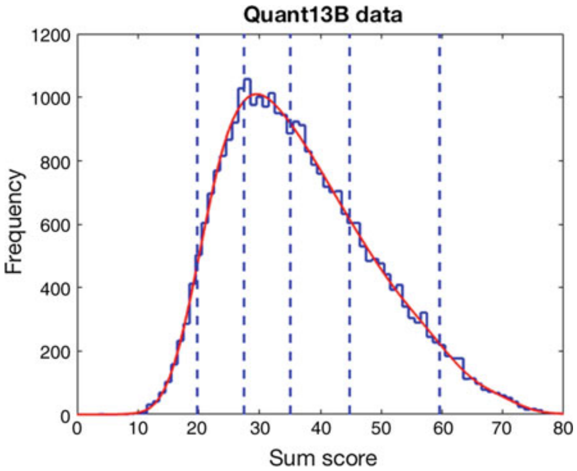


Fig. 1 The empirical distribution of sum scores. The blue histogram indicates the number of test takers within each score range, the red line indicates the smooth density function, and the blue dotted lines are 5, 25, 50, 75 and 95% quintile lines respectively

3 Three Test Scores

Let S_j denote the sum score of test taker j $(j = 1, \ldots, N)$ and define it as the number of correctly answered binary items. Let $P_i(\theta_j)$ be the probability that a test taker with ability level θ_j answered item i $(i = 1, \ldots, n)$ correctly. The expected sum scores are defined as

$$E_j = \sum_i^n P_i(\theta_j). \tag{1}$$

Note, a commonly used expected score uses parametric IRT to model $P_i(\theta_j)$.

To estimate optimal scores O_j (Ramsay and Wiberg 2017a) we focus on estimating the more convenient log-odds function

$$W_i(\theta) = \log\left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right). \tag{2}$$

To estimate $W_i(\theta)$ we can use B-spline basis function expansions

$$W_i(\theta) = \sum_k^K \gamma_{ik}\psi_{ik}(\theta), \tag{3}$$

where for each item i, γ_{ik} is the coefficient of the basis function, $\psi_{ik}(\theta) = B_k(\theta|\xi, M)$ is the B-spline basis function, ξ is a knot sequence, K is the number of spline functions and M is the order of the spline. The advantage of this approach is that B-spline basis functions are easily expanded in dimensionality and they give stable and fast computations.

The left panel of Fig. 2 contains the P_i estimates of the 80 item response functions and the right panel of Fig. 2 shows the W_i estimates of the SweSAT data. From Fig. 2 we learn that items vary in shape of their ICC and their corresponding log-odds functions W_i. Some items are very difficult, other items have low discrimination. If U_{ij} is test taker j's response (0/1) to item i and if either $P_i(\theta)$ or its counterpart $W_i(\theta)$ are either known or we can condition on estimates on them, then the left hand side of

$$\sum_i^n \left[U_{ij} - P_i(\theta)\right] \frac{\mathrm{d}W_i}{\mathrm{d}\theta} = 0 \tag{4}$$

is the derivative of the negative log likelihood

$$-\ \log L(\theta_j) = -\sum_i^n \left[U_{ji}W_i(\theta_j) - \log(1 + \exp(W_i(\theta_j)))\right].$$

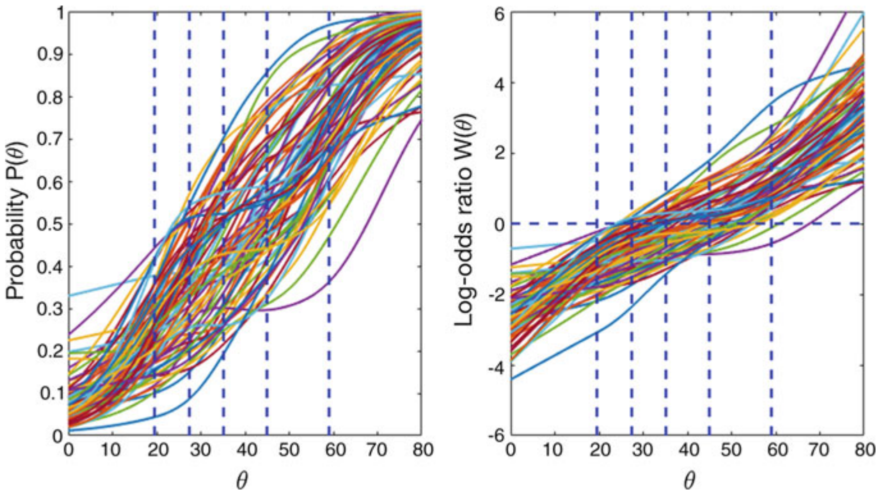


Fig. 2 The left panel displays the $P_i(\theta)$ curves for each item i estimated over the closed interval [0, 80] and the right panel displays the estimated log-odds functions W_i for the SweSAT. The vertical dashed lines are the 5, 25, 50, 75 and 95% quintiles of the empirical distribution of the sum scores

with respect to θ, and the right hand side is zero for its optimal value. Equation 4 is interesting in several aspects. The slopes of the log-odds functions $W_i(\theta)$ at the optimal θ weight the residuals $U_{ij} - P_i(\theta)$. The optimal scores thus correspond to the ability that minimizes the difference between the answers and their probabilities in which each item is weighted by its impact (or sensitivity) value. In practice, this means that high-slope items are mainly influencing the differences in scores among the test takers. The most useful items for assessing test takers at level θ have higher slopes of W_i at that location, while items having nearly flat W_i are down-weighted, which would be the case for easy items being given to high-level θ test takers. We will refer to the interaction between item weights and item performance in the weighting as the *item impact function*. The item impact curves $(dW_i/d\theta)$, corresponding to the curves in Fig. 2, are shown in Fig. 3. From Fig. 3, it is obvious that items have various weights or performances for a certain ability level θ, and one particular item's performance will change at different θ. Summing up, the optimal scoring algorithm is focused on the items that are most informative as reflected by the size of the item impact function $dW_i/d\theta$, which yields the amount of information provided by answers to item i.

$$dW_i/d\theta$$

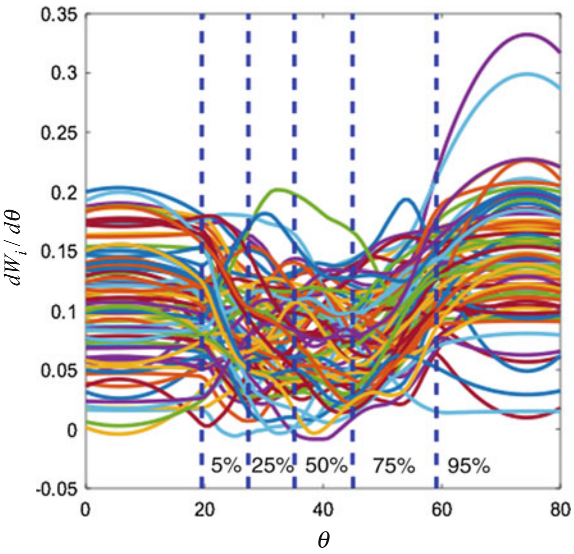


Fig. 3 The item impact curves, $dW_i/d\theta$, that provide the optimal weighting of item scores. The vertical dashed lines are the 5, 25, 50, 75 and 95% quintiles of the empirical distribution of the sum scores

4 Estimating Nonparametric ICC's

An efficient nonparametric procedure for joint estimation of the n functions W_i and the knowledge states θ_j was described in Ramsay and Wiberg (2017a). In their procedure they use parameter cascading (PC), which is a generalization of profiling that is computationally faster than marginalization over θ. Let θ_j be represented by smooth functions $\theta_j(W_1, \ldots, W_n)$. The PC optimizations performed are initialized by a fast data smoothing approach to estimate the W_i as described in Ramsay (1991). PC is a compound optimization procedure in which an inner optimization $(H(\theta|\gamma))$ of a penalized log likelihood function with respect to the θ_j is updated, each time an *outer* optimization $(F(\gamma))$ adjusts the coefficients of the B-spline basis function expansions of the W_i. In PC, the gradient plays a crucial role in the outer optimization through the implicit theorem such that an efficient search is made possible. Details of how to perform PC are provided in Ramsay and Wiberg (2017a). We emphasize that PC is different from using alternating optimization (AO) as for example the EM-algorithm. Instead of a compound optimization as in PC, AO switches between optimizing one criterion F with respect to some γ keeping θ fixed, and optimizing another criterion H with respect to θ keeping γ fixed.

5 Optimal Scores in Comparison with Sum Scores and Expected Sum Scores

5.1 Simulation Study

As a first step, the difference between optimal scores and sum scores as well as optimal scores and expected sum scores were calculated for the SweSAT data. In order to further examine the difference between sum scores, expected sum scores and optimal scores we used simulations from the populations defined by the W_i curves and the θ_j's estimated from the data. The first obstacle was how to handle the problem of identifying the distribution of θ. To make a fair comparison with the sum scores we simulated test data using a smooth estimate of the density of the sum scores based on the SweSAT empirical distribution shown in Fig. 1. As we had access to a sample of 30,000 test takers the W_i have been pre-calibrated and were considered to be known (and can be seen in Fig. 2) and thus we only simulated the test takers' responses. Root mean squared error (RMSE) of θ was used to assess recovery. The analysis was performed using PC for optimization. The 81 sum score values were used as fixed values of θ and we simulated 1000 test takers responses. Sum scores, expected sum scores and optimal scores were averaged across 1000 simulated samples for each value of θ. The average bias of θ for each test score was also used to evaluate the different test scores.

5.2 Results of the Simulation Study

The difference between optimal scores and sum scores as well as optimal scores and expected sum scores are displayed in Fig. 4 for the SweSAT data. The left panel in Fig. 4 shows a large increase in test scores for high-performing test takers if they would get an optimal score instead of a sum score. The expected sum score in the right panel is overall more similar to the sum scores than optimal scores, but the really top achievers among the test takers get penalized with an expected sum score. The sum score/optimal score and sum score/expected score differences can be as large as the size of 20% for some of the scores (for sum scores around 40, the difference can be ±8).

In Fig. 5 the empirical distributions are displayed in the left panel and the average RMSE and bias are shown in the right panel for each value of θ. The empirical distributions for the three different scores only differ slightly. For the mid 90% of the test takers the bias is close to zero regardless of the test scoring method. But low-performing test takers get higher sum scores than the corresponding θ values used to generate the data, while at the same time, high-performing test takers lose about five items using sum scores. For the 5% top- and bottom-performing test takers the bias and RMSE for sum scoring is substantial. For the mid 90% of the test takers the RMSE is larger for the sum scores than for optimal or expected sum

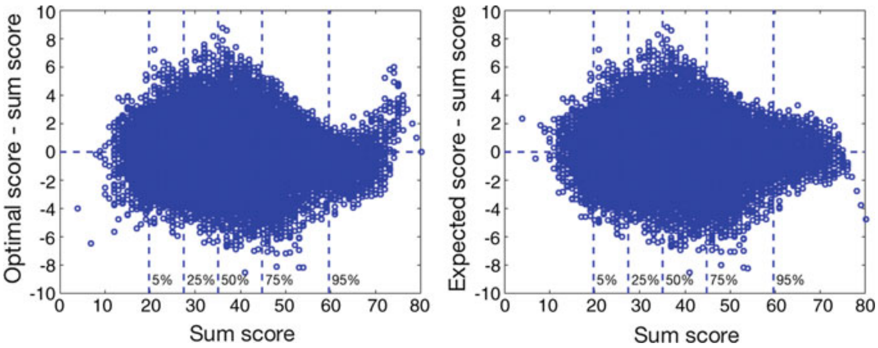


Fig. 4 The left panel displays optimal scores minus the sum scores plotted against sum scores and the right panel displays the expected sum scores minus the sum scores for the SweSAT

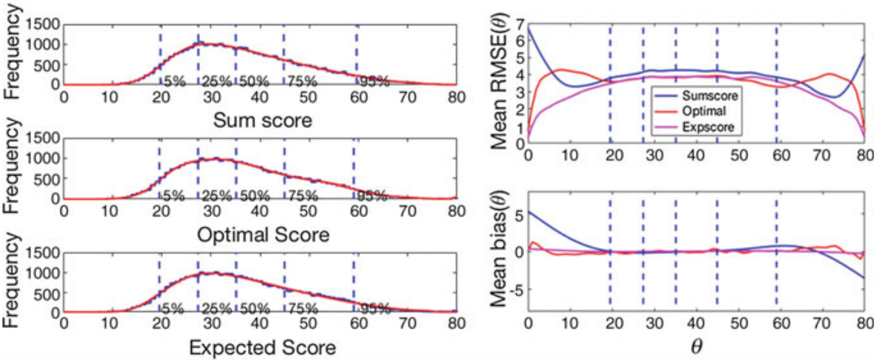


Fig. 5 The left panel displays the empirical distribution of sum scores, optimal scores and expected sum scores and the right panel displays the average RMSE of θ and average bias of θ for the three test scores. The vertical dashed lines are the quintiles of the empirical distribution of the maximum likelihood estimates

scores. From the simulations, the optimal score RMSE was on average 6.8% lower than the sum score RMSE, which corresponds to a mean squared error (MSE) of 14%. Because the MSE declines in proportion to $1/n$, we see that the sum-scored SweSAT would have to be 11 items longer than an optimally scored test in order to achieve the same average accuracy. Note that the expected sum scores have the lowest RMSE and bias at each score values, but they are expected scores and are thus not built on the observed scores as sum scores and optimal scores are. The results from the simulations for optimal scores in comparison to sum scores are in line with the results in Ramsay and Wiberg (2017a), who used simulations based on three different tests and compared optimal scores and sum scores.

6 Discussion

This paper used a large sample from a college admissions test in order to discuss optimal scores in comparison to sum scores and expected sum scores. A closed interval in terms of the range of the sum scores was used in order to model student performance differences. This choice facilitates comparisons with the sum scores and the expected sum scores in terms of bias and RMSE, and also allows for understandable interpretations for the test takers.

The simulation study indicated that the expected sum scores and optimal scores should be preferred over sum scores as their average bias and average RMSE were lower than for the corresponding sum scores. The improvement in terms of RMSE was about 6% for 90% of the test takers. Even though the expected sum scores had the lowest bias and RMSE we cannot recommend it in general as it measures something else than the optimal scores, i.e. it is an expected score instead of an observed score. It was mainly included here for sake of completeness and as expected sum scores are sometimes used in test analysis. The largest problem with sum scores is the substantial negative bias for high-performing test takers and the positive bias for low-performing test takers. The substantial improvement is important, especially in high-stakes test as the SweSAT. To get an improvement of 6% could be the difference of being accepted into the university program of one's choice or not. The improvement found in the well-designed SweSAT lead us to expect a larger benefit if we have less well-designed tests, as for example those given in classrooms. We are not stating that sum scores should never be used as they might be useful in some situations. However if we put some effort into explaining how optimal scores work it may be beneficial for both test constructors and test takers as they contain more information.

In the future it is important to continue examining the performance of optimal scoring, especially against parametric IRT as that is used all over the world by test constructors. As additional information about test takers in terms of covariates are regularly gathered when large-scale tests are given it should be interesting to examine optimal scoring with covariates as it has been used successfully in other test areas as for example test equating (Bränberg and Wiberg 2011; Wallin and Wiberg 2017; Wiberg and Bränberg 2015). Other interesting future directions include the use of optimal scoring with polytomous scored items and multidimensional tests. In order to spread the usage of optimal scoring it is crucial to develop an easy to use software. Currently the authors are developing a new version of TestGraf (Ramsay 2000) which will incorporate all the important features of optimal scoring. In summary, optimal scoring provides a number of interesting opportunities as it is built on efficient and advanced statistical methodology and technology. We need to stop the waste of valuable information and give our top-performing test takers the score they earn.

Acknowledgements This research was funded by the Swedish Research Council grant 2014-578.

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 tests* (pp. 395–479). Reading, MA: Addison-Wesley.

Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48,* 419–440.

Lee, Y.-S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 37,* 121–134.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York, NY: Springer.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Ramsay, J. O. (1997). A functional approach to modeling test data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York, NY: Springer.

Ramsay, J. O. (2000). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*. [Computer software and manual]. Department of Psychology, McGill University, Montreal Canada.

Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis in R and Matlab*. New York, NY: Springer.

Ramsay, J. O., & Silverman, B. W. (2002). Functional models for test items. In *Applied functional data analysis*. New York, NY: Springer.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer.

Ramsay, J. O., & Wiberg, M. (2017a). A strategy for replacing sum scores. *Journal of Educational and Behavioral Statistics, 42,* 282–307.

Ramsay, J. O., & Wiberg, M. (2017b). Breaking through the sum score barrier. In *Paper presented at the International Meeting of the Psychometric Society, July 11–15, Asheville, NC* (pp. 151–158).

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of the Behavioral and Educational Sciences, 27,* 291–317.

Wallin, G., & Wiberg, M. (2017). Non-equivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang. (Eds.), *Quantitative psychology—81st annual meeting of the psychometric society, Asheville, North Carolina, 2016* (pp. 309–320). New York: Springer.

Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39,* 349–361.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods, 11,* 253–270.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71,* 281–301.

Disentangling Treatment and Placebo Effects in Randomized Experiments Using Principal Stratification—An Introduction

Reagan Mozer, Rob Kessels and Donald B. Rubin

Abstract Although randomized controlled trials (RCTs) are generally considered the gold standard for estimating causal effects, for example of pharmaceutical treatments, the valid analysis of RCTs is more complicated with human units than with plants and other such objects. One potential complication that arises with human subjects is the possible existence of placebo effects in RCTs with placebo controls, where a treatment, suppose a new drug, is compared to a placebo, and for approval, the treatment must demonstrate better outcomes than the placebo. In such trials, the causal estimand of interest is the medical effect of the drug compared to placebo. But in practice, when a drug is prescribed by a doctor and the patient is aware of the prescription received, the patient can be expected to receive both a placebo effect and the active effect of the drug. An important issue for practice concerns how to disentangle the medical effect of the drug from the placebo effect of being treated using data arising in a placebo-controlled RCT. Our proposal uses principal stratification as the key statistical tool. The method is applied to initial data from an actual experiment to illustrate important ideas.

Keywords Causal inference · Placebo effects · Principal stratification

1 Introduction

Placebo-controlled, blinded randomized controlled trials (RCTs) are the standard for approving pharmaceuticals to be given to human beings in the United States, European Union, and much of the world. In fact, agencies such as the U.S. Food and

R. Mozer (✉) · D. B. Rubin
Harvard University Department of Statistics, Cambridge, MA 02138, USA
e-mail: rrose@g.harvard.edu

D. B. Rubin
e-mail: rubin@stat.harvard.edu

R. Kessels
Emotional Brain, B.V., Almere, The Netherlands
e-mail: r.kessels@emotionalbrain.nl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_2

Drug Administration (FDA) and the European Medicines Agency (EMA) usually require evidence from such trials that the drugs being proposed are safe and effective. It is a widely accepted stance in the world of drug development that if a drug is “snake oil”, meaning it is ineffective and only appears to work because of presumed expectancy effects, then the producer of the drug should not profit from its sale. It was because of this attitude that placebo controlled, double-blind randomized trials became essentially necessary for the approval of new drugs in the 1960s. That is, for a drug to be considered effective, the active drug (treatment) must be compared to an inactive drug (a placebo), which (to a user) is indistinguishable from the active drug, where assignment to the treatment versus control is random and unknown to the experimental units until the completion of the experiment; here the units are said to be “blinded” to the actual assignment. If assignment is unknown to both the experimental units and the experimenter, the experiment is considered “double-blind”.

Although randomized experiments have been used for nearly a century, for decades they were only used with unconscious units, such as plants, animals, or industrial objects, none of which presumably could be influenced by the knowledge that they were objects of experimentation. Historically, it has been recognized that humans are different and can be influenced by the knowledge that they are part of an active experiment. In some cases, that knowledge alone has been shown to influence participants behavior, as with the well-known “Hawthorne effect” (Landsberger 1958), where awareness of participation in a study influences outcomes. In other examples, the knowledge that some individuals would receive an active drug with a particular anticipated effect creates the expectation among all experimental units that this anticipated effect will be achieved among all participants, a version of so-called “expectancy effects (Rosenthal and Fode 1963; Rosenthal and Jacobson 1966). Thus, a number of complications may arise when analyzing data from randomized experiments with human subjects when the conduct of the experiment itself influences participants’ outcomes.

2 Motivation

Emotional Brain (EB) is a research company based in the Netherlands that is developing a therapy for improving sexual functioning in women, which they call Lybrido. Lybrido is designated for the treatment of a medical condition in women called Female Sexual Interest/Arousal Disorder (FSIAD). The increase in “satisfying sexual events” (SSEs) per week from baseline (before any drugs, active or placebo, have been received) is the accepted primary outcome of interest, and for approval of Lybrido, by either the FDA or EMA, there must be evidence that the drug is superior to placebo with respect to increase in SSEs from baseline, ΔSSE. As with other psychopathologies, experiments on therapies for treating this condition are believed to suffer from large placebo effects because the anticipation of effects of the drug can have obvious effects on the self-reported number of SSEs.

A variety of small, but expensive, randomized placebo-controlled double-blind trials have been conducted to study the effectiveness of Lybrido (Van Der Made et al.

2009a, b; Poels et al. 2013). In these trials, simple analyses comparing the randomized groups with each other (intention to treat analyses) generally show significant positive effects for Lybrido relative to placebo, but the large placebo effects (that is, large increases in SSEs observed in all groups) complicate the interpretation and implication of the results.

The desire to disentangle the active effects of Lybrido from its related placebo effects is important for several reasons. First, assume Lybrido has a true effect for some subset of women, but this true effect is masked by highly variable placebo effects; how do we eliminate the noise and so identify that subset of women? This is related to the current hot-topic issue of "personalized medicine", which describes selecting treatments that are tuned to patients characteristics. Another important question concerns what outcomes should be anticipated in actual medical practice, when doctors prescribe a treatment and patients are aware of the prescription they receive. In this setting, patients' outcomes will reflect both placebo effects as well as the medical effects of the active drug. Considering both types of effects may allow prescribing physicians to anticipate better the benefits a patient can expect when using the drug outside of the setting of an RCT.

The objective of this work is to disentangle active drug and placebo effects in RCTs, such as those with Lybrido. Previous attempts to address this issue using existing methods are summarized in Kessels et al. (2017), and, though some have interesting ideas, none are statistically fully satisfactory. Here we use the statistical tool called Principal Stratification (Frangakis and Rubin 2002) to estimate jointly treatment and placebo effects within the framework of causal inference based on potential outcomes, commonly called the Rubin Causal Model (Holland 1986) for a body of work done in the 1970s (Rubin 1974, 1975, 1978, 1980); a short summary of this perspective is in Imbens and Rubin (2008) and a book on it is Imbens and Rubin (2015).

In principle, we consider the administration of placebo as an intervention, just as the administration of an active drug. The placebo effect is then defined by comparing potential outcomes under assignment to placebo to potential outcomes under no treatment at all. Just as active treatment effects can vary across units, so can placebo effects, which can also vary as a function of patients' individual characteristics. Further, the effects of the active treatment can also vary with respect to characteristics of patients, including their individual placebo effects, which further complicates statistical inference.

3 The Principal Stratification Framework for Joint Estimation of Treatment and Placebo Effects

3.1 Notation

Consider an RCT with N subjects, indexed by $i = 1, \ldots, N$. Subject i is assigned treatment Z_i, which equals 1 for subjects assigned and receiving active treatment and equals 0 for subjects assigned and receiving placebo. Throughout, we assume full compliance with assignment. Interest focuses on the effect of treatment ($Z_i = 1$) compared to placebo ($Z_i = 0$) on an outcome variable, defined in terms of change from a baseline measurement Y_{i0}. For each subject we may also observe a vector of p pre-treatment covariates $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ where X is the $N \times p$ matrix of covariates for all subjects. The outcome variable takes the value $Y_i(1)$ if subject i is assigned treatment and $Y_i(0)$ if subject i is assigned placebo. The "fundamental problem facing causal inference" (Rubin 1978) is that we cannot observe both potential outcomes $Y_i(0)$ and $Y_i(1)$, but rather $Y_i^{obs} = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$, the observed outcome for subject i. Additionally, we consider a third potential outcome, $Y_i(-1)$, which is defined but never observed for any unit in the situation we consider and represents the outcome that would be observed if unit i is neither assigned nor receives either treatment or placebo and is aware of this. We then define the causal effects of interest by differences in potential outcomes, where $Y_i(0) - Y_i(-1)$ is the "placebo effect" for unit i and $Y_i(1) - Y_i(0)$ is the "medical effect" of active treatment for unit i, or for descriptive simplicity, the treatment effect.

3.2 General Modeling Strategy

Because we believe that effect of the active treatment can depend on both individual characteristics of the patient (i.e., covariate values X_i) and the magnitude of the patient's response to placebo, $Y_i(0)$, our approach is a version of the one used in Jin and Rubin (2008), which deals with "extended partial compliance", a special case of principal stratification that defines principal strata based on continuous measures of how each patient would comply with their assignment under both treatment and control.

Here, we view patients' response to placebo as roughly analagous to compliance status under active treatment, and following Jin and Rubin (2008), we define continuous principal strata according to this potential outcome, which is only partially revealed (i.e., revealed for those patients assigned placebo), but is missing for those patients assigned the active treatment. Causal effects of the active treatment versus placebo are then defined conditional on the observed covariates and the potential outcomes under placebo. Regression models (typically not linear) are used for the joint conditional distribution of potential outcomes given covariates, specified by the distribution of the placebo potential outcome (given covariates) and the conditional distribution of the potential outcome under treatment given the potential outcome under placebo (and, of course, the covariates). This is explained in greater detail in Sect. 4. For analysis, we use Bayesian models with proper prior distributions and employ Markov Chain Monte Carlo (MCMC) methods, which are only outlined in this paper. Under this framework, missing potential outcomes are multiply imputed to obtain a large number of completed data sets, from each of which, all causal estimands, including individual causal effects, can be computed. Aggregates of the estimated individual effects across the multiply imputed data sets then approximate the posterior distributions of interest.

3.3 Assumptions

Throughout this article, we assume the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980), which requires that there is no interference between units (that is, treatment assignment for an individual unit has no effect on the potential outcomes of other units) and that there are no hidden versions of treatments. We also assume ignorable treatment assignment (Rubin 1978), which requires that the treatment assignment is known to be a probabilistic function of observed values and is true by design in randomized experiments. Next, we assume that the potential outcomes under no treatment, $Y_i(-1)$, defined as the change in the outcome from its measurement at baseline, is zero for all units (i.e., $Y_i(-1) = 0$ for all $i = 1, \ldots, N$). This important assumption implies that the outcome that would be observed for each unit if they were given neither the active treatment nor placebo, and are aware that they are receiving neither, will be exactly equal to the value of that unit's outcome at baseline; assessing this assumption would require a design with such an assignment (i.e., an assignment with instructions to take nothing and continue to be followed up with measurements as if the patient had been assigned either active treatment or placebo).

All other assumptions are extensions of the classical assumptions utilized in problems involving principal stratification. In particular, we assume positive side-effect monotonicity on the primary outcome for both treatment and placebo, that is, $Y_i(1) \geq 0$ and $Y_i(0) \geq 0$ for all i, which implies that neither the treatment nor the placebo are harmful to any units, in the sense that an individual will not experience a decline in their outcome (measured as change from baseline) as a result of either intervention.

We also assume additivity of the treatment and placebo effects on some scale. This is analagous to the perfect blind assumption commonly made in causal inference, which requires that, upon receipt, the active drug is indistinguishable from the placebo except for its active effect. Under this assumption, for a unit assigned to treatment, the portion (on some scale) of the observed outcome that is attributable to the placebo effect is exactly equal to the placebo effect that would be observed if that unit had been assigned placebo. Thus, the potential outcome when assigned treatment can be viewed as the sum of the "placebo effect" and some "extra" effect achieved under treatment that is attributable to the active drug, which we call the treatment effect.

Together, these assumptions also imply that for every patient, the total response that would be observed when assigned treatment is greater than or equal to the response that would be observed when assigned placebo (i.e., $Y_i(1) \geq Y_i(0)$ for all i).

4 Model and Computation

4.1 The General Model with No Covariates

We begin by considering the simplest case of an RCT with no covariates. We first specify a distribution for the potential outcomes under control, $Y(0)$, conditional on some global parameter θ:

$$Y_i(0)|\theta \sim \mathcal{L}^{(0)}, \;\; Y_i(0) \geq 0 \;\text{ for all }\; i, \theta \tag{1}$$

where $\mathcal{L}^{(0)}$ denotes the probability law for $Y_i(0)$, governed by some parameters, which are functions of the global parameter θ. We then specify a distribution for the potential outcomes under treatment, $Y_i(1)$, conditional on the potential outcomes under control, $Y_i(0)$ and θ as

$$Y_i(1)|Y_i(0), \theta \sim \mathcal{L}^{(1)}, \;\; Y_i(1) \geq 0 \;\text{ for all }\; i, \theta \tag{2}$$

where

$$E[Y_i(1)|Y_i(0), \theta] = Y_i(0) + f(Y_i(0)). \tag{3}$$

Here, $\mathcal{L}^{(1)}$ is another probability law, and f is an arbitrary function that generally defines heterogeneous treatment effects across units as a function of potential outcomes under placebo. Under this formulation, $f(Y_i(0))$ is the treatment effect for unit i. By the assumptions stated in Sect. 3.3, $f(\cdot)$ must be chosen such that $f(Y_i(0)) \geq 0$ for all i and $Y_i(0) + f(Y_i(0))$ is monotonically non-decreasing in $Y_i(0)$, which defines a positive, monotonically non-decreasing curve, analagous to a dose-response curve, which captures the expected effect of assignment to treatment versus assignment to placebo for each possible value of placebo response.

For example, consider the specification of $f(\cdot)$ as the polynomial $f(x) = a_0 + a_1x + a_2x^2$, where a_0, a_1 and a_2 are constrained such that $f(x) \geq 0$, and $1 + f'(x) = 1 + a_1 + 2a_2x \geq 0$ for all x (thereby satisfying the monotonicity constraint). Under this specification, the parameters of interest are (a_0, a_1, a_2), where the intercept parameter a_0 is a common treatment effect across all subjects, including those who have zero response to placebo, and the parameters a_1 and a_2 capture how treatment effects vary linearly and quadratically, respectively, with the magnitude of placebo response. Figure 1 illustrates such a specification, where the left plot displays the expected medical effect of the active drug as a function of placebo response, which is relevant for drug approval, and the right plot displays the overall expected response to being assigned the active drug and taking it as a function of placebo response, which is relevant for anticipating the benefits a patient can expect when using the drug as prescribed by a doctor.

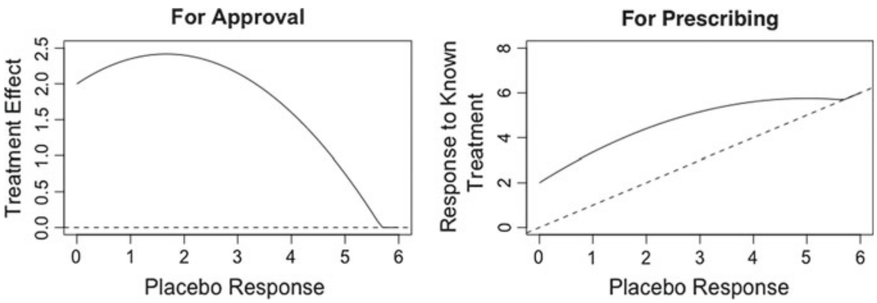


Fig. 1 Two illustrations of a possible quadratic specification of *f*

4.2 Computation

Under the general formulation above, the complete-data likelihood for the data $Y = (Y(0), Y(1))$ (meaning the likelihood if both $Y_i(1)$ and $Y_i(0)$ were observed for all units) is:

$$p(Y|\theta, Z) = \prod_i p(Y_i(1), Y_i(0)|\theta) = \prod_i p(Y_i(1)|Y_i(0), \theta)p(Y_i(0)|\theta). \tag{4}$$

For Bayesian inference, with prior distribution $p(\theta)$ on θ, the posterior distribution of θ given the complete data Y is then:

$$p(\theta|Y, Z) \propto p(\theta)p(Y, Z|\theta) = p(\theta)p(Y|\theta, Z), \tag{5}$$

where the equality follows from the randomization of Z. Posterior inference on θ can then be done using straightforward application of MCMC techniques, such as the Gibbs sampler (Geman and Geman 1984; Gelman et al. 2014). For example, in each iteration of the Gibbs sampler, we draw the missing potential outcomes Y^{mis} given the observed data Y^{obs} and the current draw of the parameter θ:

$$\begin{aligned} p(Y^{mis}|Y^{obs}, \theta, Z) &= \prod_{i\in\{Z_i=0\}} p(Y_i(1)|Y_i(0) = Y_i^{obs}, \theta) \\ &\quad \times \prod_{i\in\{Z_i=1\}} p(Y_i(0)|Y_i(1) = Y_i^{obs}, \theta) \\ &= \prod_{i\in\{Z_i=0\}} p(Y_i(1)|Y_i(0) = Y_i^{obs}, \theta) \\ &\quad \times \prod_{i\in\{Z_i=1\}} p(Y_i(1)|Y_i(0) = Y_i^{obs}, \theta)p(Y_i(0) = Y_i^{obs}|\theta) \end{aligned} \tag{6}$$

where the second equality follows from Bayes Rule. We then draw θ given the completed data $Y = (Y^{obs}, Y^{mis})$ using Eqs. 4 and 5, and we continue this process until convergence in distribution. Depending on the specifications of $\mathcal{L}^{(0)}$ and $\mathcal{L}^{(1)}$, the conditional distribution of Y^{mis} given Y^{obs} and θ, and the conditional distribution of θ given the complete data Y, may not have closed-form solutions that allow us to sample directly values of Y^{mis} or θ. In such situations, Metropolis-Hastings steps can be used to draw approximate samples from the desired conditional distributions in each iteration of the Gibbs sampler. For posterior inference on causal effects of interest, we continue this sampling procedure after approximate convergence, in each iteration drawing the missing potential outcome, $Y_i(0)$ or $Y_i(1)$, for each patient. Thus, in each iteration, we construct a completed dataset consisting of all observed potential outcomes and the imputed missing potential outcomes, and then use this completed data to calculate the implied placebo and treatment effects. Repeating this process over many such simulated datasets produces the approximate posterior distribution for all causal effects of interest. In the same way, posterior samples of θ can provide posterior estimates of the parameters of the function f, which characterizes the relationship between expected response to treatment and expected response to placebo.

Depending on the exact specification of $f(\cdot)$, the likelihood may suffer from problems with multimodality, as is common with many specifications of mixture models, such as this one. In such situations, initialization of the MCMC procedure can have an impact on convergence, and first finding regions of high posterior density (e.g., maximum likelihood estimates—MLEs) for model parameters using a method such as a variant of Expectation Maximization (EM) (Dempster et al. 1977) to inform initial values in the MCMC procedure can help. In cases of extreme multi-modality of the likelihood, one can also specify more restrictive prior distributions on the parameters governing $f(\cdot)$.

4.3 *Incorporating Covariates*

The model presented in Sect. 4.1 considers a patient's response to placebo as an underlying, psychological, characteristic that exists prior to treatment assignment. By defining heterogeneous treatment effects as a function of this characteristic, we can estimate both the expected effect of assignment to treatment versus assignment to placebo (the medical effect of the active drug) and the expected effect of assignment to placebo versus assignment to neither treatment nor placebo (the placebo effect) for each type of patient, at least under specific assumptions.

When covariates, $X_i = (X_{i1}, \dots, X_{ip})$, are observed for patients, we can specify the distribution for potential outcomes under control, $Y_i(0)$, conditional on X_i and the global parameter θ as:

$$Y_i(0)|X_i, \theta \sim \mathcal{L}^{(0)} \quad Y_i(0) \geq 0 \text{ for all } \ i, \theta. \tag{7}$$

We then model the potential outcomes under treatment, $Y_i(1)$, conditional on $Y_i(0), X_i$, and θ as

$$Y_i(1)|Y_i(0), X_i, \theta \sim \mathcal{L}^{(1)} \;\; Y_i(1) \geq 0 \text{ for all } \; i, \theta, \tag{8}$$

where $\mathcal{L}^{(1)}$ is such that

$$E[Y_i(1)|Y_i(0), X_i, \theta] = E[Y_i(0) + f(Y_i(0))|X_i]. \tag{9}$$

In general, we assume that covariate effects on $Y_i(0)$ are conditionally independent of effects on $Y_i(1)$. For example, continuing the example where f is specified using the polynomial $f(x) = a_0 + a_1x + a_2x^2$, we might consider linear regression models for covariate effects on both $Y_i(0)$ and $Y_i(1)$:

$$\begin{aligned} Y_i(0)|X_i, \theta &= \beta_0 + X_i\beta + \epsilon_i \\ Y_i(1)|Y_i(0), X_i, \theta &= Y_i(0) + X_i\gamma + a_1Y_i(0) + a_2Y_i(0)^2 + \eta_i, \end{aligned} \tag{10}$$

where ϵ_i and η_i are independent residual terms and $\beta, \gamma \in \mathcal{R}^p$ govern covariate effects. Here, we may include an intercept term for the distribution of potential outcomes under control but not for the distribution of potential outcomes under treatment. In this example, posterior inference for θ comprises two standard Bayesian regressions (Gelman et al. 2014).

5 Evaluating Treatment and Placebo Effects of Lybrido on Sexual Function

To illustrate our proposed approach, we return to our motivating example of Lybrido. Data for this example were pooled from two double-blind, placebo-controlled RCTs conducted by EB to investigate the efficacy of Lybrido among patients for whom FSIAD was believed to be caused by insensitivities in the brain to sexual cues. Because the actual results of both studies are under peer review process with an implied embargo, a subset of 67 patients was sampled from these data to be used for illustrative purposes here, 34 randomized to treatment (Lybrido) and 33 randomized to control (placebo).

The primary outcome of interest in this example is the increase from baseline in number of SSEs within a four week period during the study. In this example, no baseline measurements for SSEs are directly observed for any participants in the sample, but implicitly these values are all equal to zero, because the patients in these experiments have FSIAD and therefore suffer from low sexual desire. This likely leads to infrequent SSEs among these patients, which makes the assumed value of zero SSEs at baseline realistic. In addition to the outcome, we observe the age and body mass index (BMI) for each patient at the time of enrollment, as well as 40 other covariates collected via self-report using the Sexual Motivation Questionnaire (SMQ), as described in detail in a subsequent publication.

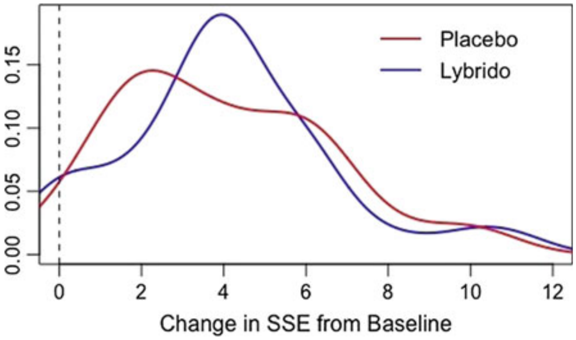


Fig. 2 Kernel density estimates for the distributions of observed potential outcomes in treatment (Lybrido) and control (placebo)

We observe an average of 4.00 SSEs over the four week study period for patients randomized to receive treatment, with a standard deviation of 2.58, and an average of 4.06 SSEs over the four week period for patients randomized to receive placebo, with a standard deviation of 2.58. Kernel density estimates of the distributions of observed potential outcomes in the treatment and control groups are shown in Fig. 2.

Using simple intention to treat (ITT) analysis (Sheiner and Rubin 1995), which compares the means of observed potential outcomes among treated units to those in control, we estimate the ITT effect of assignment to Lybrido to be $4.00 - 4.06 = -0.06$. At first glance, this result suggests that Lybrido has essentially zero effect compared to placebo and might lead to the conclusion that the drug is ineffective as a treatment for FSIAD. However, because both the placebo and treatment groups are observed to have large and highly variable responses (with standard deviations of approximately 2.58 in each group), this finding may instead suggest that any effect of the active drug is simply being masked by large placebo effects and varying treatment effects, which more sophisticated statistical analyses might be able to detect.

5.1 Model Specification

To illustrate our proposed approach on these data, we consider models both with and without the observed covariates. For both models, we specify the function $f(\cdot)$, which relates each patients' treatment effect to their expected potential outcomes under assignment to placebo, using the simple quadratic form $f(x) = a_0 + a_1 x + a_2 x^2$. In the model for $Y_i(1)$ that includes covariates, however, no intercept term is included because $Y_i(1)$ is already centered at $Y_i(0)$. For both models, we assume a truncated normal distribution for placebo response, $Y_i(0)$. With no covariates, this is:

$$Y_i(0)|\theta \sim \mathcal{N}_+(\mu_0, \sigma_0^2), \tag{11}$$

where $\mathcal{N}_+(\mu, \sigma^2)$ denotes a normal density with mean μ and variance σ^2 truncated to the interval $[0, \infty]$. Similarly, we specify a truncated normal distribution for treatment response, $Y_i(1)$, given $Y_i(0)$ as:

$$Y_i(1)|Y_i(0), \theta \sim \mathcal{N}_+(Y_i(0) + f(Y_i(0)), \sigma_1^2). \tag{12}$$

In this illustrative example, we use truncated normal distributions for both placebo response, $Y_i(0)$, and treatment response, $Y_i(1)$, to satisfy the assumption of positive side-effect monotonicity, which requires that both $Y_i(0)$ and $Y_i(1)$ be strictly non-negative for all i. However, other distributions that satisfy this constraint (e.g., Poisson) could be specified for one or both of these variables. In general, we advise researchers implementing this approach in practice to choose appropriate distributions based on domain knowledge about the treatment and population under investigation.

When including covariates, we model $Y_i(0)$ conditional on X_i as:

$$Y_i(0)|X_i, \theta \sim \mathcal{N}_+(\beta_0 + X_i\beta, \sigma_0^2), \tag{13}$$

and model $Y_i(1)$ given $Y_i(0)$ and X_i as:

$$Y_i(1)|Y_i(0), X_i, \theta \sim \mathcal{N}_+(Y_i(0) + f(Y_i(0)) + X_i\gamma, \sigma_1^2). \tag{14}$$

In the model without covariates, the global parameter is $\theta = (a_0, a_1, a_2, \sigma_0^2, \sigma_1^2)$, and with covariates we have $\theta = (a_0, a_1, a_2, \beta_0, \beta, \gamma, \sigma_0^2, \sigma_1^2)$, where β_0 is an intercept term for the regression of response to placebo, $Y_i(0)$, on the covariates X_i, and β and γ are p-dimensional vectors with components for coefficients for the covariate effects on $Y_i(0)$ and $Y_i(1)$. In both models, we use weakly informative prior distributions on all parameters, where each prior distribution is proper and fully specified.

5.2 Results

Results from the models with and without covariates are displayed in Fig. 3. When using the model without covariates, we estimate the function f as $\hat{f}(Y_i(0)) = 0.288 - 0.035Y_i(0) - 0.481Y_i(0)^2$, which suggests that Lybrido has the largest effects on patients that do not respond to placebo ($E[Y_i(1)|Y_i(0) = 0] \approx 0.288$). Further, we see that estimated treatment effects decrease with response to placebo, such that patients who have a placebo response of approximately one or more post-assignment SSEs are expected to have essentially zero treatment effects. That is, big placebo responders do not benefit from receiving the active treatment. The findings are similar when employing the model that incorporates covariates. Using this model, we obtain $\hat{f}(Y_i(0)) = 0.321 + 0.016Y_i(0) - 0.323Y_i(0)^2$ with $E[Y_i(1)|Y_i(0) = 0] = 0.321$. Among covariates considered, none were identified as significant predictors of

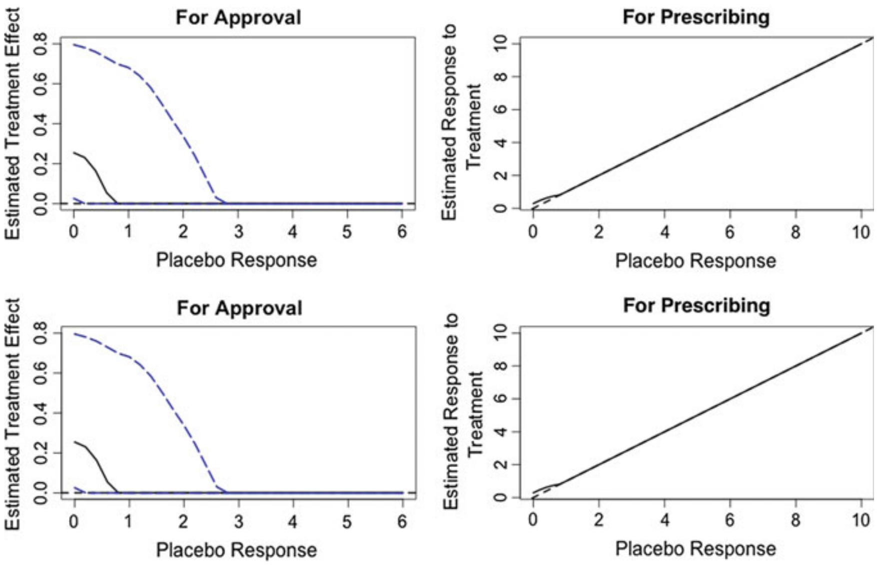


Fig. 3 Estimated treatment (Lybrido) effects and total response to treatment as a function of placebo response using model with no covariates (top), and essentially the same figures formed using the model with covariates (bottom). Dashed blue lines show 95% posterior intervals

response to placebo or active treatment, though this is possibly due to the small sample size. Contrary to the ITT estimate, overall these results provide some evidence that Lybrido may have a small, but positive, effect for a subset of patients who have a minimal response to placebo—an interesting possibility.

6 Discussion

The approach developed in this paper is intended to allow for the estimation of distinct treatment and placebo effects in randomized experiments with human subjects. We believe that this approach establishes a foundation for precise estimation of each of the effects of interest, and has practical implications for both regulatory agencies making approval decisions for new drugs and clinicians prescribing drugs to patients. However, we recognize that this foundation is only a start to elucidate placebo effects and their influence on the effects of active treatments. Although results for the applied data analysis are promising, the framework presented here relies on a number of assumptions. Further research may attempt to relax some of these assumptions, for example, by considering new experimental designs that are intended for the investigation of treatments with large expected placebo effects.

References

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1–38.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL: Chapman & Hall/CRC.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Imbens, G. W., & Rubin, D. B. (2008). *The new Palgrave dictionary of economics* (pp. 255–262). Palgrave Macmillan.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jin, H., & Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, *103*(481), 101–111.

Kessels, R., Mozer, R., & Bloemers, J. (2017). Methods for assessing and controlling placebo effects. *Statistical Methods in Medical Research*.

Landsberger, H. A. (1958). *Hawthorne revisited: Management and the worker, its critics, and developments in human relations in industry.* ERIC.

Poels, S., Bloemers, J., van Rooij, K., Goldstein, I., Gerritsen, J., van Ham, D., et al. (2013). Toward personalized sexual medicine (Part 2): Testosterone combined with a PDE5 inhibitor increases sexual satisfaction in women with HSDD and FSAD, and a low sensitive system for sexual cues. *The Journal of Sexual Medicine*, *10*(3), 810–823.

Rosenthal, R., & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Systems Research and Behavioral Science*, *8*(3), 183–189.

Rosenthal, R., & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports*, *19*(1), 115–118.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *The Proceedings of the Social Statistics Section of the American Statistical Association* (pp. 233–239).

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 34–58.

Rubin, D. B. (1980). Comment. *Journal of the American Statistical Association*, *75*(371), 591–593.

Sheiner, L. B., & Rubin, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology & Therapeutics*, *57*(1), 6–15.

Van Der Made, F., Bloemers, J., Van Ham, D., Yassem, W. E., Kleiverda, G., Everaerd, W., et al. (2009a). Anatomy/physiology: Childhood sexual abuse, selective attention for sexual cues and the effects of testosterone with or without Vardenafil on physiological sexual arousal in women with sexual dysfunction: A pilot study. *The Journal of Sexual Medicine*, *6*(2), 429–439.

Van Der Made, F., Bloemers, J., Yassem, W. E., Kleiverda, G., Everaerd, W., Van Ham, D., et al. (2009b). The influence of testosterone combined with a PDE5-inhibitor on cognitive, affective, and physiological sexual functioning in women suffering from sexual dysfunction. *The Journal of Sexual Medicine*, *6*(3), 777–790.

Some Measures of the Amount of Adaptation for Computerized Adaptive Tests

Mark D. Reckase, Unhee Ju and Sewon Kim

Abstract Computerized Adaptive Testing (CAT) is gaining wide acceptance with the ready availability of computer technology. The general intent of is to adapt the difficulty of the test to the capabilities of the examinee so that measurement accuracy is improved over fixed tests, and the entire testing process is more efficient. However, many computer administration designs, such as two-stage tests, stratified adaptive tests, and those with content balancing and exposure control, are called adaptive, but the amount of adaptation greatly varies. In this paper, several measures of the amount of adaptation for a CAT are presented along with information about their sensitivity to item pool size, distribution of item difficulty, and exposure control. A real data application is presented to show the level of adaptation of a mature, operational CAT. Some guidelines are provided for how much adaptation should take place to merit the label of an “adaptive test.”

Keywords Computerized adaptive testing · Measures of adaptation
Test design

1 Introduction

A simple definition of adaptive testing (Lord 1980) is: a test where the specific tasks that make up the test are selected for each examinee using appropriate criteria during the process of test administration to optimize a specified desired test characteristic. This simple definition includes a very broad set of tests including the oral examinations given to students before large-scale, paper-and-pencil tests became

M. D. Reckase (✉) · U. Ju · S. Kim
Michigan State University, 620 Farm Lane, East Lansing, MI 48824, USA
e-mail: reckase@msu.edu

U. Ju
e-mail: juunhee@msu.edu

S. Kim
e-mail: kimsewon@msu.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_3

common (and that are still used for evaluating doctoral candidates when they defend their dissertations) and such tests as the intelligence tests developed by Binet and Simon (1915). However, these early versions of adaptive tests did not include all the component parts that are typically included in today's large-scale use of adaptive testing. The use of the computer as a replacement for the individual person as test administrator has allowed adaptive tests to be used with large populations of examinees and has resulted in the acronym, CAT (computerized adaptive test). Now CATs include item selection algorithms, exposure control, content constraints, item pool designs, estimation procedures, etc. Mills et al. (2002) describe these components. They are also discussed in other books on the topic of the general design and development of adaptive tests (e.g., Parshall et al. 2002; van der Linden and Glas 2010).

The variety of additional issues that are considered now when CATs are designed and implemented make it possible that a computer-based test that uses the technical methodologies of adaptive testing might have so many constraints that the test is in practice hardly adaptive at all. It is also possible that item development costs and security issues might lead to using item pools that are too small to support good adaptation. It is concern about these issues that has led to the research reported here on the development of statistical indices of the amount of adaptation observed in an operational CAT.

2 Types of CATs that are the Focus of this Research

There are many variations in the details of the implementations of CATs. Because it is not the purpose of this research to include all the variations, a simple categorization of types will be used to provide an organizational structure for the field: (1) adaptation on proficiency (e.g., Parshall et al. 2002; van der Linden and Glas 2010), (2) adaptation on test length for making decisions (e.g., Spray and Reckase 1996), and (3) adaptation on latent classes (Liu et al. 2015). The first type of CAT has a basic item selection goal that is optimizing the estimation of the location of each examinee on a latent continuum. Of course, as indicated above, there may be other item selection goals in this type of CAT such as minimizing exposure of items or insuring that the proportions of items from specified content areas match a target. In the work reported here, it is assumed that optimization of the accuracy of estimation of location is an important goal of the CAT.

The second type of adaptation is selecting items to obtain classification error less than a specified value when making classification decisions. This selection criterion results in CATs of varying lengths with the length depending on how close examinees are to decision points. The third type of CAT selects test tasks to optimize the classification of examinees into some number of latent classes. These latent classes are typically not along a continuum. This third type of CAT is a relatively new application of adaptive testing.

Of the three types of CAT, the first will be the main focus of the research reported here. However, the operational CAT whose adaptability was analyzed as part of the research reported here has aspects of both the first and second types. Overall, the goal of the research reported here was to develop and evaluate ways to quantitatively describe the amount of adaptation that occurs within a CAT when examinees have a range of different locations on the target latent continuum. The other types of CAT designs are topics for future research. Note, however, that variable length CATs that seek to accurately locate examinees on a continuum are considered here. Fixed test length is not a requirement for the approach taken to evaluate the amount of adaptation.

The research reported here is of two types. The first type is based on CAT simulations with the goal of developing guidelines for what are reasonable amounts of adaptation to expect from a CAT. These simulations vary three components of a CAT: (1) the item pool size, (2) the spread of difficulties in the item pool, and (3) the type of exposure control applied in the CAT (randomesque (Kingsbury and Zara 1989), Sympson-Hetter (Sympson and Hetter 1985), and none). All simulations were done assuming the Rasch model gave a good representation of item/person interactions. Using the Rasch model simplified the interpretation of the results. The second type of research is the analysis of the amount of adaptation of operational test results from a certification/licensure test.

3 A Simple Conceptual Framework

A conceptually simple hypothetical example is presented here to provide a framework for the approach taken to quantify the adaptability of a CAT with the goal of estimating location on a latent continuum. Suppose that a CAT uses the Rasch model as the psychometric model for item selection and estimation of the latent trait. Also, assume that the location of examinee *j* on the latent continuum, θ_j, is known. Then, the optimal set of test items for confirming the known location would all have difficulty parameters (*b*'s) from the Rasch calibration of the items equal to θ_j. Of course, in this hypothetical case, the mean of the *b*-parameters administered to examinee *j* would also be equal to θ_j and the standard deviation of the *b*-parameters would be 0.

Extending this hypothetical example, suppose that there is a sample of examinees with known locations on the θ-scale. If each gets the optimal set of items as described above, then the correlation between the mean of the *b*-parameters for each of the examinees, $\bar{b}_j$, and their locations on the scale, θ_j, would be 1.0. Further, the standard deviation of the $\bar{b}_j$'s would be equal to the standard deviation of θ_j's and the ratio of the two standard deviations would yield a value of 1.0.

Of course, this example is fanciful because we never know the true location of an examinee on the θ-scale, and if we did we would not have to give the CAT. However, the example does set the limits for adaptability for an ideal CAT that is designed to estimate the location of examinees on a latent continuum. For an actual, operational CAT, each examinee will receive a set of items with variation in the *b*-parameters and the mean of the *b*-parameters may not be equal to the true location of the examinee on the θ-scale or even the final estimate of the location on the scale. But, if the test is adaptive, the correlation over examinees, $r(\bar{b}_j, \hat{\theta}_j)$, between the mean difficulty parameters, $\bar{b}_j$, and the final estimate of examinee location, $\hat{\theta}_j$, for examinee *j*, should be high and positive, showing that the examinees at different locations received sets of items that were different in difficulty and that the level of difficulty is appropriately related to the final estimate of location on the θ-scale.

Even if $r(\bar{b}_j, \hat{\theta}_j)$, is close to 1.0, the adaptability of the CAT might not be good because of limitations in the item pool or because of a problem with the item selection algorithm. For example, if the item pool has a limited range of difficulty, but the item selection algorithm is working well, the correlation may be high, but the standard deviation of the $\bar{b}_j$'s may be small compared to the standard deviation of the $\hat{\theta}_j$'s. In that case, the ratio of the standard deviation of the $\bar{b}_j$'s to the standard deviation of the $\hat{\theta}_j$'s would be less than 1.0. The opposite could also be true. There could be insufficient items in the middle range of difficulty and many at the extremes. The standard deviation of the $\bar{b}_j$'s could be large relative to that of the $\hat{\theta}_j$'s resulting in a ratio that is greater than 1.0. Thus, ratios of the standard deviations, $s_{\bar{b}_j}/s_{\hat{\theta}_j}$, that differ from 1.0 in either direction indicate a problem with adaptability.

The adaptation of the test may also be poor if there are insufficient items in the region of the scale that contains the final estimate of the examinee's location, $\hat{\theta}_j$. In such a case, the item selection algorithm may have to select items that have *b*-parameters that are some distance from the current estimate of location. It is possible that the $\bar{b}_j$ might be close to the final estimate, $\hat{\theta}_j$, but the variance of the *b*'s for that examinee might be high. A statistic that would quantify this situation has the same form as a familiar equation for reliability (Hoyt 1941), $\frac{s_b^2 - pooled\ s_{b_j}^2}{s_b^2}$, where s_b^2 is the variance of the *b*-parameters in the item pool, and $s_{b_j}^2$ is the variance of the *b*-parameters administered to examinee *j*. Because the adaptation is over a sample of examinees, the variance of b_j's is pooled over examinees. If the variance of the *b*-parameters for each examinee is 0 (constant *b*'s) and if there is variation in the difficulty of items in the pool, then this statistic is 1.0. A value less than 1.0 indicates the relative amount of variation in *b*-parameters selected for examinees compared to the amount of variation in difficulty for the full item pool. This index is labeled the Proportion Reduction in Variance, PRV.

The analysis of the hypothetical perfect case and the expectations about how an actual CAT will function lead to a proposal that three indices of adaptation be used to describe the amount of adaptation that results from the implementation of a CAT: $r(\bar{b}_j, \theta_j)$, $s_{\bar{b}_j}/s_{\hat{\theta}_j}$, and $PRV = \frac{s_b^2 - pooled\ s_{b_j}^2}{s_b^2}$. Because there may be little adaptation early in a CAT when the location of the examinee is not well estimated, these indices will be considered for the items in the last half of the CAT as well as the full set of items administered to each examinee.

4 Simulation Studies of the Three Indices

A series of simulation studies were run with two overall goals. The first was to determine if the three indices being considered would function in a reasonable way under circumstances when the amount of adaptation was expected to vary. The second goal was to develop guidelines on what is considered an acceptable amount of adaptation for a test that is labeled as an adaptive test. Clearly, if the correlation between mean *b*-parameter for each examinee and $\hat{\theta}_j$ is 0.0, there is no adaptation, and if it is 1.0, adaptation is occurring. But what does it mean if the correlation is 0.6? Are there values of the correlation, the ratio of the standard deviations, and the proportion of within examinee difficulty variance (PRV) that would indicate that the use of the term “adaptive test” is questionable?

Three types of simulation studies were done to investigate the characteristics of the indices. All the studies assumed a set of items that were well fit by the Rasch model so only the *b*-parameters for the items needed to be considered. For these studies, the *b*-parameters were assumed to be known so error in the estimation of item characteristics was not a factor in the study. The following three studies were conducted: (1) variation in item pool size; (2) variation in the spread of difficulty of the items in the pool; and (3) type of exposure control applied to item selection. Studies (1) and (2) were done because it was expected that these item pool characteristics would influence the amount of adaptation in predictable ways. Exposure control was included because it was expected to degrade the amount of adaptation. These studies would give base-rates for interpreting the statistics when applied to data from operational adaptive tests.

4.1 Variation in Item Pool Size

For this simulation study, an item level adaptive test was used that had a starting proficiency estimate for all examinees of 0.0. The item selection algorithm chose for administration the item that had the most information at the current estimate of proficiency. Maximum likelihood was used to estimate proficiency when both a correct and incorrect responses were present in the response string. When only correct or incorrect responses were in the response string, the maximum likelihood estimates are infinite. For those cases, the last proficiency estimate was incremented by 0.7 after a correct response and −0.7 after an incorrect response. The value of 0.7 was selected based on early studies of bias in proficiency estimation for adaptive tests (Reckase 1975). The adaptive test was fixed length with 30 items administered to each simulated examinee.

The item pools for the simulated CAT varied from 50 to 500 in size. They were generated in the following way. The *b*-parameters for the items in the pool were randomly sampled from a *N*(0, 1) distribution. The full set of 500 *b*-parameters was generated and then they were randomly divided into ten sets of 50 items. Then the first set of 50 was used for the simulation of a 50-item pool. Then the first set was augmented by the second set to create the 100-item pool. That was used for the second simulation. This process continued, adding a set of 50 items each time, until the simulation was run on the full set of 500 items in the pool. This approach was selected to allow us to check the form of the relationships between pool size and values of the statistics. 50 items were expected to be less than adequate to support adaptation and 500 were expected to be well more than adequate.

The three adaptation statistics were computed for the simulations for each of the item pool sizes. For each item pool size, 200 simulated examinees were sampled from a standard normal distribution. The true θ for each examinee was used to compute the probability of correct response for each item that was selected and then a random number was generated from a *U*(0, 1) distribution. If that number was less than the probability, a correct response was recorded. Otherwise, an incorrect response was assigned. This process was followed for each item that was administered.

After each set of 200 simulated examinations, the three measures of adaptation were computed using the final θ-estimate and the full set of 30 items administered to the simulee. The measures of adaptation were also computed using the final estimate and the item parameters from the last 15 items. In all cases, the results were replicated 100 times so the stability of the statistics could be computed. The standard deviations of the statistics over replications are included in parentheses in the table next to the mean over the 100 replications. These results for the first set of analyses are presented in Table 1.

Table 1 has two columns of results for each statistic. The first column is based on a random sample of 200 simulees that do not have exactly a mean of 0 and a standard deviation of 1. The mean and standard deviation have sampling variation that results in small differences. The second column for each statistic has the sample standardized so that the mean is exactly 0 and the standard deviation is exactly 1. This removes the sampling variation.

As expected, the results show that as the pool size increases, the measures of adaptation increase as well. However, for this test length and item selection procedure, there is not much increase in adaptation for pool sizes greater than 300. Of the three measures of adaptation, the ratio of the standard deviations seems to be the most sensitive. For the 50-item pool size, the value of the ratio is about 0.47 and then increase dramatically with the increase in the size of the item pool.

These results suggest that a value in the low 0.90s is an indicator of good adaptation for the correlation index, a value in the mid 0.80s indicates good adaptation for the ratio of the standard deviations, and a value about 0.80 indicates

Table 1 Evaluation of adaptation for different size item pools 30 item test length using estimated θ

Pool size	Statistic					
	$r(\bar{b}_j, \hat{\theta}_j)$		$s_{\bar{b}_j}/s_{\hat{\theta}_j}$		PRV	
	With sampling variation	Standardized	With sampling variation	Standardized	With sampling variation	Standardized
50	0.89 (0.01)	0.88 (0.02)	0.47 (0.01)	0.46 (0.04)	0.62 (0.01)	0.62 (0.02)
100	0.92 (0.01)	0.91 (0.01)	0.74 (0.02)	0.71 (0.03)	0.81 (0.01)	0.78 (0.02)
150	0.92 (0.01)	0.92 (0.01)	0.79 (0.02)	0.79 (0.03)	0.82 (0.01)	0.79 (0.02)
200	0.93 (0.01)	0.93 (0.01)	0.82 (0.02)	0.83 (0.02)	0.82 (0.01)	0.79 (0.02)
250	0.93 (0.01)	0.93 (0.01)	0.84 (0.02)	0.85 (0.02)	0.81 (0.01)	0.79 (0.01)
300	0.93 (0.01)	0.94 (0.01)	0.85 (0.02)	0.87 (0.02)	0.80 (0.01)	0.79 (0.01)
350	0.93 (0.01)	0.94 (0.01)	0.85 (0.02)	0.89 (0.02)	0.79 (0.01)	0.79 (0.01)
400	0.94 (0.01)	0.94 (0.01)	0.86 (0.02)	0.89 (0.02)	0.79 (0.01)	0.79 (0.01)
450	0.94 (0.01)	0.94 (0.01)	0.87 (0.02)	0.90 (0.02)	0.78 (0.01)	0.79 (0.01)
500	0.94 (0.01)	0.94 (0.01)	0.87 (0.02)	0.91 (0.02)	0.78 (0.01)	0.79 (0.01)

good adaptation for the PRV. These values will be refined further as more examples are considered.

Table 2 presents the results for the three indicators for the last 15 items of the 30-item test. As with the analysis of the full test length, these results suggest that there is little improvement in adaptation for item pools larger than 300. Although the required item pool size for a CAT is dependent on the distribution of proficiency for the examinee population and the number of examinees who take the CAT (Reckase 2010), it has typically been recommended that an item pool should be at least 10–12 times larger than the length of the CAT (Stocking 1994). Considering these previous recommendations, the observation that there is little improvement in adaptation when the item pool is larger than ten times the test length (30 × 10 = 300) provides more support. For that pool size, the value of the correlation was 0.96, the ratio of the standard deviations was around 0.87 (note the variation in the values), and the PRV value is about 0.97. These are higher values than for the full test because the items selected are more homogeneous because there is a reasonably good estimate of the final θ after the first 15 items.

4.2 *Variation in Item Pool Spread*

Another possible characteristic of item pools that could influence the amount of adaptation is the amount of spread of difficulty of the items. If the difficulty of the items is in a restricted range, even if the item pool is large, the test cannot be appropriately adapted for examinees that are outside that range. To check the

Table 2 Evaluation of adaptation for different size item pools 30 item test length using last 15 items and estimated θ

Pool size	Statistic					
	$r(\bar{b}_j, \hat{\theta}_{\mathrm{j}})$		$s_{\bar{b}_j}/s_{\hat{\theta}_{\mathrm{j}}}$		PRV	
	With sampling variation	Standardized	With sampling variation	Standardized	With sampling variation	Standardized
50	0.66 (0.02)	0.65 (0.03)	0.37 (0.02)	0.37 (0.05)	0.71 (0.01)	0.78 (0.06)
100	0.92 (0.01)	0.91 (0.01)	0.70 (0.02)	0.67 (0.04)	0.93 (0.00)	0.94 (0.01)
150	0.94 (0.01)	0.94 (0.01)	0.79 (0.02)	0.77 (0.03)	0.96 (0.00)	0.96 (0.01)
200	0.95 (0.01)	0.95 (0.01)	0.83 (0.02)	0.83 (0.03)	0.97 (0.00)	0.97 (0.00)
250	0.95 (0.01)	0.96 (0.01)	0.85 (0.02)	0.87 (0.02)	0.97 (0.00)	0.97 (0.00)
300	0.96 (0.01)	0.96 (0.01)	0.87 (0.02)	0.89 (0.02)	0.97 (0.00)	0.97 (0.00)
350	0.96 (0.01)	0.97 (0.00)	0.87 (0.02)	0.91 (0.02)	0.97 (0.00)	0.97 (0.00)
400	0.96 (0.01)	0.97 (0.00)	0.88 (0.02)	0.92 (0.02)	0.97 (0.00)	0.98 (0.00)
450	0.96 (0.01)	0.97 (0.00)	0.89 (0.02)	0.93 (0.02)	0.97 (0.00)	0.98 (0.00)
500	0.96 (0.01)	0.97 (0.00)	0.90 (0.02)	0.93 (0.02)	0.97 (0.00)	0.98 (0.00)

performance of the measures of adaptation for these conditions, adaptive tests were simulated for item pools that had standard deviations of the *b*-parameters from 0.1 to 1.5 at 0.1 intervals. In all cases, the item pools were centered on 0.0 and the simulated examinees were sampled from a standard normal distribution. The size of the item pools was 300 and the test length was 30. As for the previous analyses, the indicators of adaptation were computed for the full test length and for the last 15 items.

As with the results for item pool size, the results for the spread of the item pool difficulty showed that the indices of adaptation improved as the spread of the item pool increased (see Table 3). It is interesting to note that for the 30-item test the values of the indices continued to improve as the standard deviations of the item pool increased beyond the standard deviations of the examinees (1.0). This suggests that the variation of difficulty parameters for an item pool should be greater than the variation in estimated θ so that there will be sufficient items for those examinees at the extremes of the θ-range. However, when the last 15 items were used for the analysis, the results did not improve as much as for the full 30-item test. This may mean that early in the test it is helpful to have a greater range of difficulty to help determine the approximate location of proficiency, but it is not as important once good estimates of location have been obtained.

The results for the spread of the item pool study were consistent with the pool size study when selecting values for the statistics. For the 30-item test, correlation in the low 0.90s, ratio of standard deviations in the mid 0.80s, and the PRV around 0.80. For the last 15 items, correlation around 0.96, ratio of standard deviation around 0.86 and PRV about 0.97 (Because of space limitations, the full table for the last 15 items is not shown.).

Table 3 Evaluation of adaptation for item pools with different standard deviations 30 item test length using estimated θ

Pool SD	Statistic					
	$r(\bar{b}_j, \hat{\theta}_\mathrm{j})$		$s_{\bar{b}_j}/s_{\hat{\theta}_\mathrm{j}}$		PRV	
	With sampling variation	Standardized	With sampling variation	Standardized	With sampling variation	Standardized
0.1	0.82 (0.01)	0.83 (0.01)	0.13 (0.00)	0.13 (0.00)	−0.29 (0.07)	−0.27 (0.08)
0.2	0.83 (0.02)	0.83 (0.01)	0.23 (0.01)	0.25 (0.01)	0.08 (0.05)	0.00 (0.06)
0.3	0.84 (0.01)	0.85 (0.01)	0.35 (0.01)	0.37 (0.01)	0.21 (0.05)	0.27 (0.04)
0.4	0.86 (0.01)	0.87 (0.01)	0.47 (0.02)	0.48 (0.02)	0.36 (0.03)	0.40 (0.03)
0.5	0.88 (0.01)	0.88 (0.01)	0.56 (0.02)	0.59 (0.02)	0.50 (0.02)	0.48 (0.02)
0.6	0.89 (0.01)	0.90 (0.01)	0.63 (0.02)	0.67 (0.02)	0.55 (0.02)	0.55 (0.02)
0.7	0.90 (0.01)	0.91 (0.01)	0.70 (0.02)	0.73 (0.02)	0.63 (0.02)	0.65 (0.01)
0.8	0.91 (0.01)	0.92 (0.01)	0.75 (0.02)	0.78 (0.02)	0.69 (0.02)	0.70 (0.01)
0.9	0.93 (0.01)	0.93 (0.01)	0.81 (0.02)	0.83 (0.02)	0.74 (0.01)	0.76 (0.01)
1.0	0.94 (0.01)	0.94 (0.01)	0.85 (0.02)	0.88 (0.02)	0.78 (0.01)	0.79 (0.01)
10.1	0.94 (0.01)	0.94 (0.01)	0.86 (0.02)	0.90 (0.02)	0.82 (0.01)	0.82 (0.01)
1.2	0.95 (0.01)	0.94 (0.01)	0.89 (0.02)	0.93 (0.02)	0.84 (0.01)	0.84 (0.01)
1.3	0.95 (0.01)	0.95 (0.01)	0.91 (0.02)	0.94 (0.02)	0.86 (0.01)	0.87 (0.01)
1.4	0.95 (0.01)	0.95 (0.01)	0.92 (0.02)	0.95 (0.02)	0.88 (0.00)	0.88 (0.01)
1.5	0.95 (0.01)	0.95 (0.01)	0.93 (0.02)	0.95 (0.02)	0.89 (0.00)	0.90 (0.00)

4.3 Influence of Exposure Control

Exposure control is an important aspect of CAT when the methodology is used for large scale assessment. To make full use of the available computer facilities, test sessions tend to be scheduled multiple times per day and every day of the week. That means that examinees can communicate with each other before and after an examination session threatening test security. The examinees can share information about the test questions they have seen and the answers to those questions.

Because of relatively continuous testing, procedures have been developed to limit the number of test questions that examinees will have in common. These procedures are called exposure control procedures (see Georgiadou et al. (2007) for a review of procedures). All the exposure control procedures have the effect of possibly selecting items that are not the best match for the current estimate of proficiency, but are almost as good as the best match. The selection of these "almost as good" test items might reduce the level of adaptation of a CAT, so it is important to determine their effect.

Two commonly used exposure control procedures were used in this research. The first is called the randomesque procedure. It was first suggested by Kingsbury and Zara (1989). This procedure identifies the N items that are best for gaining information at the current estimate of proficiency and randomly selects one of them for administration. When this procedure is used, examinees with the same estimate of proficiency will have a $1/N^2$ chance of being administered the same item.

The second exposure control procedure is the Sympson-Hetter method (Sympson and Hetter 1985). This method uses a simulation of the CAT process with the actual item pool to determine how often items will be selected for administration with the expected sample of examinees. An exposure parameter is then estimated for each item that is the probability that the item will be administered if it is selected. Items that are projected to be used frequently are given lower exposure parameters. However, when some items have low exposure parameters, other items will be selected more often. Therefore, the process is repeated after each change in the exposure parameters until all the exposure parameters reach stable values.

Because the simulation of item pool size suggested a pool size of 300 provided good adaptation, this simulation was done using an item pool of 306 that was designed using the bin-and-union procedures described in Reckase (2010). The target distribution of item difficulty parameters that was developed using the procedure is provided in Table 4. This distribution of difficulty parameters is somewhat flatter and wider than a standard normal distribution.

For the randomesque exposure control procedure, the item to be administered was randomly sampled from the five items that had b-parameters closest to the current estimate of proficiency. For the Sympson-Hetter procedure, the goal was to have a maximum item exposure of 0.20. The exposure parameters were estimated

Table 4 Item pool design for the exposure control study

Bin boundaries	Frequency	Mean b-parameter
$-5.1 \geq b > -4.5$	1	−4.68
$-4.5 \geq b > -3.9$	4	−4.19
$-3.9 \geq b > -3.3$	14	−3.62
$-3.3 \geq b > -2.7$	21	−3.01
$-2.7 \geq b > -2.1$	23	−2.35
$-2.1 \geq b > -1.5$	24	−1.75
$-1.5 \geq b > -0.9$	26	−1.20
$-0.9 \geq b > -0.3$	27	−0.54
$-0.3 \geq b > 0.3$	26	−0.02
$0.3 \geq b > 0.9$	27	0.59
$0.9 \geq b > 1.5$	26	1.18
$1.5 \geq b > 2.1$	24	1.82
$2.1 \geq b > 2.7$	23	2.37
$2.7 \geq b > 3.3$	21	2.99
$3.3 \geq b > 3.9$	14	3.68
$3.9 \geq b > 4.5$	4	4.01
$4.5 \geq b > 5.1$	1	4.75
Total	306	0.01

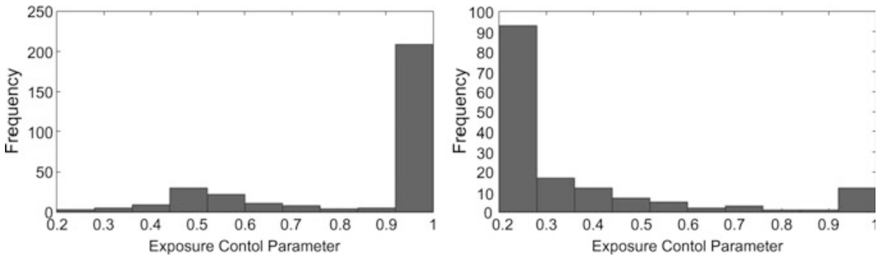


Fig. 1 Distribution of exposure control parameters for the Sympson-Hetter procedure for the 306 (left) and 153 (right) item pools

Table 5 Influence of exposure control on adaptation full test length of 30 items

		Statistic		
		$r(\bar{b}_j, \hat{\theta}_{\mathrm{j}})$	$s_{\bar{b}_j}/s_{\hat{\theta}_{\mathrm{j}}}$	PRV
Full item pool	No exposure control	0.95 (0.00)	0.97 (0.01)	0.95 (0.00)
	Randomesque procedure	0.95 (0.00)	0.97 (0.01)	0.95 (0.00)
	Sympson-Hetter method	0.96 (0.00)	0.97 (0.01)	0.94 (0.00)
Sub pool 1	No exposure control	0.96 (0.00)	0.96 (0.01)	0.93 (0.00)
	Randomesque procedure	0.96 (0.00)	0.96 (0.01)	0.93 (0.00)
	Sympson-Hetter method	0.95 (0.00)	1.18 (0.02)	0.38 (0.01)
Sub pool 2	No exposure control	0.96 (0.00)	0.95 (0.01)	0.94 (0.00)
	Randomesque procedure	0.96 (0.00)	0.94 (0.01)	0.93 (0.00)
	Sympson-Hetter method	0.94 (0.00)	0.99 (0.02)	0.32 (0.01)

through a number of iterations of the CAT process with the item pool specified in Table 4 until the exposure parameters stabilized. The distribution of the estimated exposure control parameters is presented in the left side of Fig. 1. This figure shows that for this item pool, over 200 of the items had exposure control parameters of 1.0 indicating that no exposure control was needed. About 70 items had exposure control parameters around 0.50.

The results of the exposure control study are shown in Table 5. The results for the no exposure control condition showed that the CAT procedure worked very well with the item pool designed for this study. All the statistical indices were very high. The results were also very good for both exposure control procedures. There were only slight and hardly noticeable declines in the statistics. These results are consistent with the distribution of exposure control parameters from the Sympson-Hetter method (see Fig. 1 left panel) that indicated that over 200 items did not need any exposure control. The items would be administered less than 20% of the time without any exposure control. This is probably because the test length of 30 was only 10% of the pool.

Figure 2 (left panel) shows the variation in item exposure for the three conditions. That figure shows the number of times items at different levels of difficulty

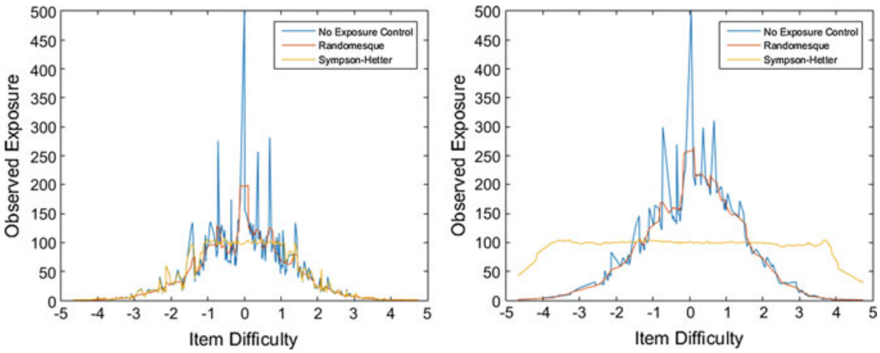


Fig. 2 Observed count of item administrations for the 306- (left) and 153-item (right) pools

were administered for the three conditions. The no-exposure-control condition has peaks for items early in the test when all examinees have the same ability estimates with all 500 examinees taking the first item with *b*-parameter closest to the starting proficiency estimate of 0.0. The randomesque procedure had much lower peaks and the Sympson-Hetter kept the frequency of item usage to about 100 administrations or lower.

Because these results were so positive, an additional study was conducted using two different random half samples of the full pool. These consisted of 153 items each. The test length remained at 30 which was now about 20% of the pool. The target exposure remained at 0.20. The results for the randomesque procedure remained very good, but the level of adaptation for the Sympson-Hetter procedure was considerably reduced (see Table 5). The correlation of the average *b*-parameter for examinees and the final θ-estimate remained high indicating that there was adaptation, but the PRV statistic and the ratio of the standard deviations showed that the items were less well targeted to the current ability estimate.

The results for the adaptation of the last 15 items (see Table 6) more dramatically show the effect on adaptation for the Sympson-Hetter procedure when the

Table 6 Influence of exposure control on adaptation last 15 items

		Statistic		
		$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV
Full item pool	No exposure control	0.98 (0.00)	1.02 (0.01)	0.99 (0.00)
	Randomesque procedure	0.98 (0.00)	1.02 (0.01)	0.99 (0.00)
	Sympson-Hetter method	0.99 (0.00)	1.03 (0.01)	0.98 (0.00)
Sub pool 1	No exposure control	0.99 (0.00)	1.02 (0.01)	0.96 (0.00)
	Randomesque procedure	0.99 (0.00)	1.02 (0.01)	0.96 (0.00)
	Sympson-Hetter method	0.87 (0.01)	1.48 (0.03)	0.08 (0.02)
Sub pool 2	No exposure control	0.99 (0.00)	0.99 (0.01)	0.97 (0.00)
	Randomesque procedure	0.99 (0.00)	0.99 (0.01)	0.97 (0.00)
	Sympson-Hetter method	0.82 (0.02)	1.21 (0.03)	−0.08 (0.02)

smaller pool size of 153 was used. For the full pool, the results are very good for all procedures. The results for the smaller pool also show the sensitivity of the Sympson-Hetter procedure to the characteristics of the item pool.

All three statistics showed a reduction in the amount of adaptation for the Sympson-Hetter procedure. The correlation between average *b*-parameter administered and the final θ-estimate was in the 0.80s compared to the high 0.90s for the randomesque procedure, the ratios of the standard deviations were 1.48 and 1.21 for the smaller item pools, and the PRV statistics were 0.08 and −0.08. This suggests that the items being selected for an examinee were almost as widely spread as the entire item pool. Figure 1 (right panel) shows the distribution of exposure parameters for one of the 153-item pools. This distribution is dramatically different from the results for the full item pool. About 90 items had exposure control parameters near 0.20. Figure 2 (right panel) shows the counts of item administrations for the three conditions. The Sympson-Hetter procedure provided better exposure control than the other procedures, but at the expense of less adaptation.

4.4 Summary of the Simulation Study Results

The simulation studies provided several important results that help to understand the amount of adaptation that is obtained when using CAT methodology. First, the studies gave some guidelines for the interpretation of the statistics that are suggested for evaluating the adaptability of a CAT. When the full-length test is analyzed, a correlation in the low 0.90s, a ratio of the standard deviation in the mid 0.80s, and a PRV of about 0.80 indicate a high level of adaptation. For the simulated 30-item test, these statistical indicators were very stable with a sample size of 300. All the empirical standard deviations (i.e., estimates of the standard error of the statistic) were small. The ratio of the standard deviations had the largest standard deviations, but these were still small. When only the last half of the test is analyzed, higher values are expected because those early items used to determine the rough location of the examinee on the scale are not included.

The results also support the guideline that the item pool should be about ten times the length of the test to support good adaptation. Also, more spread in difficulty is better than low spread, and with a good quality item pool, there is little effect of exposure control. However, when the item pool is too small, the Sympson-Hetter exposure control procedure degrades the level of adaptation. The randomesque procedure does not reduce adaptation when the 153-item pool was used, but many items in the pool had high exposure. For good quality adaptation to occur when exposure control is used, a good quality item pool is needed.

5 Operational Data Analysis

Two types of data are needed for computing the statistical indicators of the amount of adaptation. The first is the final proficiency estimate for each examinee. The second is the list of item parameters for the items administered to each examinee. The former is the usual output from a CAT. The latter is something that needs to be stored in a convenient format for computing the mean and standard deviation of the difficulty values for each examinee.

Data for the operational use of the measures of adaptation were available for the NCLEX nursing licensure test (NCSBN 2016). The total sample for an administration period was large (about 30,000) so it was also possible to take multiple random samples of 500 from the full sample to evaluate the stability of the quality of adaptation statistics. The CAT test used for this analysis was variable length with a minimum test length of 60 items and a maximum length of 250 items. The items for an individual test were selected from a large item pool of approximately 1,500 items with a difficulty distribution peaked around the decision point on the IRT scale. The mean *b*-parameter for the pool was −0.17 and the standard deviation was 1.00. Testing stopped when it was determined that an examinee was significantly different than a preset passing score or 250 operational items were administered. The testing procedure also included content balancing (eight content areas) and exposure control using the randomesque procedure—randomly selecting from 15 items with the most information at the most recent proficiency estimate. The CAT procedure was based on the Rasch model.

The results for the evaluation of the adaptability of this test are given in Table 7. For all the statistics, this test met the guidelines suggested by the simulation studies. The correlation of the mean *b*-parameter and the final estimate was in the low 0.90s, the ratio of the standard deviations was better than the guideline of the mid 0.80s, and the PRV exceed the guideline of about 0.80. These statistics indicate that this test clearly deserves the label of an adaptive test. This is even the case when the test had a variable-length stopping rule, content balancing, and exposure control. Good adaptation results from having a good item selection algorithm along with a high quality, well designed item pool.

Table 7 Adaptability statistics for an operational test

NCLEX	Statistic					
	$r(\bar{b}_j, \hat{\theta}_{\mathrm{j}})$		$s_{\bar{b}_j}/s_{\hat{\theta}_{\mathrm{j}}}$		PRV	
	Mean	SD	Mean	SD	Mean	SD
Total	0.92	0.01	0.96	0.02	0.84	0.01
Benchmark values	Low 0.90s		Mid 0.80s		0.80	

Note SD = empirical standard deviation

6 Discussion

The purpose of the research reported here was to evaluate some statistical indicators of the amount of adaptation that occurs when a test is labeled as an adaptive test. The research was stimulated by a concern that some tests labeled as CATs are so constrained in their item selection and/or have such limited item pools that the amount of adaptation may be minimal. Calling them adaptive tests might be misleading. The research comprised of a series of simulation studies designed to determine if the selected statistics were sensitive to item pool characteristics and features of a CAT that would affect adaptation. Those studies supported the use of the selected statistics and provided guidelines for interpreting the statistical indicators.

The three statistics that were selected—(1) the correlation between the mean difficulty for an examinee and the final proficiency estimate, (2) the ratio of the standard deviation of the mean difficulties and the standard deviation of the proficiency estimates, and (3) the proportion of reduction of item difficulty variance (PRV) brought about by the use of the CAT—were then applied to data from an operational CAT. The statistics indicated that this test was very adaptive even though it used content balancing and exposure control, and it used a variable-length stopping rule. The strong results were due to a well-designed, high quality item pool and a test length that was long relative to many other operational CATs.

The research reported here is initial work in this area. Future research will consider other adaptive testing designs such as multi-stage tests and those based on the three-parameter logistic model instead of the Rasch model used here. Also, other operational test data will be analyzed to determine the amount of variation that exists in the adaptability of existing tests.

References

Binet, A., & Simon, T. (1915). *A method of measuring the development of intelligence of young children* (3rd ed.). (C. H. Town, Trans.). Chicago: Chicago Medical Book Company.

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment, 5*(8), 1–39.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*(3), 153–160.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359–375.

Liu, J., Ying, Z., & Zhang, S. (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika, 80*(2), 468–490.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2002). *Computer-based testing: building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates.

NCSBN. (2016). *NCLEX-RN examination: Detailed test plan for the National Council licensure examination for registered nurses*. Chicago: NCSBN.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127–141.

Reckase, M. D. (1975, April). *The effect of item choice on ability estimation when using a simple logistic tailored testing model*. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bays procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*(4), 405–414.

Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools (ETS Research Report No. 93–2). Educational Testing Service: Princeton, NJ.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the Annual Meeting of the Military Testing Association. Navy Personnel Research and Development Center: San Diego, CA.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York: Springer.

Investigating the Constrained-Weighted Item Selection Methods for CD-CAT

Ya-Hui Su

Abstract Cognitive diagnostic computerized adaptive testing (CD-CAT) not only provides useful cognitive diagnostic information measured in psychological or educational assessments, but also obtains great efficiency brought by computerized adaptive testing. At present, there are only a limited numbers of previous studies examining how to optimally construct cognitive diagnostic tests. The cognitive diagnostic discrimination index (CDI) and attribute-level discrimination index (ADI) have been proposed for item selection in cognitive diagnostic tests. Zheng and Chang (Appl Psychol Measure 40:608–624, 2016) proposed the modified version of these two indices, an extension of the Kullback-Leibler (KL) and posterior-weighted KL (PWKL) methods, and suggested that they could be integrated with the constraint management procedure for item selection in CD-CAT. However, the constraint management procedure hasn't been investigated in CD-CAT yet. Therefore, the aim of this study is two fold (a) to integrate the indices with the constraint management procedure for item selection, and (b) to investigate the efficiency of these item selection methods in CA-CAT. It was found that the constraint-weighted indices performed much better than those without constraint-weighted procedure in terms of constraint management and exposure control while maintaining similar measurement precision.

Keywords Cognitive diagnostic models · Item selection · Constraint-weighted Computerized adaptive testing

Y.-H. Su (✉)
Department of Psychology, National Chung Cheng University, Taiwan,
168 University Rd., Minhsiung Township, Chiayi County 62102, Taiwan
e-mail: psyyhs@ccu.edu.tw

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_4

1 Introduction

Cognitive diagnostic models (CDMs) can be used to assess if students have mastered or have not mastered specific skills. Many CDMs have been proposed to obtain diagnostic information (Hartz 2002; Junker and Sijtsma 2001; Mislevy et al. 2000; Rupp et al. 2010; Tatsuoka 1983). One application of CDMs is integrating CDMs with computerized adaptive testing (CAT), denoted as cognitive diagnostic CAT (CD-CAT; Cheng 2009; Huebner 2010). The CD-CAT approach not only provides useful cognitive diagnostic information measured in psychological or educational assessments, but also obtains great efficiency brought by CAT. It provides diagnostics information to parents, teachers, and students, which can be used to direct additional instruction to the areas needed mostly by individual students.

One of the important issues in CD-CAT is how to develop the item selection algorithms. At present, there are only a limited numbers of previous studies examining how to optimally construct cognitive diagnostic tests. The cognitive diagnostic discrimination index (CDI; Henson and Douglas 2005) and attribute-level discrimination index (ADI; Henson et al. 2008) have been proposed to assemble tests followed by CDMs. The CDI measures the overall discrimination power of an item by using Kullback-Leibler (KL) information to correctly classifying the students' true status; however, the CDI itself does not provide any information about the item's discrimination power for a specific attribute (Henson et al. 2008). Therefore, ADI is proposed to measure the discrimination power of an item with respect to each of the attributes. When the attribute relationships are assumed to be nonhierarchical, the CDI and ADI have been shown to be efficient in constructing tests.

For greater generality to attribute hierarchy structure, Kuo et al. (2016) proposed the modified CDI (MCDI) and modified ADI (MADI) by considering attribute hierarchical structure and including the ratio of test length to the number of attributes. However, it was found that item usage from different attributes was still unbalanced for nonhierarchical structure. To investigate item selection in the framework of CD-CAT, Zheng and Chang (2016) proposed the posterior-weighted CDI (PWCDI) and posterior-weighted ADI (PWADI), which can be considered as an extension of the KL and posterior-weighted KL (PWKL) methods. Zheng and Chang found that the PWCDI and PWADI could obtain results as fast as the PWKL method, and suggested they can be used with constraint management procedures.

In additional to statistical optimization, the construction of assessments usually involves meeting various statistical and non-statistical constraints. For example, content balancing (selecting proportionate numbers of items from different content areas), key balancing (distributing correct answers evenly between options A, B, C, etc.), limiting specific types of items (such as those with negative stems), etc. Because items are selected sequentially, it is challenging to meet various constraints simultaneously in CAT context. Many item selection methods have been proposed to handle these constraints in CATs. One popular constraint management procedure is the maximum priority index (MPI), which can be used to monitor constraints simultaneously and efficiently in unidimensional and multidimensional CATs (Cheng and Chang 2009; Cheng et al. 2009; Su 2015, 2016; Su and Huang 2015; Yao 2011, 2012, 2013). Besides, the MPI procedure can be implemented easily and computed efficiently so it is widely used in operational CATs (Cheng and Chang 2009). However, the constraint management procedure hasn't been investigated in CD-CAT yet.

Because of the fundamental nature of the CDI and ADI approach, item information is the only thing to be considered for item selection. Such approach could easily lead to some items overexposed and bad pool usage. Many constraints are commonly required to increase validity on test scores while constructing tests. However, none of the previous studies (Henson and Douglas 2005; Henson et al. 2008; Kuo et al. 2016; Zheng and Chang 2016) included constraint management procedures in their studies. Since Zheng and Chang (2016) suggested the CDI and ADI approach can be used with constraint management procedures in a straightforward manner, it is important to integrate the CDI and ADI approach with the MPI procedure to achieve better test security and test validity. Therefore, this study has two fold (a) to integrate PWCDI and PWADI with the MPI procedure for item selection, and (b) to investigate the efficiency of these item selection methods in CD-CAT.

1.1 The Cognitive Diagnostic Discrimination Index (CDI) and Attribute-Level Discrimination Index (ADI)

Henson and Douglas (2005) proposed the CDI for test construction in CDMs. To extend the concept of Kullback-Leibler information (Chang and Ying 1996), the CDI of item j for any two distinct cognitive patterns $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ is defined as follows:

$$\mathrm{CDI}_j = \frac{\sum_{u \neq v} \left[h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1} D_{juv} \right]}{\sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1}}, \tag{1}$$

where

$$h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v) = \sum_{k=1}^{K} (\boldsymbol{\alpha}_{uk} - \boldsymbol{\alpha}_{vk})^2, \tag{2}$$

and

$$D_{juv} = E_{\boldsymbol{\alpha}_u} \left[\log \left[\frac{P_{\boldsymbol{\alpha}_u}(X_j)}{P_{\boldsymbol{\alpha}_v}(X_j)} \right] \right] = P_{\boldsymbol{\alpha}_u}(1) \log \left[\frac{P_{\boldsymbol{\alpha}_u}(1)}{P_{\boldsymbol{\alpha}_v}(1)} \right] + P_{\boldsymbol{\alpha}_u}(0) \log \left[\frac{P_{\boldsymbol{\alpha}_u}(0)}{P_{\boldsymbol{\alpha}_v}(0)} \right]. \tag{3}$$

In Eqs. (1), (2), and (3), $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ are $1 \times K$ attribute vectors, $P_{\boldsymbol{\alpha}_u}(1)$ and $P_{\boldsymbol{\alpha}_u}(0)$ are the probabilities of a correct response and an incorrect response given $\boldsymbol{\alpha}_u$, respectively, and $P_{\boldsymbol{\alpha}_v}(1)$ and $P_{\boldsymbol{\alpha}_v}(0)$ are the corresponding probabilities given $\boldsymbol{\alpha}_v$. X_j is the response of item j. The CDI$_j$ can be summed across items to obtain test-level CDI. To assemble a test with a good discrimination between mastery and non-mastery, items with large CDI should be selected.

To address an item's discrimination power for a specific attribute, Henson et al. (2008) defined ADI as follows:

$$\mathrm{ADI}_j = \frac{d_{j1} + d_{j0}}{2} = \frac{\sum_{k=1}^{K} d_{jk1} + \sum_{k=1}^{K} d_{jk0}}{2^K}. \tag{4}$$

For item j, d_{jk1} is the power to discriminate masters from non-masters on attribute k for item j where d_{jk0} is the power to discriminate non-masters from masters on attribute k for item j. The ADI$_j$ can be summed across items to obtain test-level ADI. To assemble a test with a good discrimination between mastery and non-mastery, items with large ADI should be selected.

1.2 The Posterior-Weighted CDI (PWCDI) and Posterior-Weighted ADI (PWADI) Methods

Zheng and Chang (2016) proposed the posterior-weighted version for the CDI and ADI, denoted as PWCDI and PWADI. These two indices can be considered as an extension of the KL and PWKL methods. For item j, the posterior-weighted **D** (PWD) matrix can be defined as follows:

$$\mathrm{PWD}_{juv} = E_{\alpha_u}\left[\pi(\boldsymbol{\alpha}_u) \times \pi(\boldsymbol{\alpha}_v) \times \log\left(\frac{P(X_j|\boldsymbol{\alpha}_u)}{P(X_j|\boldsymbol{\alpha}_v)}\right)\right], \tag{5}$$

where $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ are the updated cognitive pattern posteriors. Then, the PWCDI and PWADI are defined as follows:

$$\mathrm{PWCDI}_j = \frac{1}{\sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1}} \sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1} \mathrm{PWD}_{juv}, \tag{6}$$

and

$$\mathrm{PWADI}_j = \frac{1}{2^K} \sum_{all\_relevant\_cells} \mathrm{PWD}_{juv}, \tag{7}$$

respectively.

1.3 The Maximum Priority Index (MPI)

To improve measurement precision, test security, and test validity, the MPI method was proposed to monitor many statistical and non-statistical constraints simultaneously (Cheng and Chang 2009). Denote the constraint relevancy C_{jk}, where j is the number of items in the pool and k is the total number of constraints. $c_{jk} = 1$ represents constraint k relevant to item j and $c_{jk} = 0$ otherwise. Each constraint k is associated with a weight w_k. Usually, major constraints such as content balancing are put larger weights than others. The priority index of item j can be computed with

$$\mathrm{PI}_j = I_j \prod_{k=1}^{K} (w_k f_k)^{c_{jk}}, \tag{8}$$

where I_j represents the Fisher information of item j evaluated at the current θ and f_k measures the scaled 'quota left' of constraint k. For a content constraint k, the PI can be considered in a certain content area. After x_k items have been selected, the resulting scaled 'quota left' is

$$f_k = \frac{(X_k - x_k)}{X_k}. \tag{9}$$

Note that when $c_{jk} = 0$, meaning item j is not restricted by constraint k, the term $w_k f_k$ will not contribute to the final product PI_j. For every available item in the pool, the PI can be computed according to Eq. (8). Instead of the largest Fisher information, the item with the largest PI value will be chosen to administer.

When item exposure control is considering during item selection, assume constraint k requires that the item exposure rates of all items to be lower than or equal to $r_{\max}$, f_k can be defined as

$$f_k = \frac{1}{r_{\max}} (r_{\max} - \frac{n}{N}), \tag{10}$$

where n/N is the provisional exposure rate of item j after N examinees have taken the CATs.

When flexible content balancing constraints are required, Cheng and Chang (2009) suggested the MPI method need to be used jointly with the two-phase item selection strategy (Cheng et al. 2007). Each flexible content balancing constraint involves a lower bound l_k and an upper bound u_k. Denote the number of items μ_k to be selected from content area k. Then,

$$\sum_{k=1}^{K} \mu_k = L, \tag{11}$$

where K (k = 1, 2, …, K) and L are the total number of the content areas and the test length, respectively. In the first phase, l_k items are selected from each content area to meet the lower bound constraints such that $L_1 = \sum_{k=1}^{K} l_k$. After x_k items have been selected, the resulting scaled 'quota left' is

$$f_k = \frac{1}{l_k}(l_k - x_k). \tag{12}$$

Then, in the second phase, the remaining $L_2 = L - L_1$ items are selected within the upper bounds of each content area. The f_k can be computed by

$$f_k = \frac{1}{u_k}(u_k - x_k). \tag{13}$$

The MPI item selection method had fewer constraint violations and better exposure control while obtaining the same level of measurement precision. When flexible content balancing constraints is considered, the one-phase item selection strategy was proposed by incorporating both upper bounds and lower bounds. The PI becomes

$$\mathrm{PI}_j = I_j \prod_{k=1}^{K} (f_{1k} f_{2k})^{c_{jk}}, \tag{14}$$

where

$$f_{1k} = \frac{1}{u_k}(u_k - x_k - 1), \tag{15}$$

and

$$f_{2k} = \frac{(L - l_k) - (t - x_k)}{L - l_k}, \tag{16}$$

where t is the number of items that have already been administered and $t = \sum_{k=1}^{K} x_k$. The f_{1k} in Eq. (15) measures how close from the upper bound. The $L - l_k$ in Eq. (16) is the maximum number of items that can be selected from other content areas. When f_{2k} equals to 0, it represents that the items from other content areas have reached its maximum.

2 Method

The deterministic input, noisy, and gate (DINA; Haertel 1989; Junker and Sijtsma 2001) model is considered in the study. The DINA model assumes that each attribute measured by the item must be successfully applied to obtain a correct answer. The probability of getting a correct answer is defined as

$$P(X_{ij} = 1 \mid s_j, g_j, \eta_{ij}) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})}, \tag{17}$$

where

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{18}$$

indicates if examinee i has mastered all the required attributes of item j. s_j is the slip parameter, which measures the probability that an examinee with all the required attributes misses to answer the item j correctly. g_j is the guessing parameter, which measures the probability that an examinee without all the required attributes answers the item j correctly.

2.1 Simulation Design

A fixed-length CD-CAT simulation study was carried out to evaluate the efficiency of the item selection methods. Three factors were manipulated in this study: test length (short vs. long), item bank quality (low vs. high), and item selection methods (four methods). Three thousand examinees were generated by the DINA model, and every examinee had a 50% chance of mastering each attribute. The item bank had 500 items followed by five-attribute DINA model, which was similar to the previous studies (Cheng 2009; Zheng and Chang 2016). Each item had 30% chance to measuring each attribute. For low-quality item bank, item parameters s_j and g_j were generated from U(0.10, 0.30); for high-quality item bank, these two item parameters were generated from U(0.05, 0.25). The length for short and long tests was set as 5 and 10 items, respectively.

Four item selection methods were included in the study. Two were with constraint-weighted MPI and the other two were without constraint-weighted MPI. Besides the PWCDI and PWADI indices, the PWCDI and PWADI indices were integrated with the constraint-weighted MPI for item selection, denoted as CW-PWCDI, and CW-PWADI, respectively. When the CW-PWCDI, and CW-PWADI were used for item selection in CD-CAT, the Fisher information in Eq. (8) was replaced with the PWCDI and PWADI indices. Six constraints were considered in the study, including item exposure control and five attributes balanced. The maximum item exposure rates were 0.2 in the study while the constraint-weighted item selection methods were applied. To make sure items selected evenly across five attributes, the constraint of content balance was considered in the study. The efficiency of the PWCDI and PWADI item selection methods were compared with the CW-PWCDI, and CW-PWADI methods in CD-CAT through simulations in terms of constraint management, measurement precision, and exposure control.

2.2 Evaluation Criteria

The results of the simulation study were analyzed and discussed based on the following criteria: (a) constraint management, (b) measurement precision, and (c) exposure control. The constraint management was used to check whether the test sequentially assembled for each examinee meets all the specified test-construction constraints. The number of constraints violated in each test was recorded, and then the proportion of tests violating a certain number of constraints was calculated. Finally, the effectiveness of constraint management was evaluated by the averaged number of violated constraints ($\bar{V}$):

$$\bar{V} = \frac{\sum_{n=1}^{N} V_n}{N}, \tag{19}$$

where V_n represented the number of constraint violations in the nth examinees' test.

The measurement precision was evaluated by attribute correct classification rate (ACCR) and mastery pattern correct classification rate (PCCR), which were defined as follows:

$$\mathrm{ACCR}_k = \sum_{i=1}^{3000} I(\alpha_{ik} = \hat{\alpha}_{ik})/3000, \tag{20}$$

and

$$\mathrm{PCCR} = \sum_{i=1}^{3000} (\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i)/3000. \tag{21}$$

With respect to exposure control, the maximum item exposure rate, the number of overexposed items (i.e. items with exposure rate are higher than 0.20), and the number of unused items were reported. Besides, the χ^2 statistic was used to measure the skewness of item exposure rate distribution (Chang and Ying 1999)

$$\chi^2 = \frac{1}{L/500}\sum_{j=1}^{500}\left(r_j - L/500\right)^2, \tag{22}$$

where r_j is the exposure rate of item j and L is the test length. It was a good index for the efficiency of item pool usage by qualifying the discrepancy of item exposure between the observed and the expected pool usage under uniform distribution. The smaller the χ^2 statistic, the better the item exposure control.

3 Results

The results of the simulations were summarized according to constraint management, measurement precision, and exposure control in Tables 1, 2, and 3, respectively. Six constraints were considered in the study, including five attributes balanced and item exposure control. Since the violation was considered at each examinee level, six constraints were included to evaluate the efficiency of the constraint management. The proportions of assembled tests violating a certain number of constraints and the average number of violated constraints for different

Table 1 The constraint management for four item selection methods in various test lengths

Bank quality	Test length	Item selection methods		Numbers of violations						
			0	1	2	3	4	5	6	Averaged
Low	5	CW-PWCDI	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		CW-PWADI	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		PWCDI	0.323	0.323	0.354	0.000	0.000	0.000	0.000	1.031
		PWADI	0.317	0.333	0.350	0.000	0.000	0.000	0.000	1.033
	10	CW-PWCDI	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		CW-PWADI	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		PWCDI	0.327	0.323	0.350	0.000	0.000	0.000	0.000	1.023
		PWADI	0.311	0.334	0.355	0.000	0.000	0.000	0.000	1.044
High	5	CW-PWCDI	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		CW-PWADI	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		PWCDI	0.370	0.315	0.315	0.000	0.000	0.000	0.000	0.945
		PWADI	0.356	0.300	0.344	0.000	0.000	0.000	0.000	0.988
	10	CW-PWCDI	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		CW-PWADI	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		PWCDI	0.380	0.320	0.300	0.000	0.000	0.000	0.000	0.920
		PWADI	0.329	0.350	0.321	0.000	0.000	0.000	0.000	0.992

Note. For a certain number of the violated constraints (from 0 to 6), the proportions of violating tests were recorded. The average number of violated constraints for different item selection methods lists in the last column of the table

Table 2 The ACCR and PCCR for four item selection methods in various test lengths

Bank quality	Test length	Item selection methods	ACCR					
			A1	A2	A3	A4	A5	PCCR
Low	5	CW-PWCDI	0.740	0.740	0.738	0.741	0.742	0.556
		CW-PWADI	0.771	0.752	0.761	0.763	0.749	0.573
		PWCDI	0.739	0.768	0.737	0.741	0.716	0.542
		PWADI	0.725	0.780	0.736	0.740	0.700	0.554
	10	CW-PWCDI	0.816	0.820	0.818	0.821	0.823	0.736
		CW-PWADI	0.810	0.812	0.811	0.810	0.811	0.733
		PWCDI	0.811	0.814	0.814	0.808	0.802	0.713
		PWADI	0.800	0.810	0.806	0.796	0.800	0.712
High	5	CW-PWCDI	0.889	0.890	0.888	0.891	0.892	0.705
		CW-PWADI	0.921	0.901	0.911	0.913	0.899	0.723
		PWCDI	0.888	0.918	0.887	0.890	0.866	0.692
		PWADI	0.875	0.918	0.886	0.890	0.850	0.704
	10	CW-PWCDI	0.966	0.970	0.968	0.971	0.973	0.887
		CW-PWADI	0.959	0.962	0.961	0.959	0.962	0.883
		PWCDI	0.961	0.965	0.964	0.958	0.952	0.863
		PWADI	0.951	0.960	0.956	0.946	0.950	0.861

Table 3 The item exposure control for four item selection methods in various test lengths

Bank quality	Test length	Item selection methods	Maximum rate	Overexposed items	Unused items	Chi-square
Low	5	CW-PWCDI	0.167	0	8	1.389
		CW-PWADI	0.171	0	5	1.380
		PWCDI	0.358	11	20	11.765
		PWADI	0.421	13	21	10.354
	10	CW-PWCDI	0.170	0	7	1.381
		CW-PWADI	0.180	0	6	1.379
		PWCDI	0.423	15	21	13.754
		PWADI	0.433	16	19	12.833
High	5	CW-PWCDI	0.181	0	6	0.891
		CW-PWADI	0.182	0	5	0.913
		PWCDI	0.430	14	35	28.987
		PWADI	0.432	15	30	32.011
	10	CW-PWCDI	0.182	0	5	0.971
		CW-PWADI	0.183	0	5	0.959
		PWCDI	0.315	21	36	39.491
		PWADI	0.431	19	32	46.322

item selection methods list in Table 1. In general, the constraint-weighted item selection methods (i.e. CW-PWCDI and CW-PWADI) performed much better than the PWCDI and PWADI for different bank quality and different test length in terms of constraint management. When the item bank quality is low, the CW-PWCDI and CW-PWADI yielded zero in the averaged violations whereas the PWCDI and PWADI obtained the averaged violations ranging from 1.023 to 1.044. When the item bank quality is high, the CW-PWCDI and CW-PWADI still yielded zero in the averaged violations whereas the PWCDI and PWADI obtained the averaged violations ranging from 0.920 to 0.992. No matter the bank quality is high or low, the PWADI yielded slightly larger rates in the averaged violations than the PWCDI for both test length.

With respect to measurement precision, the ACCR and PCCR for four different item selection methods in various test lengths list in Table 2. The ACCR was calculated on the basis of five-attribute levels, including A1, A2, A3, A4, and A5. In general, the constraint-weighted item selection methods (i.e. CW-PWCDI and CW-PWADI) performed slightly better than the PWCDI and PWADI in terms of measurement precision. That is, the CW-PWCDI and CW-PWADI yielded slightly higher ACCR and PCCR than the PWCDI and PWADI. The longer tests, the higher ACCR and PCCR would be. The higher test quality, the higher ACCR and PCCR would be. For 5-item tests, the PWCDI performed slightly worse than the PWADI, and the CW-PWCDI performed slightly worse than the CW-PWADI. For 10-item tests, however, the PWCDI performed slightly better than the PWACDI, and the CW-PWCDI performed slightly better than the CW-PWACDI.

With respect to exposure control, the actual item exposure rates of each item were recorded. The maximum item exposure rate, the number of overexposed items, the number of unused items, and the chi-square statistic measuring the skewness of the item exposure rate distribution were calculated. The results of exposure control for different item selection methods list in Table 3. In general, the constraint-weighted item selection methods (i.e. CW-PWCDI and CW-PWADI) outperformed the other two methods for both test lengths and both bank quality in terms of item exposure control, especially when the bank quality is high. The CW-PWCDI and CW-PWADI yielded lower maximum item exposure rates, less overexposed items, less unused items, and smaller chi-square statistics than the PWCDI and PWADI. The performance of the CW-PWCDI and CW-PWADI was very similar. The longer test length, the worse the PWCDI and PWADI would be. The higher test quality, the worse the PWCDI and PWADI would be.

4 Discussions

The CD-CAT provides useful cognitive diagnostic information measured in psychological or educational assessments. It also obtains great efficiency brought by computerized adaptive testing. However, there are only a limited numbers of previous studies examining how to optimally construct cognitive diagnostic tests.

This study proposed to integrate the constraint-weighted MPI with the PWCDI and PWADI for item selection in CD-CAT. It was found that the CW-PWCDI and CW-PWADI outperformed the PWCDI and PWADI in terms of constraint management and exposure control while maintaining similar measurement precision to the PWCDI and PWADI. The constraint-weighted item selection methods (i.e. the CW-PWCDI and CW-PWADI) has great potential for item selection in operational CD-CAT.

Some future research lines are addressed as follows. First, only the fixed-length CD-CAT is considered in the study. Each examinee has different measurement precision when a fixed-length stopping rule is considered. It might result in a high misclassification rate, which might be unfair to some examinees. To achieve the same level of measurement precision to all examinees, some examinees may need to take more items and some may need to take fewer items. However, some research questions need to be investigated when a stopping rule of measurement precision is considered. It is important to investigate the constraint-weighted item selection methods in variable-length conditions for CD-CAT in the future. Second, this study only considered the simulated item bank with five-attribute DINA model, which was similar to previous studies (Cheng 2009; Zheng and Chang 2016). Besides, six constraints were considered in the study. It would be worth to investigate the efficiency of the CW-PWCDI and CW-PWADI item selection methods in an operational CD-CAT pool with different number of attributes, different number of constraints, other cognitive diagnosis models, and other constraint-weighted procedures.

References

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213–229.

Chang, H.-H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23,* 211–222.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74,* 619–632.

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.

Cheng, Y., Chang, H.-H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement, 31,* 467–482.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 301–321.

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

Henson, R. A., & Douglas, J. (2005). Test construction for cognitive diagnostics. *Applied Psychological Measurement, 29,* 262–277.

Henson, R. A., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32,* 275–288.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research and Evaluation, 15*(3), 1–7.

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

Kuo, B.-C., Pai, H.-S., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement, 40,* 315–330.

Mislevy, R., Almond, R., Yan, D., & Steinberg, L. (2000). *Bayes nets in educational assessment: Where do the numbers come from?* Princeton, NJ: CRESST/Educational Testing Service.

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

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). Switzerland: 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). Switzerland: Springer.

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

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.

Zheng, C., & Chang, H.-H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 40,* 608–624.

Modeling Accidental Mistakes in Multistage Testing: A Simulation Study

Thales A. M. Ricarte, Mariana Cúri and Alina A. von Davier

Abstract Stress in tests may cause individuals to underperform. In an adaptive test context, earlier mistakes due to stress can raise the risk of administering inadequate items to the examenees leading to an underestimation of their ability. In this paper, the effects of accidental mistakes on the first stage of an Multistage Adaptive Testing (MST) were analyzed in a simulation study. Two Item Response Theory models were used in this study: the Two-Parameter Logistic and the Logistic Positive Exponent models. Two groups were created: one group had a probability of making accidental mistakes and one did not have this probability. Comparison of latent trait estimates accuracy and the impact on the item selection process of the MST (Routing) between these two models were made. Results shows that both models had similar performance with slightly differences depending on the procedures to simulate the responses.

Keywords Logistic positive exponent · Multistage adaptive testing
Routing · Accidental mistakes

1 Introduction

Stressing about obtaining a good result on an important test can lead examinees to score less than they would had scored if there were no stress involved. Beilock (2010) stated that “Although people may certainly be motivated to perform their best under stress, these environments can cause people to perform at their worst”.

T. A. M. Ricarte (✉)
University of São Paulo, Surface mail: Rio Negro lane, 1030, apt 2113, Barueri, SP CEP: 06454-000, Brazil
e-mail: thalesamr@gmail.com

M. Cúri
University of São Paulo, São Carlos, SP, Brazil

A. A. von Davier
ACT, Iowa City, IA 52244, USA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_5

In Computerized Adaptive Testing (CAT), if an examinee performs well on an item of a specific difficulty, (s)he will be presented with a more difficult question. Or, if (s)he performed poorly, then a simpler question will be presented. Consequently, a worse performance than expected (according to his/her latent trait level) would lead to a test with easier items and an underestimation of the latent trait. Chang and Ying (2008) showed that subsequent items administered in adaptive tests (that are easier and more discriminating) are ineffective to move the estimate close to the true latent trait value, unless the test is sufficiently long.

The focus of this paper is on a specific type of adaptive testing, namely the Multistage Adaptive Test (MST, Yan et al. 2014). This type of test is composed of pre-assembled short linear tests called *modules* that are administered in stages (minimum of 2), each stage contains one or more modules. These modules have different levels of difficulty.

Figure 1 shows a diagram of a three-stage MST (represented by the levels of the diagram) and seven modules (represented by the yellow boxes). The first stage contains one module (named Routing module), the second and third stages have three modules each.

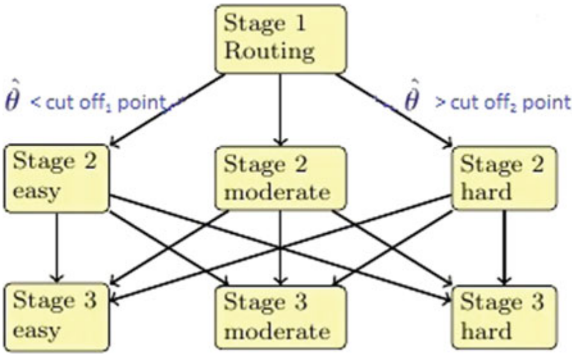
Usually, the first stage of an MST has one module. Since, not much information about θ is known in this stage, the first module can be composed of items with large range of difficulty levels.

After a stage is completed (unless it is the last one), a new module from the next stage is administered to the examinee. This module is chosen by a selection method that is based on the individual performance on previous stages. The module selection criteria is called routing.

Several approaches can be implemented as routing rules. For example: number of correct responses or cut-off points for the θ estimates. In Fig. 1, the routing rule is represented by "$\hat{\theta}$ < cut-off$_1$ point" and "$\hat{\theta}$ > cut-off$_2$ point". The arrows represent the possible modules that could be administered to the examinee depending on his/her $\hat{\theta}$.

Because of this structure, the MST has some practical advantages over CAT: (a) MST is easier to implement than CAT (for example, MST does not need to be

Fig. 1 Diagram of an example of an MST with three stages



administrated via computer); (b) whereas test administrations are constructed item by item in CAT, they are constructed with few fixed modules in MST, which makes it easier to validate the test content and fairness, and (c) in MST, individuals can review their responses within each module, which is not possible in CAT.

In this paper, the MST will be based on Item Response Theory (IRT) models to describe the probability of an individual with a latent trait level to correctly respond an item. Two models are compared: the Two-Parameter Logistic (2PL) model which is a well known IRT model and the Logistic Positive Exponent (LPE) model (Samejima 2000) that adds an exponential parameter to the 2PL. Items under the LPE model have an asymmetric Item Characteristic Curve (ICC).

The 2PL model was chosen because their item parameters are easier to estimate than more complex models like the Three-Parameter Logistic (3PL) model. The 2PL is implemented in high-stakes tests like Test of English as a Foreign Language (TOEFL, About the TOEFL Test 2017) and The Graduate Record Examination (GRE, About the GRE general test 2017).

The second model adopted in this study is the LPE and it can be written as

$$P_{LPE}(X_{ij} = 1 \mid \theta_i, a_j, b_j, \lambda_j) = \left[\frac{1}{1 + \exp(-a_j(\theta_i - b_j))}\right]^{\lambda_j}, \quad (1)$$

where X_{ij} is a binary random variable that assumes the value of 1 if the examinee with latent trait $\theta_i, i \in \{1, ..., I\}$, chooses the correct response for the item $j \in \{1, ..., J\}$, and 0 otherwise; $P_{LPE}(X_{ij} = 1 \mid a_j, b_j, \lambda_j, \theta_i)$ is the LPE probability of the examinee to correctly respond to the item; $a_j > 0$, b_j, $\lambda_j > 0$ are the discrimination, the difficulty and the acceleration parameters, respectively.

Notice that the 2PL model, $P_{2pl}(X_{ij} = 1 \mid \theta_i, a_j, b_j)$, can be obtained by fixing the $\lambda = 1$ in (1). Moreover, if the a parameter is also fixed to 1, the Rasch model is obtained.

The reason the LPE model was chosen in this study can be found in Ricarte (2016), where a particular case of the model (considering $a_j = 1, j \in \{1, ..., J\}$) was implemented in several MST simulations. It was shown that for items with $\lambda > 1$, the right answers to more difficult items can have greater positive impact on the individual's ability estimate than in the 2PL. This could be useful to help individuals recover from accidental mistakes in the beginning of the test. In his dissertation, it was observed that the item parameter estimation of the LPE is complicated using the Marginal Maximum Likelihood as well as a Bayesian MCMC approach. For this reason, in this study, the exponential parameter was fixed.

For a better understanding of the LPE model, Fig. 2 shows examples of the Item Characteristic Curve (ICC) for the LPE model with different parameter values. In these examples, b is fixed at 0, $a = 1$ (black curves) or $a = 2$ (red curves), and λ assumes 0.5, 1 and 2 values. Notice that for $a = 1$ and $\lambda = 1$ the Rasch model ICC is reached and for $a = 2$ and $\lambda = 1$ the curve represents a 2PL's ICC. For $\lambda = 0.5$, the LPE's ICC are dislocated to the left (in relation to the curves with $\lambda = 1$) and the

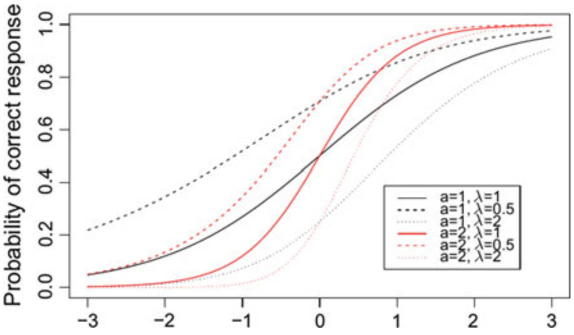


Fig. 2 Examples of LPE's ICC with parameters $a = 1$ (black curves) or $a = 2$ (red curves), $b = 0$ and different λ values. Rasch is represented when LPE with $a = 1$ and $\lambda = 1$

curve are less steep. For $\lambda = 2$, the LPE's ICC are on the right side and the curve are steeper.

Notice that $\lambda \neq 1$ causes the LPE's ICC to be asymmetric. As consequence, items with $\lambda < 1$ cause wrong answers to easier items to have greater negative impact in the individual's ability estimate than the 2PL. In contrast, for items with $\lambda > 1$, right answers to more difficult items have greater positive impact on the individual's ability estimate than in the 2PL (Samejima 2000).

The Present Study

The aim of this study is to compare the effects of accidental mistakes on the parameter estimation of 2PL and LPE models, and on the selection of modules in an MST. For this purpose, three simulation studies using different criteria to simulate individual responses were made and three LPE models with different fixed values for λ ($\lambda = 2, 4$, and 6 for all items) were used for each simulation.

For all simulations, the individual samples were segmented in two groups: half of the individuals, named No mistakes group, were considered not to be susceptible to make causal mistakes, while the other half, named Mistakes group, had a probability to make them. This division was made because the stress caused on a high-stakes test can vary to each individual.

Additionally, in a test composed of multiple-choice items, individuals have a probability of giving a correct response by chance to an item (probability of guessing), this was taken into account in two simulations as well.

To easily simulate the set of responses with or without accidental mistakes and guessing, a Four-Parameter Logistic (4PL) model (Barton and Lord 1981) was used. This model has item parameters that accounts for both situations and can be written as

$$P_{4PL}(X_{ij} = 1 \mid \theta_i, a_j, b_j, c_j, d_j) = c_j + \frac{d_j - c_j}{1 + \exp(-a_j(\theta_i - b_j))}, \tag{2}$$

where i, j, X_{ij}, a_j and b_j are defined in (1), $P_{4PL}(X_{ij} = 1 \mid \theta_i, a_j, b_j, c_j, d_j)$ is the 4PL probability of the examinee to correctly respond to the item j, $0 < c_j < 1$ is the guess-

Table 1 4PL's *c* and *d* parameter values used in the three simulations of individual responses for each group, other parameters were randomly generated

	Simulation 1		Simulation 2		Simulation 3	
Group	c	d	c	d	c	d
No mistake	0	1	0.2	1	0.2	1
Mistake	0	0.8	0	0.8	0.2	0.8

ing parameter of item j, $0 < d_j < 1$ is the accidental mistake parameter. Notice that the higher the d_j the lower probability of accidental mistake. If $d_j = 1$ then 3PL model is obtained. Additionally, the 2PL can be reached if $d_j = 1$ and $c_j = 0$.

Table 1 illustrates the *c* and *d* parameters of the 4PL used in the three simulation studies to represent different scenarios. The *a* was sampled from the log-normal (0, 0.5), *b* and θ parameters were both sampled from the standard Normal distribution. Notice that, because of how the groups were defined, No mistake and Mistake *d* parameter values are equal to 1 and 0.8, respectively, for all simulations.

In the first simulation, guessing was not considered for neither group ($c = 0$ for both groups). In second simulation, a probability of guessing was considered for the No mistakes group, but not for the Mistakes group ($c = 0.2$ for No mistakes group and $c = 0$ for the Mistakes group). In the final simulation, both groups had probability to correctly guess the response for all items ($c = 0.2$ for both groups).

2 Methods

In this Section, details of the simulation specifications of this study are segmented in three parts. In Sect. 2.1, the estimation methods used to fit the 2PL and LPE models are described. In Sect. 2.2, the MST structure, response generation, item and individual parameters used in the study are explained. In Sect. 2.3, measures to aid in the comparison of the models analyzed are presented.

2.1 Estimation

In this Subsection, the methods to estimate the item parameters and latent traits are described. First, a Bayesian MCMC method to estimate the item parameters is presented. Afterwards, the EAP method to estimate the latent traits is shown.

2.1.1 Bayesian MCMC Estimation

In this approach, the parameters given the data (posterior distribution) have a distribution composed of the likelihood function and a distribution that reflects on prior knowledge of the parameters (prior distribution).

Chains of values are sampled for each parameter from the posterior distribution. The resultant chains were used to infer about the parameters of the model. In this paper, the sampling was done using a Metropolis-Hasting Algorithm using the Winbugs Software.

For both 2PL and LPE model, the prior distribution used in the estimation algorithm were $Log - Normal(0, 0.5)$, $Normal(0, 2)$ and $Normal(0, 1)$ for a, b and θ parameters, respectively. Three chains were generated with 50000 samples each, burn-in of 10000 and thinning of 10 were made.

2.1.2 Latent Trait Estimation

Usually, in the application of an MST, the item parameters are estimated before the test (by administering the items to a sample of the target population). In the test administration, these estimates are used as fixed values for the item parameters and a method is used to estimate the latent traits. In this paper, the estimation method used is the Expected a Posteriori which consists in calculating the expected value of the θ's posterior distribution. The standard normal distribution was used as prior for θ.

2.2 Model Specifications

In this study, only the module on Stage 1 and the first routing are specified. Because of that, the specification of later stages and models were not considered. Three simulations were made to study the differences between 2PL and LPE models in a first stage of an MST scenario.

For all simulations, 5000 individual responses to 20 items were generated. The individuals, parameters were sampled from the standard Normal distribution.

For all simulated items, the a and b parameters were sampled from the Log-Normal (0, 0.5) and the standard Normal distributions, respectively. These items compose the module used in this study.

The individuals were divided in 2 groups: 2500 individuals are not susceptible to make accidental mistakes and 2500 individuals can make mistakes. These groups were denominated as No mistakes and Mistakes groups, respectively. The differences between groups are specified by the model used to simulate their responses (see Table 1).

The 2PL and three cases of LPE models with fixed acceleration parameter values were fitted to the data. The three fixed values used for LPE's λ parameter were $\lambda = 2, 4$ and 6 for all items.

2.3 Dependent Variables

To evaluate the precision of the latent trait estimate results, two measures were calculated for each group. The bias is written as

$$\text{Bias}(\hat{\theta}) = \frac{\sum_{i=1}^{2500}(\theta - \hat{\theta}_i)}{2500}, \tag{3}$$

where θ_i is the latent trait parameter for the i-th individual, $\hat{\theta}_i$ is the estimated value of θ_i for the i-th individual. The other measure is the RMSE that was calculated as

$$\text{RMSE}(\hat{\theta}) = \sqrt{\frac{\sum_{i=1}^{2500}(\theta - \hat{\theta}_i)^2}{2500}}. \tag{4}$$

To study the effects of the models in routing on the earliest stage of an MST, a range of cut-off points were specified for both easy and hard modules (ranges from −1.5 to 0 and 0 to 1.5 for the easy and hard modules, respectively). Then, both the true and estimated values of θ were routed accordingly. The proportion of agreement on the routing (for example: the true and estimated θ were routed to the easy module or both of them were routed to the hard module) were calculated for all cut-off points.

The cut-off points (one for easy and one for hard modules) with the highest proportions accordance (among the ones in the range studied) were selected and their proportion of agreement were denominated as correct easy route and correct hard route for the routing to the easy and hard modules, respectively.

To analyze the assessment rating quality using the 2PL and LPE models, the true and estimate θ ranks for each model were compared by calculating the Spearman correlation coefficient and the Weighted Cohen Kappa (Cohen 1968), which measures the inter-rate agreement between two classifiers and also accounts for the degree of their disagreement. The Weighted Cohen Kappa in our study was written as

$$\kappa = 1 - \frac{\sum_{i=1}^{2500}\sum_{j=1}^{2500} w_{ij}o_{ij}}{\sum_{i=1}^{2500}\sum_{j=1}^{2500} w_{ij}e_{ij}}, \tag{5}$$

where o_{ij} and e_{ij} are the observed and expected accuracy, respectively, and $w_{ij} = (i - j)^2$ are the elements of the weight matrix.

3 Results

Table 2 shows bias, RMSE, the correct easy route and correct hard route for the No mistakes and Mistakes groups for all three simulations previously described.

Table 2 Bias, RSME, ranking, correct module routes, Spearman correlations and Cohen Kappa coefficient for the ranks for the 2PL and LPE models fitted for all simulations

		LPE			2PL
		$\lambda = 6$	$\lambda = 4$	$\lambda = 2$	
Simulation 1					
No mistakes (c = 0, d = 1.0)	Bias	−0.17	−0.17	−0.17	−0.17
	RMSE	0.33	0.34	0.34	0.34
	Correct easy route	0.82	0.82	0.81	0.81
	Correct hard route	0.96	0.96	0.97	0.97
	Spearman cor.	0.94	0.94	0.94	0.94
	W. Cohen Kappa	0.94	0.94	0.94	0.94
Mistakes (c = 0, d = 0.8)	Bias	0.16	0.16	0.16	0.17
	RMSE	0.39	0.39	0.39	0.39
	Correct easy route	0.43	0.43	0.43	0.42
	Correct hard route	0.55	0.56	0.56	0.56
	Spearman cor.	0.89	0.89	0.90	0.89
	W. Cohen Kappa	0.89	0.89	0.89	0.89
Simulation 2					
No mistakes (c = 0.2, d = 1.0)	Bias	−0.40	−0.40	−0.41	−0.41
	RMSE	0.53	0.53	0.53	0.53
	Correct easy route	0.61	0.61	0.60	0.60
	Correct hard route	0.97	0.97	0.97	0.97
	Spearman cor.	0.86	0.86	0.86	0.86
	W. Cohen Kappa	0.89	0.89	0.89	0.90
Mistakes (c = 0, d = 0.8)	Bias	0.41	0.41	0.40	0.40
	RMSE	0.53	0.53	0.53	0.53
	Correct easy route	0.30	0.30	0.30	0.30
	Correct hard route	0.65	0.65	0.66	0.66
	Spearman cor.	0.87	0.87	0.87	0.87
	W. Cohen Kappa	0.79	0.79	0.80	0.80

(continued)

Table 2 (continued)

		LPE			2PL
		$\lambda = 6$	$\lambda = 4$	$\lambda = 2$	
Simulation 3					
No mistakes (c = 0.2, d = 1.0)	Bias	−0.24	−0.23	−0.23	−0.24
	RMSE	0.42	0.42	0.42	0.42
	Correct easy route	0.74	0.74	0.74	0.74
	Correct hard route	0.94	0.94	0.94	0.94
	Spearman cor.	0.89	0.89	0.89	0.90
	W. Cohen Kappa	0.86	0.86	0.86	0.86
Mistakes (c = 0.2, d = 0.8)	Bias	0.24	0.24	0.24	0.24
	RMSE	0.54	0.54	0.54	0.53
	Correct easy route	0.40	0.40	0.40	0.40
	Correct hard route	0.60	0.60	0.60	0.60
	Spearman cor.	0.79	0.79	0.79	0.79
	W. Cohen Kappa	0.87	0.87	0.87	0.87

Additionally, the Spearman correlation and weighted Cohen Kappa for the No mistakes and Mistakes groups were also displayed.

In simulation 1, the results showed that the bias in the No mistakes group were negative for all cases, meaning that the latent traits were overestimated. The opposite effect occurs in the group that was susceptible to make mistakes. The RMSE of the latent trait estimates was greater for the Mistakes group in all models. The Spearman correlation and Cohen Kappa were very similar in all models for both the No mistakes and Mistakes groups. Only about half of the routing was done correctly for the Mistakes group in all models. However, almost 90% of the individuals on the No mistakes group were routed correctly. There weren't big performance differences among models.

In the second simulation, the same interpretation of the bias in the first simulation holds. The No mistakes group latent trait estimates had negative biases, while the Mistakes group's ones were positive. However, the magnitude of the biases in all models considered were higher in the second simulation than in the first. In both groups, the latent trait estimates using the 2PL and LPE models had similar RSME, Spearman correlation and Cohen Kappa, as in the first simulation. Comparing to the previous simulation, the routing for the easy module of the No mistakes group were worse in the second simulation than in the first (approximately 0.60 versus 0.80 in average). This also was the case for the correct easy route of the Mistakes group. However, the correct hard route values were higher in the second simulation than in the first one.

In the third simulation, similar to the previous ones, the biases of the estimates indicate that in the No mistakes group θ was overestimated, while in the Mistakes groups θ was underestimated. No other important differences were found among models considering bias, RMSE, correlation and Kappa. The third simulation routing performance for the No mistakes group was better than in the second simulations, but worse than in the first. This same pattern occurs in the correct easy route for the Mistakes group. The correct hard route values in the third simulation were higher than the in the first simulation but lower than in the second one.

4 Discussion and Conclusion

In this paper, the LPE with fixed λ values, (2, 4 and 6) was compared with the 2PL model in simulations that emulates the consequences of accidental mistakes in the first stage of an MST.

For that purpose, three simulations were made and two groups of 2500 individuals each were created. The first group, denominated as No mistake group, responses were generated considering the 2PL or 3PL model. The second group, denominated as Mistake group, responses were generated using the 4PL model with $d = 0.8$ to simulate the occurrence of accidental mistakes.

The results showed that the 2PL and LPE models with fixed λ values had similar results for all simulations. Even though, as mentioned in Sect. 2, items under the LPE model with $\lambda > 1$ will reward more the correct answer to more difficult items than under the 2PL, which means that at least in theory these items would be more forgiving to accidental mistakes. This characteristics would also make the items to be lenient towards guessing. Because of that, when the c parameter was added to Simulations 2 and 3, it was expected that the performance of LPE model would be worse comparing to the 2PL, but this was not observed either.

For all simulations, the bias of the latent trait estimates were negative for No mistakes group and positive for the Mistakes group and their absolute values were similar. The positive bias in the Mistakes group could be explained by the accidental mistakes to the items, contributing to the underestimations of the value of the latent traits. Consequently, the joint estimation of the item parameters for both groups caused an overestimation of the No mistakes group's latent traits.

The absolute value of the biases were higher in the second simulation in comparison to the first. In the former, it was considered that the No mistakes group had a probability of guessing correctly the response to an item. Because of this, it was expected that the No mistakes group's $\hat{\theta}$ would be even higher in relation to the real θ values than in Simulation 1. For the Mistake group, the magnitude of the latent trait estimate bias was also higher in Simulation 2 than in the Simulation 1, indicating a compensation in the latent trait distribution between both groups.

In Simulation 3, where both groups could guess the correct response of the items, it was observed that the absolute values of the bias were not as high as in the second simulation. It is reasonable to conclude that there was a balance between the bias values of both groups when they had the same probability of guessing.

Ricarte (2016) mentioned that the LPE's λ parameter influences both inclination and position of the ICC. The fact that the combination of LPE's item parameters could result in similar ICC to the 2PL might explain the similarities of the models' performance in our simulations.

In this paper, no significant differences between LPE and 2PL were found. However, future studies for different scenarios need to be done to reach a final conclusion.

Studies that could to be done to improve this analysis: (a) different ways to construct the Routing Module and routing; (b) considering estimating λ instead of fixing it (specially in scenarios with guessing); and (c) comparison of the "Rasch" LPE (LPE with a parameter fixed at 1) and Rasch models may be of interest, given that there would be no influence of the a parameter on the inclination of the curve.

References

About the GRE general test. (2017). Retrieved December 7, 2017 from ETS: https://www.ets.org/gre/revised_general/about/?WT.ac=grehome_greabout_b_150213

About the TOEFL Test. (2017). Retrieved December 7, 2017 from ETS: http://kb.mit.edu/confluence/pages/viewpage.action?pageId=3907305

Barton, M., Lord, F. (1981). An upper asymptote for the Three-Parameter Logistic model, *ETS Research Report Series, 1981* (Vol. 1, pp. 1–8). https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Beilock, S. (2010). *Choke: What the secrets of the brain reveal about getting it right when you have to* (p. 153). New York: Free Press.

Chang, H.-H., Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*(3), 441–450. https://EconPapers.repec.org/RePEc:spr:psycho:v:73:y:2008:i:3:p:441-450

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Ricarte, T. (2016). *Multistage adaptive testing based on logistic positive exponent model*. Ph.D. dissertation, Inter-institute Statistics of ICMC and UFSCar São Paulo University, São Carlos. http://www.teses.usp.br/teses/disponiveis/104/104131/tde-24032017-101011/

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, *65*(3), 319–335. https://doi.org/10.1007/BF02296149.

Yan, D., von Davier, A., Lewis, C. (2014). *Computerized Multistage Testing Theory and Applications*. Chapman and Hall/CRC.

On the Usefulness of Interrater Reliability Coefficients

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

Abstract For four data sets of different measurement levels, we computed 20 coefficients that estimate interrater reliability. The results show that the coefficients provide very different numerical values when applied to the same data. We discuss possible explanations for the differences among coefficients and suggest further research that is needed to clarify which coefficient a researcher should use to estimate interrater reliability.

Keywords Agreement · Interrater reliability coefficients · Estimates of interrater reliability

1 Introduction

Interrater reliability (IRR) entails the degree of agreement, consistency, or shared variance among two or more raters assessing the same subjects, expressed as a number between 0 (no agreement) and 1 (perfect agreement). On September 27, 2017, the term "inter-rater reliability"—including quotation marks—returned 173,000 hits on Google Scholar, which illustrates its academic importance. IRR also has societal relevance. For example, in the Netherlands an officer of Child Protection Services (Raad voor de Kinderbescherming) assesses the recidivism risks, risk factors, and protective factors of each juvenile delinquent (Van der Put et al. 2011). For the juvenile delinquent, the stakes are high because the assessment by the officer of Child Protection Services determines the district attorney's

D. ten Hove · T. D. Jorgensen · L. A. van der Ark (✉)
Research Institute of Child Development and Education, University of Amsterdam,
P. O. Box 15776, 1001 NG Amsterdam, The Netherlands
e-mail: L.A.vanderArk@uva.nl

D. ten Hove
e-mail: D.tenHove@uva.nl

T. D. Jorgensen
e-mail: T.D.Jorgensen@uva.nl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_6

sentencing recommendation. If the IRR of the assessment procedure were low, the sentencing recommendation would largely depend on the officer who did the assessment, which is highly undesirable.

In our experience, most researchers associate IRR with Cohen's (1960) kappa, but there is an abundance of coefficients available. Just for nominal data, Popping (1988) identified over 38 coefficients. Zhao et al. (2013) discussed 22 of these coefficients and found several were mathematically equivalent, resulting in 11 unique coefficients. The R package `irr` (Gamer et al. 2012) contains 17 different coefficients for various types of data that estimate the IRR. Some coefficients have different versions, which increases the number of coefficients even further. For example, the intraclass correlation coefficient (ICC) can be calculated using a one-way or two-way model, to estimate the consistency or agreement of either a single rating or the average across raters. Due to the abundance of coefficients, we found that preferring a particular coefficient to estimate IRR is hard to justify. Despite review articles on IRR (e.g., Gwet 2014; Hallgren 2012), it is unknown to what degree the estimated IRR depends on the coefficient.

It would be desirable if coefficients that can be applied to data with the same measurement level (e.g., nominal data) produce similar results. Therefore, this paper investigates to what degree the choice of coefficient affects the estimated IRR. In the discussion, we attempt to explain some of the differences among coefficients, and suggest research that is needed to answer the question: "Which coefficient should a researcher use to estimate interrater reliability?".

2 Methods

2.1 Data

We selected four datasets that are freely available from the R package `irr` (see Table 1; Gamer et al. 2012). Each dataset contained the ratings of R raters observing S subjects. The dataset *Diagnoses* (Fleiss 1971) consists of ratings by six psychiatrists classifying 30 patients into one of five nominal diagnostic categories:

Table 1 Characteristics of the four datasets

Dataset	*S*	*R*	*NR*	Min	Max	Level
Diagnoses	30	6	180	1	5	Nominal
Vision	7477	2	14954	1	3	Ordinal
Video	20	4	80	2	5	Interval
Anxiety	20	3	60	1	6	Interval

Note S = number of subjects; R = number of raters; NR = number of ratings ($S \times R$); *Min* = minimum score; *Max* = maximum score depression, personality disorder, schizophrenia, neurosis, or other. The dataset *Vision* (Stuart 1953) consists of the distance-vision performance of 7477 subjects using their left eye and their right eye. The two eyes are considered the two instruments (i.e., two raters). The ratings were measured on a scale from 1 (*low performance*) to 4 (*high performance*), which we treat as ordinal. The dataset *Video* is an artificial dataset consisting of four raters rating the credibility of 20 videotaped testimonies. Ratings could vary from 1 (*not credible*) to 6 (*highly credible*), though observed scores only ranged from 2 to 5. Technically, rating scales cannot yield interval-level data unless it can be known that the distance between adjacent integers is equivalent for any pair of adjacent integers across the range of the scale; however, unbiased results may be obtained by treating Likert-type rating scales containing at least five points as interval-level rather than ordinal-level data (Rhemtulla et al. 2012). Therefore, we treated the ratings as interval-level data. The dataset *Anxiety* is also an artificial dataset, in which three raters rated the anxiety of 20 subjects on a scale from 1 (*not anxious at all*) to 6 (*extremely anxious*). The measurement level of these ratings was also treated as interval.

2.2 IRR Coefficients

We considered 20 IRR coefficients from the R package `irr` (version 0.84; Gamer et al. 2012). We considered nine coefficients for nominal ratings (Table 2, top panel). Cohen's kappa (κ; Cohen 1960) can be used only for nominal ratings with two raters. Weighted versions of κ have been derived that can also be used only for nominal ratings with two raters (Cohen 1968). The weights reflect the amount of disagreement between the raters. We calculated two weighted κ versions: κ with equal weights (κ_W) and with squared weights (κ_{W^2}). Three generalizations of κ were available to assess nominal data with more than two raters: Fleiss' kappa (κ_{Fleiss}; Fleiss 1971), Conger's exact kappa (κ_{Exact}; Conger 1980), and Light's kappa (κ_{Light}; Light 1971). The percent agreement, Krippendorff's (1980) alpha, and coefficient iota (Janson and Olson 2001) each have a version for several measurement levels, including nominal-level ratings. Their coefficients for nominal ratings are denoted PA_N, α_N, and ι_N, respectively.

We considered four coefficients for ordinal ratings (Table 2, central panel). Kendall's (1948) W and the mean of Spearman's rank-order correlation ($\bar{\rho}$; Spearman 1904) have been designed specifically for ordinal data, whereas the percent agreement and Krippendorff's (1980) alpha have a version for ordinal ratings. The latter two coefficients are denoted PA_O and α_O, respectively.

We considered seven coefficients for interval-level ratings (Table 2, bottom panel). Each coefficient can also be applied to ratio-level ratings. The percent agreement, Krippendorff's (1980) alpha, and coefficient iota (Janson and Olson 2001) have a version for interval ratings. These coefficients are denoted PA_I, α_I, and ι_I respectively. For the Finn (1970) coefficient and the ICC (Shrout and Fleiss

Table 2 Characteristics of the 20 IRR coefficients used in this study

Symbol	Name	*SE*	NHST	Miss	$R > 2$
Nominal level					
κ	Cohen's kappa	●	●		
κ_W	Weighted kappa (equal weights)	●	●		
κ_{W^2}	Weighted kappa (squared weights)	●	●		
κ_{Fleiss}	Fleiss' kappa	●	●		●
κ_{Exact}	Conger's exact kappa	●	●		●
κ_{Light}	Light's kappa	●	●		●
PA_N	Percent agreement	●			●
α_N	Krippendorff's alpha	●		●	●
ι_N	Coefficient iota				●
Ordinal level					
W	Kendall's W		●		●
$\bar{\rho}$	Mean Spearman's rank correlation				●
PA_O	Percent agreement	●			●
α_O	Krippendorff's alpha	●		●	●
Interval level					
PA_I	Percent agreement	●			●
α_I	Krippendorff's alpha	●		●	●
ι_I	Coefficient iota				●
Finn_2	Finn's coefficient (two-way)		●		●
ICC_2	Intraclass correlation coefficient (two-way)	●	●		●
$\bar{r}$	Mean Pearson's correlation				●
A	Robinson's A				●

Note SE = standard errors are available; *NHST* = null-hypothesis significance test is available; *Miss* = missing data can be handled by other methods than listwise deletion; $R > 2$ = the method can handle more than two raters

1979), we specified two-way models to treat both raters and subjects as each being randomly drawn from a population, which is often the case in social and behavioral research. In addition, for the ICC we computed the level of consistency rather than the level of absolute agreement. Furthermore, we computed the mean of Pearson's product-moment correlation coefficients ($\bar{r}$; Pearson 1895) and Robinson's measure of agreement (A; Robinson 1957).

We excluded three coefficients of the R package `irr` from our analyses, because they clearly measured something different than the IRR: the Stuart-Maxwell coefficient (Maxwell 1970) and the Bhapkar (1966) coefficient assess homogeneity in marginal distributions, and the coefficient of Eliasziw et al. (1994) estimates intrarater reliability (i.e., consistency of repeated ratings from the same rater).

2.3 Analyses

For the nominal dataset (*Diagnoses*), we applied only nominal IRR coefficients. For the ordinal dataset (*Vision*), we applied all ordinal, nominal, and interval-level IRR coefficients, with the exception of α_N and α_I. The results of interval-level coefficients are interesting because researchers frequently treat Likert-type scales as though they are continuous. The results of nominal IRR coefficients are interesting when the ordering is not of primary interest in the application at hand. Therefore, for the interval-level datasets (*Video* and *Anxiety*), we also computed all nominal, ordinal, and interval-level IRR coefficients, with the exception of $PA_N, PA_O, \alpha_N, \alpha_O,$ and ι_N.

We investigated the range of values obtained by these coefficients. We also investigated whether the choice of coefficient affects the conclusion about the IRR using the heuristic labels suggested by Landis and Koch (1977) for the use of κ: negative values indicate a poor IRR, values between 0 and 0.20 indicate a slight IRR; values between 0.21 and 0.40 indicate a fair IRR; values between 0.41 and 0.60 indicate a moderate IRR; values between 0.61 and 0.80 indicate a substantial IRR, and values between 0.81 and 1.00 indicate an almost perfect IRR.

Furthermore, we investigated the following aspects of the IRR coefficients in Table 2, by checking the literature and the functions of the package `irr`: Are standard errors available? Is it possible to conduct null-hypothesis significance testing? Are missing data allowed? And if so, how can missing data be handled? How many raters are allowed?

3 Results

Table 3 shows the variability of the evaluated IRR coefficients as estimated for the four datasets. For the nominal-level dataset *Diagnoses*, the six available IRR coefficients ranged from 0.17 (*PA*) to 0.46 (κ_{Light}; $M = 0.40$, $SD = 0.11$). For the ordinal-level dataset *Vision*, the IRR coefficients ranged from 0.60 (several coefficients) to 0.85 (*W*; $M = 0.69$, $SD = 0.09$), but from 0.71 (several coefficients) to 0.85 if only ordinal IRR coefficients are considered. For the interval-level dataset *Video*, the IRR coefficients ranged from 0.04 (κ_{Fleiss}) to 0.92 (Finn; $M = 0.26, SD = 0.24$), but from 0.10 (α_I) to 0.92 if only interval-level IRR coefficients are considered. For the interval-level dataset *Anxiety*, the IRR coefficients ranged from −0.04 (κ_{Fleiss}) to 0.54 (*W*; $M = 0.22$, $SD = 0.21$), but from 0.00 (PA_I) to 0.50 (Finn$_2$) if only interval-level IRR coefficients are considered.

Table 3 (cf. the asterisks next to the values) also shows that the interpretation of the IRR of a dataset by means of the benchmarks of Landis and Koch (1977) depends on the choice of coefficient. For the dataset *Diagnoses*, the IRR could be labelled either slight, fair, or moderate; for the dataset *Vision*, the IRR could be labelled either moderate, substantial, or almost perfect; for the dataset *Video*, the

Table 3 IRR estimates for 20 coefficients on 4 datasets

Coefficient	Diagnoses	Vision	Video	Anxiety
Nominal level				
κ	[a]	0.60*	[a]	[a]
κ_W	[a]	0.65**	[a]	[a]
κ_{W^2}	[a]	0.60*	[a]	[a]
κ_{Fleiss}	**0.43***	0.60*	0.04	−0.04
κ_{Exact}	**0.44***	0.60*	0.10	−0.02
κ_{Light}	**0.46***	0.60*	0.07	−0.02
PA_N	**0.17**	[b]	[b]	[b]
α_N	**0.43***	[b]	[b]	[b]
ι_N	**0.44***	[b]	[b]	[b]
Ordinal level				
W	[c]	**0.85*****	0.39	0.54*
$\bar{\rho}$	[c]	**0.71****	0.24	0.34
PA_O	[c]	**0.71****	[b]	[b]
α_O	[c]	**0.71****	[b]	[b]
Interval level				
PA_I	[c]	[b]	**0.35**	**0**
α_I	[c]	[b]	**0.10**	**0.16**
ι_I	[c]	0.60*	**0.15**	**0.19**
Finn$_2$	[c]	0.78**	**0.92*****	**0.50***
ICC_2	[c]	0.70**	**0.16**	**0.20**
$\bar{r}$	[c]	0.70**	**0.24**	**0.28**
A	[c]	0.85***	**0.40**	**0.48***
Ranges of values				
Range[d]	**0.17 – 0.46**	**0.71 – 0.85**	**0.10 – 0.92**	**0.00 – 0.50**
Range[e]	0.17 – 0.46	0.60 – 0.85	0.04 – 0.92	−0.04 – 0.54

Note *coefficient greater than 0.40 (moderate IRR)

**coefficient greater than 0.60 (substantial IRR)

***coefficient greater than 0.80 (almost perfect IRR)

[a] coefficient cannot be computed because the number of raters is greater than 2

[b] coefficient was not computed because a version of the coefficient that applies to another measurement level was computed

[c] coefficient was not computed because the measurement level of data is nominal

[d] range of all IRR coefficients that match the measurement level of the ratings

[e] range of all IRR coefficients

Estimates that correspond to the correct measurement level are printed in boldface

IRR could be labeled anywhere from slight to almost perfect; and for dataset *Anxiety*, the IRR could be labelled either poor, slight, fair, or moderate.

For 13 of the 20 coefficients, standard errors were available (Table 2). To the best of our knowledge, for the other coefficients, standard errors are not available.

For nine coefficients, a test statistic is available that tests whether the coefficient equals zero.

Although no dataset contained missing values, it is worth noting that the package `irr` handles missing data differently for different coefficients. Coefficients α_N, α_O, and α_I use all available data by counting disagreements among any observed pair of ratings on the same subject (i.e., pairwise deletion). Coefficients ι_N and ι_I do not allow missing ratings (i.e., the software will return a missing value for the coefficient when any ratings are missing), whereas all other coefficients handle missing data by listwise deletion.

4 Discussion

The results showed that the coefficients provide very different numerical values when applied to the same dataset. Depending on the choice of the coefficient, the IRR label for a single dataset can range from poor to almost perfect. This seriously questions the usefulness of IRR coefficients. We limited ourselves to coefficients available in the R packages `irr` (Gamer et al. 2012), so the ranges may be even wider if more coefficients were included. This problem should be investigated further.

The usefulness of the coefficients in this paper can be investigated only if IRR has a sound definition; however, a clear definition seems to be absent. Some coefficients (e.g., the ICC) are based on variance decomposition, which is compatible with the framework of generalizability theory (e.g., Vangeneugden et al. 2005), whereas other coefficients (e.g., *PA*) are derived from the concept of literal agreement. Coefficients that stem from different conceptualizations of IRR cannot all measure the same thing. In a recent discussion with Feng (2015), Krippendorff (2016) wrote: “I contend Feng discusses reliability measures with seriously mistaken conceptions of what reliability is to assure us of” (p. 139). We need to distinguish the different theories behind the IRR coefficients and come up with a more accurate terminology to identify competing conceptualizations of IRR. Only if the theories and models behind IRR are sorted out, we can start investigating why some IRR coefficients produce higher values than others, and we can separate the wheat from the chaff. In that respect, we believe the work of Zhao et al. (2013) is a valuable contribution. They explain, for example, the flaws of chance-corrected coefficients such as κ. Once we have selected estimates for different conceptualizations of IRR, we can deal with other issues identified in this study.

Another major problem is that few coefficients can handle missing data. This is problematic because ratings in the social and behavioral sciences can be expensive. For example, an assessment of a juvenile delinquent by an officer of Child Protection Services in The Netherlands (see our Introduction) takes approximately 6–8 h. A study investigating the IRR must allow for planned missingness because it is financially and practically impossible to have all officers assess all juvenile delinquents. Hence, a useful coefficient must be estimable with missing data.

We also found that for some coefficients, standard errors and confidence intervals cannot be computed and null-hypothesis testing is impossible. These standard errors, confidence intervals, and hypothesis tests should first be derived. Then the bias of all standard errors, the coverage of all confidence intervals, and the Type I error rate of all hypothesis tests should be investigated.

Finally, we used the benchmarks of Landis and Koch (1977). These benchmarks are considered to be the single most often used benchmarks (e.g., Gwet 2014, p. 164). The 42,000+ citations of the Landis and Koch paper on Google Scholar indicate at least their widespread use. A relevant question may be whether these benchmarks, which were designed for κ, can be used for coefficients stemming from different conceptualizations of IRR. In future research, it should be investigated whether different sets of heuristic rules should be provided for different types of coefficients.

References

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association, 61,* 228–235. https://doi.org/10.2307/2283057.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. https://doi.org/10.1177/001316446002000104.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213–220. https://doi.org/10.1037/h0026256.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin, 88,* 322–328. https://doi.org/10.1037/0033-2909.88.2.322.

Eliasziw, M., Young, S. L., Woodbury, M. G., & Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: Using goniometric measurements as an example. *Physical Therapy, 74,* 777–788. https://doi.org/10.1093/ptj/74.8.777.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology, 11,* 13–22. https://doi.org/10.1027/1614-2241/a000086.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30,* 71–76. https://doi.org/10.1177/001316447003000106.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–382. https://doi.org/10.1037/h0031619.

Gamer, M., Lemon, J., & Fellows, I., & Singh, P. (2012). irr: Various coefficients of interrater reliability and agreement [computer software]. https://CRAN.R-project.org/package=irr.

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*, 23–34. http://www.tqmp.org/RegularArticles/vol08-1/p023/p023.pdf.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61,* 277–289. https://doi.org/10.1177/00131640121971239.

Kendall, M. G. (1948). *Rank correlation methods*. London, UK: Griffin.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology, 12,* 139–144. https://doi.org/10.1027/1614-2241/a000119.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174. https://doi.org/10.2307/2529310.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76,* 365–377. https://doi.org/10.1037/h0031643.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry, 116,* 651–655. https://doi.org/10.1192/bjp.116.535.651.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London, 58*, 240–242. http://www.jstor.org/stable/115794.

Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research* (pp. 90–105). London, UK: Palgrave Macmillan. https://doi.org/10.1007/978-1-349-19051-5_6.

Rhemtulla, M., Brosseau-Laird, 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.

Robinson, W. S. (1957). The statistical measurement of agreement. *American Sociological Review, 22*, 17–25. http://www.jstor.org/stable/2088760.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101. https://doi.org/10.2307/1412159.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika, 40,* 105–110. https://doi.org/10.2307/2333101.

Van der Put, C. E., Spanjaard, H. J. M., van Domburgh, L., Doreleijers, T. A. H., Lodewijks, H. P. B., Ferwerda, H. B., et al. (2011). Ontwikkeling van het Landelijke Instrumentarium Jeugdstrafrechtketen (LIJ) [development of the national assessment procedure for youth criminal justice]. *Kind & Adolescent Praktijk, 10*, 76–83. http://www.tqmp.org/RegularArticles/vol08-1/p023/p023.pdf.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics, 61,* 295–304. https://doi.org/10.1111/j.0006-341X.2005.031040.x.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association, 36,* 419–480. https://doi.org/10.1080/23808985.2013.11679142.

An Evaluation of Rater Agreement Indices Using Generalizability Theory

Dongmei Li, Qing Yi and Benjamin Andrews

Abstract This study compared several rater agreement indices using data simulated using a generalizability theory framework. Information from previous generalizability studies conducted with data from large-scale writing assessments was used to inform the variance components in the simulations. Rater agreement indices, including percent agreement, weighted and unweighted kappa, polychoric, Pearson, Spearman, and intraclass correlations, and Gwet's AC_1 and AC_2, were compared with each other and with the generalizability coefficients. Results showed that some indices performed similarly while others had values that ranged from below 0.4 to over 0.8. The impact of the underlying score distributions, the number of score categories, rater/prompt variability, and rater/prompt assignment on these indices was also investigated.

Keywords Rater agreement · Generalizability · Inter-rater reliability

1 Introduction

Rater agreement is an important factor affecting the reliability of test scores involving subjective rater scoring. Numerous rater agreement indices have been proposed to measure the consistency of rater scores (Banerjee et al. 1999; Gwet 2014). However, these indices are based on different assumptions which are rarely met in practice and can result in paradoxes and abnormalities (Zhao et al. 2013). Except for a few guidelines that were developed for Cohen's kappa (e.g., Altman 1991; Fleiss 1981; Landis and Koch 1977), there is little guidance in the literature

D. Li (✉) · Q. Yi · B. Andrews
ACT, Inc, Iowa City, IA 52243, USA
e-mail: dongmei.li@act.org

Q. Yi
e-mail: qing.yi@act.org

B. Andrews
e-mail: benjamin.andrews@act.org

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_7

regarding how to interpret the values for most inter-rater agreement indices. Furthermore, raters are usually not the only source of error. In large-scale writing assessments, for example, research has repeatedly shown that the sampling of tasks tends to be an even bigger source of error variance than the sampling of raters (Breland et al. 1999). Therefore, rater agreement indices alone are not able to provide an accurate estimation of test score reliability if other sources of error are known to exist. Generalizability (G) theory (Brennan 2001), on the other hand, provides a comprehensive framework for investigating the reliability of test scores by allowing researchers to differentiate multiple sources of error.

The purpose of this study was to evaluate several rater agreement indices commonly used in the context of large-scale writing assessments using a G theory framework. A couple more recently proposed indices intended to overcome some undesirable features of earlier indices were also included. Specifically, this study was intended to answer the following research questions:

1. How do the rater agreement indices compare to the generalizability and dependability coefficients from G theory analyses? How do the indices compare to one another?
2. How does the number of rating categories affect the various rater agreement indices and the G theory coefficients?
3. What is the impact of the distribution of the underlying scores on the performance of these rater agreement indices?
4. What is the impact of rater/prompt assignment and rater/prompt variability on these indices?

These research questions were investigated based on data simulated with realistic parameters (i.e., variance components) obtained from earlier writing score generalizability research with real data. In the following sections, the rater agreement indices investigated in this study are first introduced. Then the data simulation procedures and the G theory coefficients used as criteria for comparison are described. Results are presented and discussed at the end.

2 Methods

2.1 Rater Agreement Indices

Four types of agreement indices were included in the study: (1) percent agreement, (2) Cohen’s kappa (i.e., kappa, linear and quadratic weighted kappa), (3) correlations (i.e., polychoric, Pearson, Spearman, and intraclass), and (4) two newer indices: Gwet’s AC_1 and AC_2. Below is a brief description of these indices and some known relationships among them.

Percent agreement. Percent agreement, the most intuitive indicator of rater agreement, is the percentage of cases where raters gave exactly the same ratings.

Sometimes both the perfect agreement and the adjacent agreement are reported, but this study considered only perfect agreement.

Kappa, linear and quadratic weighted kappa. Kappa (Cohen 1960) is calculated as $\frac{p_a - p_e}{1 - p_e}$, where p_a represents the percent agreement across all the categories, and p_e represents the percent agreement expected by chance. The weighted versions of kappa (Cohen 1968) take the same basic form as kappa, but weights are applied to each cell of the agreement matrix when calculating p_a and p_e. Whereas kappa is most appropriate for nominal scales, the weighted versions are for ordinal or interval scales. Vanbelle (2016) suggested that the linear and quadratic weighted kappa coefficients provide complementary information regarding position and variability, respectively, and recommended that both coefficients be reported.

Correlations. Correlations quantify the relationship between two variables. The Pearson product-moment correlation measures the linear relationship between continuous variables, and the Spearman rank-order correlation measures the monotonic relationship between two continuous or ordinal variables based on the rank orders of each variable. The polychoric correlation estimates the relationship between two normally distributed latent variables from their observed ordinal values.

Previous research showed the equivalence between the quadratic weighted kappa and some of the correlation coefficients under restricted conditions. For example, Cohen (1968) showed that for a general m × m table with identical marginal distributions, the weighted kappa is equal to the Pearson product-moment correlation coefficient obtained when the nominal categories are scaled so that the first category is scored 1, the second category 2, and so on. According to Schuster (2004), the Pearson product-moment correlation is insensitive to differences in rater means and variances, but quadratic weighted kappa is sensitive to both mean and variance differences between raters.

Intraclass correlations (ICC), though viewed as a type of correlation, are usually conceptualized within the framework of analysis of variance (ANOVA), and expressed as a proportion of variance, i.e., $\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$. When used in the context of rater agreement, σ_α^2 represents the variance due to true differences and σ_ε^2 represents the variance due to raters. Fleiss and Cohen (1973) showed that if the categories are scaled as described above, the quadratic weighted kappa is equivalent to the ICC coefficient where the mean differences between the raters are included as a component of variability. This study included this type of ICC.

AC$_1$ and AC$_2$. All kappa statistics depend on the marginal distributions and the true prevalence of a trait, which may cause paradoxes (Brenner and Kliebsch 1996; Yang and Chinchilli 2011; Warrens 2012). Gwet (2008) proposed AC_1 as a "paradox-resistant" alternative to the kappa coefficient, and later developed AC_2 (Gwet 2010), which is a weighted version of AC_1. Like the kappas, AC_1 is for nominal scales and AC_2 is for ordinal or interval scales. These two statistics are not commonly used to report rater agreement in large-scale writing assessments, but are included in this study because of their demonstrated superiority over kappa (Gwet 2014).

2.2 G Theory

As pointed out by Brennan (2001), G theory liberalizes and extends traditional notions of reliability by allowing researchers to identify and quantify the sources of errors in a measurement procedure. It provides both a conceptual framework and a statistical framework for evaluating the consistency of scores. Below is a brief introduction to a few important G theory concepts. See Brennan (2001) for details.

The universe of admissible observations are all acceptable conditions or qualifications for each facet of the measurement procedure, such as raters and prompts in the measurement of writing proficiency. The purpose of a G study is to obtain estimates of variance components associated with a universe of admissible observations. The universe of generalization is the universe to which a decision maker wants to generalize based on the results of a particular measurement procedure. The purpose of a D study is to provide estimates of variance components and score properties for well-specified measurement procedures, including universe score variance, error variances, and two reliability-like coefficients—the G coefficient ($E\rho^2$) and the index of dependability (Φ). Whereas $E\rho^2$ is the ratio of universe score variance to itself plus relative error variance, Φ is the ratio of universe score variance to itself plus absolute error variance. Note that these coefficients, as proportions of variances, are actually different types of ICCs.

2.3 Data Generation and Evaluation Criteria

Data were simulated using a G theory framework that includes both rater and prompt variabilities. In the terminology of G theory, the universe of admissible observations for the researchers contains a large number of potential writing prompts with similar characteristics and a large number of potential raters with similar trainings, and with a potential pairing of any rater with any prompt for each individual in the population of examinees. Therefore, the population (p) and the rater and prompt facets are fully crossed. Let r represent raters and i represent writing prompts. The fully crossed G-study design can be represented as $p \times i \times r$, and each observed score (X_{pir}) for a single prompt evaluated by a single rater can be represented by (Brennan 2001, p. 6):

$$X_{pir} = \mu + v_p + v_i + v_r + v_{pi} + v_{pr} + v_{ir} + v_{pir}, \tag{1}$$

where μ represents the grand mean, and v represents the various effects, including the uncorrelated rater and prompt main effects and all the interaction effects.

Variance components with different rater and prompt variability. Each effect was normally distributed with a mean of 0 and a standard deviation from variance components observed in previous studies on writing assessments using the $p \times i \times r$ design. Two sets of variance components were used in the simulation: one with

Table 1 Variance components used for data simulation and expected D-study results for combinations of different numbers of raters and prompts

			D study results				
			$n_i^{'}$	1	1	1	1
G study variance components			$n_r^{'}$	1	2	1	2
	Data 1	Data 2		Data 1		Data 2	
σ^2 (p)	1	1	σ^2 (p)	1	1	1	1
σ^2 (i)	0.01	0.09	σ^2 (I)	0.01	0.01	0.09	0.09
σ^2 (r)	0.01	0.09	σ^2 (R)	0.01	0.01	0.09	0.05
σ^2 (pi)	0.36	0.36	σ^2 (pI)	0.36	0.36	0.36	0.36
σ^2 (pr)	0.04	0.04	σ^2 (pR)	0.04	0.02	0.04	0.02
σ^2 (ir)	0.01	0.01	$\sigma^2(IR)$	0.01	0.01	0.01	0.01
σ^2 (pir)	0.25	0.25	σ^2 (pIR)	0.25	0.13	0.25	0.13
			$\sigma^2(\delta)$	0.65	0.51	0.65	0.51
			$\sigma^2(\Delta)$	0.68	0.53	0.84	0.65
			$E\rho^2$	0.61	0.66	0.61	0.66
			$\sigma^2(\Delta)$	0.60	0.66	0.54	0.61

smaller rater and prompt variability (denoted Data 1), and one with larger variability for these two effects (denoted Data 2). These two sets of variance components only differ in terms of prompt and rater main effects, meaning that raters and prompts are more variable in Data 2. Table 1 provides details of the G study variance components used as parameters for data simulation, as well as D study results when the final scores are based on one or two raters. The following notation, which is similar to Brennan's (2001), is used in Table 1: $n_i^{'}$ and $n_r^{'}$ for the D-study sample sizes for raters and prompts, respectively; $\sigma^2(\delta)$ and $\sigma^2(\Delta)$ for the relative and absolute error variances, respectively; and $E\rho^2$ and Φ for the generalizability and dependability coefficients, respectively. Note that $E\rho^2$ did not change across the two different rater/prompt pool compositions because the interaction effects did not change, and only interaction effects involving p are used in the calculation of relative error.

Population distributions with different levels of skewness. Equation (2) is a moments (mean and standard deviation) preserving transformation that can be applied to manipulate the skewness of score distributions by changing the values of c (Reardon and Ho 2014).

$$x^{*} = t(x) = -\frac{sgn(c)}{\sqrt{e^{c^2} - 1}}\left(1 - e^{cx - \frac{c^2}{2}}\right) \tag{2}$$

Three different distributions of the population were generated, a standard normal distribution, and two levels of skewness that were obtained by applying Eq. (2) to the examinee scores (x) generated from the standard normal distribution with c being set to 0.5 or 1.

Fixed or random facets. Brennan (2001) called rater agreement indices obtained based on scores assigned by the same two raters on the same task "standardized" and those obtained from scores assigned by the same two raters on different tasks "nonstandardized". In large-scale writing assessments, standardized rater agreement indices are rarely reported. Instead, rater agreement indices are often reported under a variety of nonstandardized situations. For example, in the data used for the calculation of rater agreement indices, raters for the examinees are often randomly assigned, and the prompts taken by the examinees may be the same or different. Data in this study were simulated to mimic such situations. Observed scores for each examinee were simulated from each of the following situations: (1) a single prompt taken by all examinees and evaluated by the same two raters (FPFR), (2) a single prompt taken by all examinees and evaluated by two random raters (FPRR), and (3) each examine takes a randomly selected prompt and is evaluated by two randomly selected raters (RPRR). Scores on one prompt were simulated for the calculation of the rater agreement indices. To be able to conduct generalizability analyses under the $p \times i \times r$ design, however, scores on a second prompt were simulated to be used only for these analyses.

Number of categories. The above simulations assumed that the underlying scores are all continuous variables. However, writing assessments are often scored using a limited number of score categories. Discrete scores were created by partitioning the continuous scale into different numbers (i.e., 2, 3, 4, 5, or 6) of equally spaced intervals. For example, when partitioned into 6 intervals, scores within each interval will be 1 to 6, respectively. In this study, all agreement indices were calculated and compared based on the categorized data.

Evaluation Criteria. One major difficulty in evaluating rater agreement indices is that there is no consensus in the literature regarding what a good measure of rater agreement is. This makes it hard to establish a good criterion for the evaluation. This study changes the focus of the comparisons from which is a better rater agreement index to how the rater agreement indices compare to each other and how they compare to the generalizability coefficients when both rater and prompt variability are taken into account.

As pointed out by Brennan (2001), rater agreement indices characterize the consistency between rater scores but they do not represent the reliability of scores when other sources of error are involved or when the number of raters or prompts is more than one in the final scoring. The generalizability coefficients for scores based on one prompt and one rater were used as an appropriate baseline for comparison because error from raters and prompts are both taken into account.

Often times, the parameters used to generate data are used to calculate appropriate baselines for comparison. Because of how the data were generated in this study, however, there could be differences between the generalizability coefficients using the data generation parameters and those estimated from the simulated data.

The different numbers of categories, different population distributions, or other factors could all play a role in potential differences. Consequently, different generalizability coefficients may be more reasonable for certain comparisons. Three different sets of generalizability coefficients were considered in this study. The first set is the coefficients that were calculated using the population parameters used to simulate data. These were the best baseline when the number of categories was being investigated because as the number of categories increases, it would be expected that the estimated coefficients would be closer to the population values. The second set was the coefficients estimated from the simulated data for the $p \times i \times r$ design. These served as a reasonable baseline when comparing indices for data with a particular number of score categories or when the underlying score distribution is not normal. The third set of coefficients was estimated from data generated for the first prompt using a $p \times r$ design that treated the ratings as a facet that was fully crossed with students when ratings came from random raters. These generalizability coefficients are expected to be closer to the rater agreement indices because they only take into account rater variability.

It should be noted that there are some instances when there is no reasonable baseline. In some cases, it would be expected that certain indices would be higher or lower than any of the baseline values based on the assumptions of the indices, the data characteristics or other factors. The generalizability coefficients in this study are intended to serve simply as baselines and not values that any of the indices should necessarily closely approximate. The comparisons among the indices are potentially more informative than their relation to the baseline.

Study conditions, sample sizes, and replications. Table 2 summarizes the various factors that were taken into consideration in data generation. In short, the indices were compared under 90 combinations of the conditions (i.e., 5 numbers of score categories × 3 levels of skewness of the underlying score distributions × 3 types of rater/prompt assignment × 2 levels of rater/prompt variability).

A sample size of 1,000 was used for the generation of examinees. As mentioned earlier, scores from two raters on two prompts were generated for each examinee, though the calculation of each agreement index only used scores on the first prompt.

Table 2 Summary of study conditions

Factors	Conditions
Number of score category	2, 3, 4, 5, and 6
Skewness of underlying distributions	Normal Slightly skewed, with $c = 0.5$ in Eq. (2) (Skew1) Moderately skewed, with $c = 1$ in Eq. (2) (Skew2)
Prompt and rater assignment	Same prompt and same two raters (FPFR) Same prompt and random raters (FPRR) Random prompt and random raters (RPRR)
Prompt and rater variability	Smaller variability (Data 1) Larger variability (Data 2)

The generalizability coefficients based on the $p \times i \times r$ design were obtained based on scores from both prompts. One hundred replications were conducted for each study condition.

3 Results

Results for each agreement index and the G theory coefficients were first summarized across the 100 replications for each of the 90 combinations of study conditions. Then, to show the overall impact of a certain factor, such as the number of categories, results were aggregated across other factors. For example, to show the overall impact of the number of categories, for each index the results were summarized across all replications and then across all the other conditions. These aggregated results are shown in Table 3. Results for each combination of conditions under FPFR are presented in the appendix, along with the average standard deviations of each index across replications and across conditions.

In these tables, the generalizability and dependability coefficients for scores based on one rater and one prompt analyzed under the $p \times i \times r$ design are denoted as "Gen_1_pir" and "Phi_1_pir", respectively. Those analyzed under the $p \times r$ design are denoted as "Gen_1_pr" and "Phi_1_pr", respectively. "Gen_2_pir",

Table 3 Mean values of the indices across different conditions

		Distribution			Number of Categories					Data 1			Data 2		
Indices	Mean	Normal	Skew1	Skew2	2	3	4	5	6	FPFR	FPRR	RPRR	FPFR	FPRR	RPRR
Perfect Agree	0.61	0.61	0.61	0.60	0.77	0.69	0.60	0.51	0.45	0.62	0.62	0.62	0.59	0.59	0.59
Kappa	0.40	0.43	0.41	0.36	0.54	0.45	0.39	0.33	0.28	0.42	0.42	0.42	0.38	0.37	0.39
Linear Kappa	0.52	0.56	0.53	0.47	0.54	0.52	0.52	0.52	0.52	0.54	0.54	0.54	0.50	0.50	0.52
Quadratic Kappa	0.67	0.71	0.69	0.62	0.54	0.61	0.67	0.70	0.72	0.66	0.66	0.66	0.63	0.62	0.65
Polychoric	0.77	0.81	0.78	0.72	0.76	0.77	0.77	0.77	0.77	0.79	0.78	0.78	0.79	0.73	0.75
Pearson	0.65	0.70	0.67	0.60	0.54	0.61	0.68	0.71	0.73	0.67	0.66	0.66	0.66	0.62	0.65
Spearman	0.64	0.69	0.65	0.58	0.54	0.61	0.66	0.69	0.70	0.65	0.65	0.65	0.65	0.61	0.63
Intraclass	0.65	0.69	0.66	0.59	0.54	0.61	0.67	0.70	0.72	0.66	0.66	0.66	0.63	0.62	0.65
AC1	0.47	0.48	0.48	0.46	0.56	0.58	0.48	0.41	0.35	0.49	0.49	0.49	0.46	0.46	0.45
AC2	0.80	0.80	0.80	0.80	0.56	0.82	0.86	0.88	0.89	0.81	0.81	0.81	0.79	0.79	0.79
Gen_1_pir	0.42	0.49	0.45	0.33	0.32	0.40	0.44	0.47	0.48	0.44	0.43	0.43	0.43	0.41	0.39
Phi_1_pir	0.41	0.48	0.43	0.32	0.31	0.39	0.43	0.46	0.47	0.43	0.43	0.43	0.39	0.39	0.39
Gen_2_pir	0.51	0.58	0.53	0.41	0.41	0.49	0.53	0.55	0.56	0.53	0.52	0.51	0.52	0.50	0.47
Phi_2_pir	0.49	0.57	0.52	0.39	0.40	0.48	0.51	0.53	0.54	0.52	0.51	0.51	0.48	0.47	0.47
Gen_1_pr	0.65	0.70	0.67	0.60	0.54	0.61	0.67	0.71	0.73	0.67	0.66	0.66	0.66	0.62	0.65
Phi_1_pr	0.65	0.69	0.66	0.59	0.54	0.61	0.67	0.70	0.72	0.66	0.66	0.66	0.63	0.62	0.64
Gen_2_pr	0.79	0.82	0.80	0.75	0.70	0.76	0.80	0.83	0.84	0.80	0.79	0.79	0.79	0.76	0.78
Phi_2_pr	0.78	0.81	0.79	0.74	0.70	0.75	0.80	0.82	0.84	0.79	0.79	0.79	0.77	0.76	0.78

“Phi_2_pir”, “Gen_2_pr” and “Phi_2_pr” denote coefficients of scores based on one prompt but two raters.

The values in these tables were color coded to indicate how they compare with the generalizability coefficient (i.e., 0.61) based on the variance components used for data simulation. Compared to this target value, those in dark red were more than 0.05 higher, those in blue were more than 0.05 lower, and those in green were within 0.05. These results are discussed in more detail below as part of the findings for each of the research questions.

3.1 Research Question 1: How Do the Rater Agreement Indices Compare to the Coefficients from G Theory Analyses? How Do the Indices Compare to One Another?

The second column in Table 3 shows the overall average of each of the indices across the 100 replications and across all 90 combinations of conditions. The rest of the table shows the average values of these indices for the different levels of skewness in the population distribution and for the different numbers of categories. The values across the different indices ranged from 0.40 to 0.80. Compared with the generalizability coefficient expected from the data generation parameters (i.e., 0.61), some indices tended to produce higher values. This was expected because rater agreement indices do not take into account prompt variability. However, some indices (kappa, linear kappa, and AC_1) did produce values lower than 0.61.

Compared with the generalizability results from the $p \times i \times r$ design, almost all indices had higher values than both the generalizability and dependability coefficients. The quadratic weighted kappa results were similar to the correlations (except for the polychoric correlation) and also similar to the generalizability coefficients based on the $p \times r$ design.

3.2 Research Question 2: How Does the Number of Rating Categories Affect the Various Rater Agreement Indices and the G Theory Coefficients?

As shown in the “Number of Categories” section of Table 3, all indices tended to increase with the increase in the number of categories, except for the perfect agreement, kappa, linear weighted kappa, and AC_1. Linear weighted kappa stayed stable with the increase in the number of categories, but the perfect agreement, kappa, and AC_1 decreased with the increase in the number of categories. Note that when there were only two categories, kappa, linear and quadratic kappa would produce the same results. This is also true for AC_1 and AC_2, as well as the Pearson and Spearman correlations.

The average generalizability coefficients obtained for the categorized data were noticeably lower than the generalizability coefficients (0.61) calculated with the variance components used for data simulation. The tables in the appendix show that results from the categorized data were close to the values calculated for continuous data only when the data were normally distributed and when the number of categories was large.

3.3 Research Question 3: What Is the Impact of the Distribution of the Underlying Scores on the Performance of These Rater Agreement Indices?

As shown in the "Distribution" section of Table 3, with increased skewness in the population score distributions, all indices tended to decrease, except for AC_2. This finding is consistent with findings by Quarfoot and Levine (2016) that AC_2 is more robust to distribution changes. AC_1 and the percent agreement also stayed relatively stable with the increase in skewness.

3.4 Research Question 4: What Is the Impact of Rater/ Prompt Assignment and Rater/Prompt Variability on the Rater Agreement Indices?

The "Data 1" and "Data 2" sections of Table 3 present the average values of the indices across replications aggregated across the other conditions for the three different rater/prompt assignment designs for the two sets of variance components for rater/prompt variability. The different assignment made little difference for all the indices, probably due to the small amount of variability among raters and among prompts. For Data 2, which had more variability in rater and prompt main effects, slightly greater differences across the different rater/prompt assignments were observed, but the values for each index were still very close among the three different designs.

As shown in Table 1, the only differences between the two sets of variance components used for data simulation were in the variances for rater and prompt main effects. Based on results from previous research, these variances were small for large-scale writing assessments, probably due to the strict rater training and prompt selection procedures. The small differences found in this study for different rater and prompt assignments were consistent with findings from some earlier research (e.g., Lee and Kantor 2005).

4 Discussion

One unique characteristic of this study is that rater agreement indices were evaluated in comparison with test score reliability indices which appropriately took into account additional sources of error. With this design, it was expected that rater agreement indices would overestimate test score reliabilities by ignoring other sources of error.

Although it is common that a rating scale only consists of a small number of categories, it is reasonable to assume that the rating scale is often a categorization of an underlying continuous score scale. The current study showed that the categorization of a continuous scale tended to decrease the values of the agreement indices—the smaller the number of categories, the lower the values. Yet there were two groups of exceptions. The indices in the first group, including percent agreement, kappa, and $AC_{1,}$ tended to decrease with the increase in the number of categories. This is likely because they are intended for nominal categories, and thus do not differentiate between different extents of disagreement. For these indices, the larger the number of categories, the less likely it is to have a perfect agreement, whether chance agreement is corrected or not. The second group, including polychoric correlation and linear weighted kappa tended to remain stable with the change in the number of categories. This was expected for polychoric correlation because it is intended to measure the association of the underlying continuous variables. The reason for this behavior of linear kappa may need further investigation.

The different rater agreement indices investigated in the study produced a wide range of results (from 0.40 for kappa to 0.80 for AC_2), even after aggregating over all the conditions. More variability among the indices was found under specific conditions (See tables in the Appendix). The inconsistency of rater agreement indices is not a new finding, but the high values of AC_2 are worth mentioning because this index has rarely been used in the context of large-scale writing assessments, which is the main context of interest in this research. Also worth mentioning is the similarity of results between the quadratic weighted kappa and some of the other indices. Besides the Pearson and Spearman correlations and the intraclass correlation, the results also showed that the Spearman correlation and the generalizability coefficients based on the $p \times r$ design also had similar values, especially under the design where raters were randomly assigned to each simulated examinee.

This study also touched upon the issue of forcing sparse assignment of raters and prompts into a fully crossed design by not differentiating specific raters or prompts, but treating ratings or test occasions as a facet. The small differences found in this study are probably due to the small rater and prompt main effects assumed. Further research is needed to draw conclusive conclusions regarding this.

Finally, it is important to note that the study was based on simulated data with specific assumptions about the contributions of rater and prompt variability to writing score reliability. Although the parameters were based on previous relevant research and were intended to be realistic, these assumptions may not hold in other contexts where raters are not well trained or prompts are diverse in terms of topics, genres, types, etc.

Appendix: Detailed Results for FPFR

Table A1 Results for FPFR on Data 1

	Normal					Skewed (c=0.5)					Skewed (c=1)					
	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6	Mean SD
Perfect Agree	0.80	0.71	**0.62**	0.54	0.47	0.79	0.71	**0.61**	0.54	0.47	0.75	0.71	**0.61**	0.53	0.47	0.02
Kappa	**0.60**	0.51	0.44	0.38	0.32	**0.57**	0.48	0.42	0.36	0.31	0.49	0.42	0.37	0.32	0.27	0.03
Linear Kappa	**0.60**	**0.58**	**0.58**	**0.58**	**0.58**	**0.57**	0.55	0.55	0.55	0.55	0.49	0.48	0.49	0.49	0.49	0.02
Quadratic Kappa	**0.60**	0.67	0.72	0.75	0.77	**0.57**	**0.64**	0.69	0.73	0.75	0.49	**0.56**	**0.63**	0.67	0.69	0.02
Polychoric	0.82	0.82	0.82	0.82	0.82	0.79	0.80	0.80	0.80	0.80	0.71	0.74	0.74	0.74	0.74	0.02
Pearson	**0.61**	0.68	0.73	0.76	0.78	**0.58**	**0.64**	0.70	0.74	0.76	0.49	**0.57**	**0.64**	0.68	0.70	0.02
Spearman	**0.61**	0.67	0.72	0.75	0.77	**0.58**	**0.64**	0.69	0.72	0.73	0.49	**0.56**	**0.60**	**0.63**	**0.65**	0.02
Intraclass	**0.61**	0.67	0.72	0.75	0.77	**0.57**	**0.64**	0.69	0.73	0.75	0.49	**0.56**	**0.63**	0.67	0.69	0.02
AC1	**0.61**	**0.59**	0.50	0.43	0.37	**0.58**	**0.60**	0.50	0.43	0.38	0.51	**0.61**	0.51	0.43	0.37	0.03
AC2	**0.61**	0.82	0.87	0.88	0.89	**0.58**	0.83	0.87	0.89	0.90	0.51	0.85	0.88	0.90	0.91	0.01
Gen_1_pir	0.41	0.49	0.54	**0.56**	**0.57**	0.36	0.43	0.49	0.51	0.53	0.23	0.32	0.37	0.40	0.42	0.02
Phi_1_pir	0.41	0.48	0.53	0.55	**0.56**	0.36	0.43	0.48	0.51	0.52	0.23	0.31	0.36	0.39	0.41	0.02
Gen_2_pir	0.51	**0.58**	**0.62**	**0.64**	**0.64**	0.46	0.53	**0.57**	**0.59**	**0.60**	0.31	0.40	0.45	0.48	0.49	0.03
Phi_2_pir	0.51	**0.58**	**0.61**	**0.63**	**0.64**	0.45	0.52	**0.56**	**0.58**	**0.59**	0.31	0.40	0.44	0.47	0.48	0.03
Gen_1_pr	**0.61**	0.67	0.73	0.76	0.78	**0.58**	**0.64**	0.70	0.74	0.76	0.49	**0.57**	**0.63**	0.68	0.70	0.02
Phi_1_pr	**0.60**	0.67	0.72	0.75	0.77	**0.57**	**0.64**	0.69	0.73	0.75	0.49	**0.56**	**0.63**	0.67	0.69	0.02
Gen_2_pr	0.76	0.81	0.84	0.86	0.87	0.73	0.78	0.82	0.85	0.86	0.66	0.72	0.78	0.81	0.82	0.01
Phi_2_pr	0.75	0.80	0.84	0.86	0.87	0.73	0.78	0.82	0.84	0.86	**0.66**	0.72	0.77	0.80	0.81	0.02

Table A2 Results for FPFR on Data 2

	Normal					Skewed (c=0.5)					Skewed (c=1)					
	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6	Mean SD
Perfect Agree	0.79	0.69	**0.58**	0.50	0.44	0.77	0.68	**0.58**	0.50	0.44	0.74	0.67	**0.57**	0.49	0.43	0.06
Kappa	**0.57**	0.47	0.40	0.33	0.28	0.53	0.44	0.38	0.32	0.27	0.45	0.38	0.33	0.28	0.24	0.08
Linear Kappa	**0.57**	0.54	0.54	0.54	0.54	0.53	0.51	0.51	0.51	0.51	0.45	0.44	0.45	0.45	0.45	0.07
Quadratic Kappa	**0.57**	**0.64**	0.69	0.72	0.74	0.53	**0.60**	0.66	0.69	0.71	0.45	0.53	**0.59**	**0.63**	**0.65**	0.06
Polychoric	0.82	0.83	0.83	0.83	0.83	0.79	0.80	0.80	0.80	0.80	0.71	0.74	0.74	0.74	0.74	0.02
Pearson	**0.59**	0.67	0.73	0.76	0.78	0.55	**0.63**	0.70	0.73	0.75	0.47	0.56	**0.63**	0.67	0.70	0.03
Spearman	**0.59**	0.67	0.72	0.75	0.77	0.55	**0.63**	0.69	0.72	0.73	0.47	0.55	**0.61**	**0.63**	**0.65**	0.03
Intraclass	**0.57**	**0.64**	0.69	0.72	0.74	0.53	**0.60**	0.66	0.69	0.71	0.45	0.53	**0.59**	**0.63**	**0.65**	0.06
AC1	**0.60**	0.55	0.46	0.39	0.33	**0.56**	0.55	0.46	0.39	0.33	0.50	**0.56**	0.46	0.39	0.33	0.09
AC2	**0.60**	0.80	0.85	0.87	0.87	**0.56**	0.81	0.85	0.87	0.88	0.50	0.83	0.86	0.88	0.89	0.06
Gen_1_pir	0.39	0.48	0.53	0.56	**0.57**	0.34	0.43	0.48	0.51	0.52	0.21	0.31	0.36	0.39	0.41	0.03
Phi_1_pir	0.37	0.45	0.49	0.51	0.52	0.31	0.39	0.44	0.46	0.47	0.20	0.28	0.32	0.35	0.36	0.05
Gen_2_pir	0.49	**0.58**	**0.62**	**0.63**	**0.64**	0.43	0.52	**0.57**	**0.58**	**0.60**	0.29	0.40	0.44	0.47	0.48	0.03
Phi_2_pir	0.47	0.54	**0.57**	**0.59**	**0.60**	0.41	0.48	0.52	0.54	0.55	0.26	0.36	0.40	0.42	0.43	0.05
Gen_1_pr	**0.58**	0.67	0.73	0.76	0.78	0.55	**0.63**	0.70	0.73	0.75	0.47	0.56	**0.63**	0.67	0.70	0.03
Phi_1_pr	**0.57**	**0.64**	0.69	0.72	0.74	0.53	**0.60**	0.66	0.69	0.71	0.45	0.53	**0.59**	**0.63**	**0.65**	0.06
Gen_2_pr	0.74	0.80	0.84	0.86	0.87	0.71	0.77	0.82	0.85	0.86	**0.64**	0.71	0.77	0.80	0.82	0.02
Phi_2_pr	0.72	0.78	0.82	0.84	0.85	0.69	0.75	0.80	0.82	0.83	**0.62**	0.69	0.74	0.77	0.79	0.05

References

Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman & Hall/CRC.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27*(1), 3–23.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework*. (College Board Report No. 99-3; GRE Board Research Report No. 96-12R; ETS RR No. 99-3).

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Brenner, H., & Kliebsch, U. (1996). Dependence of weighed kappa coefficients on the number of categories. *Epidemiology, 7,* 199–202.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213–220.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61,* 29–48.

Gwet, K. L. (2010). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (2nd ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Report MS-31, RR-05-14). http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2005.tb01991.x/epdf.

Quarfoot, D., & Levine, R. A. (2016). How robust are multi-rater inter-rater reliability indices to changes in frequency distribution? *The American Statistician, 70*(4), 373–384.

Reardon, S. F., & Ho, A. D. (2014). Practical issues in estimating achievement gaps from coarsened data. https://cepa.stanford.edu/sites/default/files/reardon%20ho%20practical%20gap%20estimation%2025feb2014.pdf.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement, 64*(2), 243–253.

Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika, 81* (2), 399–410.

Warrens, M. J. (2012). Some paradoxical results for the quadratically weighted kappa. *Psychometrika, 77,* 315–323.

Yang, J., & Chinchilli, V. M. (2011). Fixed-effects modeling of Cohen's weighted kappa for bivariate multinomial data. *Computational Statistics & Data Analysis, 55,* 1061–1070.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association, 36*(1), 419–480.

How to Select the Bandwidth in Kernel Equating—An Evaluation of Five Different Methods

Gabriel Wallin, Jenny Häggström and Marie Wiberg

Abstract When using kernel equating to equate two test forms, a bandwidth needs to be selected. The bandwidth parameter determines the smoothness of the continuized score distributions and has been shown to have a large effect on the kernel density estimate. There are a number of suggested criteria for selecting the bandwidth, and currently four of them have been implemented in kernel equating. In this paper, all four of the existing bandwidth selectors suggested for kernel equating are evaluated and compared against each other using real test data together with a new criterion that implements leave-one-out cross-validation. Although the bandwidth methods generally were similar in terms of equated scores, there were potentially important differences in the upper part of the score scale where critical admission decisions are typically made.

Keywords Kernel equating · Continuization · Bandwidth selection
Cross-validation

1 Introduction

Observed-score test equating is the statistical procedure of adjusting test scores from different administrations to facilitate fair comparisons between examinees (González and Wiberg 2017). Kernel equating (KE; von Davier et al. 2004) is one

G. Wallin (✉) · J. Häggström · M. Wiberg
Department of Statistics, Umeå University, USBE, 90187 Umeå, Sweden
e-mail: gabriel.wallin@umu.se

J. Häggström
e-mail: jenny.haggstrom@umu.se

M. Wiberg
e-mail: marie.wiberg@umu.se

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_8

of the more recent equating frameworks, and offers a unified approach to the equating procedure. KE uses the equipercentile transformation to equate test scores (Braun and Holland 1982), which is based on the percentiles of the score distributions. However, if the equipercentile transformation is to be used in practice, the score distributions need to be continuous and monotonically increasing. This is generally not the case since test scores most often are discrete. KE utilizes kernel smoothers to solve this problem. There are a number of kernel functions that can be used, but regardless of the choice, a smoothing parameter always needs to be selected. This parameter, called the bandwidth, has been shown to be more important than the choice of kernel for density estimation (Wasserman 2006), but there is a lack of research on the impact of the bandwidth on the equipercentile transformation.

At the moment, there are four different methods suggested for bandwidth selection in KE, including the penalty method (von Davier et al. 2004), the double smoothing method (DS; Häggström and Wiberg 2014), the cross-validation method (Liang and von Davier 2014), and Silverman's rule of thumb method (SRT; Andersson and von Davier 2014). Since this paper introduces another bandwidth selection method that is based on cross-validation, the cross-validation method will be referred to as the likelihood method to make them easier to distinguish. The KE estimator that uses the penalty method (von Davier et al. 2004) was the first to be proposed within KE and has been the reference of comparison for the other methods. However, each study have used different evaluation criteria. Häggström and Wiberg (2014) evaluated the performance of DS by comparing the first two moments of the equated scores with those of the test scores from the old test form. Liang and von Davier (2014) evaluated the likelihood method in terms of bias and variance of the estimated kernel density function, and Andersson and von Davier (2014) compared the relative performance of SRT to the penalty method in terms of the difference in equated scores to the percentile rank method (Angoff 1971). For every study, only small differences from using the penalty method have been found. Moreover, some of the previous studies have varied the score distributions, others have varied the data collection design, and yet others have varied the number of examinees. Thus there is no possibility to draw conclusions about the performance of each method in comparison with the other methods based on the existing studies on bandwidth selection.

This study aimed to compare all existing bandwidth selection methods within KE in terms of equated scores, percent relative error (PRE; von Davier et al. 2004), standard error of equating (SEE; von Davier et al. 2004) and standard error of equating difference (SEED; von Davier et al. 2004) using real data from a standardized test. Furthermore, a new method for selecting the bandwidth is suggested that utilizes the leave-one-out cross-validation (LCV; Stone 1974) technique that is common in kernel density estimation.

2 Kernel Equating

KE is an equating framework that explicitly defines every step of the equating procedure. It comprises the following five steps: (1) Presmoothing, (2) Estimation of score probabilities, (3) Continuization, (4) Equating, and (5) Calculation of evaluation measures (von Davier et al. 2004; González and Wiberg 2017).

Consider two test forms X and Y, where the task is to equate the former to the latter. The scores generated from these test forms are denoted X and Y, respectively, and are considered to be random variables with observed values $x_1, \ldots, x_J$ and $y_1, \ldots, y_K$, respectively. By letting the cumulative distribution functions (CDFs) of X and Y be denoted $F_X(\cdot)$ and $G_Y(\cdot)$, respectively, and assuming they are continuous functions, an equivalent score y to a score x can be found using the equipercentile transformation:

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

However, the CDFs of X and Y are rarely continuous functions, and test scores are for the most part discrete (e.g. the number of correctly answered items). This means that for most score values it will not be true that $F_X(x) = u = G_Y(y)$ for $u \in [0, 1]$, and thus there will be a large set of score values on test form X for which Eq. 1 will not give unique score equivalents on test form Y. Different solutions to this problem have been suggested, where the percentile rank method (Angoff 1971), which uses linear interpolation, has been a common choice. More recently, the use of kernel smoothing techniques have been suggested to address this problem, which, in comparison to the percentile rank method, are not limited by the ranges of the score scales when equating two test forms. It is common practice in KE to approximate the score CDFs using a Gaussian kernel at the same time continuizes the CDFs and preserves the first two moments of the score random variables. For this purpose, let the variance of X be denoted by σ_X^2, let the bandwidth be denoted by h_X, $Z \sim N(0, 1)$, let $r_j = \Pr(X = x_j|T)$ for the target population T, and let $\mu_X = \sum_j x_j r_j$. In terms of the test score X, KE using a Gaussian kernel replaces X with the random variable $X(h_X) = a_X(X + h_X Z) + (1 - a_X)\mu_X$, where $a_X = \sqrt{\sigma_X^2/(h_X^2 + \sigma_X^2)}$. The CDF of $X(h_X)$ is given by

$$F_{h_X}(x; \mathbf{r}) = \Pr(X(h_X) \le x) = \sum_j r_j \Phi\left(\frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X}\right), \tag{2}$$

where $\mathbf{r} = (r_1, \ldots, r_J)^{\mathrm{T}}$ and $\Phi(z)$ is the standard normal distribution function. Letting $\mathbf{s} = (s_1, \ldots, s_K)^{\mathrm{T}}$ and $s_k = \Pr(Y = y_k|T)$, the corresponding CDF of $Y(h_Y)$, the continuous random variable replacing Y, is denoted by $G_{h_Y}(y; \mathbf{s})$. By letting $\hat{F}_{h_X}(x) = F_{h_X}(x; \hat{\mathbf{r}})$ and $G_{h_Y}(y) = G_{h_Y}(y; \hat{\mathbf{s}})$, the equipercentile transformation can be estimated by

$$\hat{\varphi}_Y(x) = \hat{G}_{h_Y}^{-1}\left(\hat{F}_{h_X}(x)\right). \tag{3}$$

The SEE is commonly used to evaluate the KE estimator $\hat{\varphi}_Y(x)$, and constitutes of three parts; the Jacobian of $\hat{\varphi}_Y(x)$, the Jacobian of the design function, which is the function that maps the estimated score distributions into the score probabilities of the target population, and the matrix **C** from which the covariance of the score distributions is formed. Denoting these three parts by $\hat{\mathbf{J}}_{\varphi_Y}$, $\hat{\mathbf{J}}_{\mathrm{DF}}$ and **C**, respectively, the SEE is given by

$$\mathrm{SEE}_Y(x) = \left\|\hat{\mathbf{J}}_{\varphi_Y}\hat{\mathbf{J}}_{\mathrm{DF}}\mathbf{C}\right\|.$$

If the goal is to compare two equating transformations by calculating $\varphi_Y(x) - \varphi_Y^*(x)$, where $\varphi_Y^*(x)$ differs from $\varphi_Y(x)$ only by the bandwidths, the SEED can be calculated as

$$\mathrm{SEED}_Y(\mathrm{x}) \ = \sqrt{\mathrm{Var}(\varphi_Y(x) - \varphi_Y^*(x))} = \left\|\mathbf{J}_{\varphi}\mathbf{J}_{\mathrm{DF}}\mathbf{C} - \mathbf{J}_{\varphi^*}\mathbf{J}_{\mathrm{DF}}\mathbf{C}\right\|.$$

3 Bandwidth Selection in Kernel Equating

In this section, the four previously proposed bandwidth selection methods will be briefly described together with a new proposed method. This paper makes the restriction to only compare these bandwidth selection methods, and excludes e.g. adaptive kernels which González and von Davier (2017) studied in the context of KE, and fixed values of the bandwidth that could be motivated from the goal to, for example, approximate a linear equating.

The Penalty Method

Let $\hat{f}_{h_X}(x) = \hat{F}_{h_X}^{'}$ denote the kernel density estimate yielded by differentiating $\hat{F}_{h_X}(x)$. The penalty method selects as bandwidth the value of h_X such that

$$\mathrm{PEN}(h_X) = \ \sum_j (\hat{r}_j - \hat{f}_{h_X}(x_j))^2 \ + \kappa \cdot \ \sum_j A_j \tag{4}$$

is minimized, where $A_j = 1$ if $[(\hat{f}_{h_X}^{'}(x_j - w) > 0) \cap (\hat{f}_{h_X}^{'}(x_j + w) < 0)]$ or $[(\hat{f}_{h_X}^{'}(x_j - w) < 0) \cap (\hat{f}_{h_X}^{'}(x_j + w) > 0)]$, and $A_j = 0$ otherwise (Lee and von Davier 2011; von Davier 2013). The term κ is a weight usually set to either 0 or 1 depending on if A_j is to be used or not. The term w is manually selected and regulates the interval around x_j for which $A_j = 1$. It is most often set to 0.25 (see e.g. von Davier et al. 2004; Häggström and Wiberg 2014; Andersson and von Davier 2014).

Silverman’s Rule of Thumb Method

A theoretically optimal bandwidth minimizes the asymptotic mean integrated squared error (Jones et al. 1996; Silverman 1986). If n_X denotes the sample size and X is normally distributed, the optimal bandwidth using SRT is given by $h_X \approx 1.06\sigma_X n_X^{-1/5}$ (Scott 1992). Motivated from the characteristics of test score data, Andersson and von Davier (2014) suggested an adjusted version of h_X, given by

$$\mathrm{SRT}(h_X) = \frac{9\sigma_X}{\sqrt{100 n_X^{2/5} - 81}}.$$

The Double Smoothing Method

Let g_X denote a large, pilot bandwidth and $\phi(z)$ denote the standard normal density function. In the first step of DS, a very smooth estimate of the kernel density of $X(h_X)$ is calculated as

$$\hat{f}_{g_X}(x) = \sum_{j=1}^{J} \hat{r}_j \phi\left(\frac{x - \hat{a}_X^{g_X} x_j - (1 - \hat{a}_X^{g_X})\hat{\mu}_{XT}}{g_X \hat{a}_X^{g_X}}\right) \frac{1}{g_X \hat{a}_X^{g_X}} \text{with}\, \hat{a}_X^{g_X} = \sqrt{\hat{\sigma}_{XT}^2 / (\hat{\sigma}_{XT}^2 + g_X^2)}$$

at each score value and at the values that lie in the middle of each score. To get a first DS estimate of f_{h_X}, $\hat{f}_{g_X}(x)$ is used instead of the score probabilities at each score value, which gives

$$\hat{f}^*_{h_X}(x) = \sum_{j=1}^{J} \hat{f}_{g_X}(x_j) \phi\left(\frac{x - \hat{a}_X x_j - (1 - \hat{a}_X)\hat{\mu}_{XT}}{h_X \hat{a}_X}\right) \frac{1}{h_X \hat{a}_X}.$$

This first step prevents undersmoothing the estimated score distribution. The second step of DS makes sure that the estimated distribution tracks the shape of the relative frequency distribution. This means minimizing the squared difference between the estimated score probabilities and the estimated score distribution. The DS method to select the bandwidth h_X can be compactly written as

$$\mathrm{DS}(h_X) = \sum_{l=1}^{2J-1} (\hat{r}_l^* - \hat{f}^*_{h_X}(x_l^*))^2, \text{where}\, \hat{r}_l^* = \begin{cases} \hat{r}_{\frac{l+1}{2}}, & \text{if}\, l \,\text{is odd} \\ \hat{f}^*_{h_X}(x_l^*), & \text{if}\, l \,\text{is even.} \end{cases}$$

The Likelihood Method

The likelihood method starts by splitting the sample for both of the test forms into two subsamples. In terms of the X test scores, f_{h_X} is estimated using $\hat{F}'_{h_X}$ for a set of bandwidths ranging from 0.01 to 5 with increments of 0.01. For each density estimate and each score value, the observed density value is plugged in as the intensity parameter in a Poisson likelihood function, where the frequencies are taken from the other subsample. For a score frequency of k for test score x_j, and a size of the first subsample of n_1, this maximization can be expressed as

$$\text{Likelihood}(h_X) = \max_h L(k;\hat{f}_{h_X}) = \max_h \prod_{j=0}^{J} \frac{e^{-n_1\hat{f}_{h_X}(x_j)}\left(n_1\hat{f}_{h_X}(x_j)\right)^k}{k!}.$$

The choice of bandwidth that maximizes the likelihood function is stored, and this procedure is repeated 1,000 times. The median of the 1,000 stored bandwidths is finally selected and used in the respective estimations of F_{h_X} and G_{h_Y}.

The Leave-One-Out Cross-Validation Method

There are a few issues with the existing bandwidth selection methods in KE. The penalty method, when adding the penalty function, is not a differentiable function. SRT relies on normally distributed data, and the likelihood method is computationally infeasible because it maximizes a target function with 1,000 repetitions. Other plug-in estimators that have been suggested within the general field of kernel density estimation, and that are inherently computationally fast, suffer from requiring an estimation of the second derivative of the density or requiring that the density to be estimated is very smooth (Wasserman 2006). A good trade-off could be to use LCV, which has a longstanding history in kernel density estimation, is theoretically justified (see e.g. Stone 1984), and is not as computationally expensive as the likelihood method.

Let

$$\hat{f}_{h_x}^{-j}(x_j) = \sum_{\substack{l=1 \\ l\neq j}}^{J} \hat{r}_l \phi\left(\frac{x_j - \hat{a}_x x_l - (1-\hat{a}_x)\hat{\mu}_{xT}}{h_x\hat{a}_x}\right)\frac{1}{h_x\hat{a}_x}$$

be the kernel density estimate of $f_{h_X}(x_j)$ with (x_j, r_j) left out of the estimation, i.e., $f_{h_X}(x_j)$ is estimated using LCV. In this fashion, overfitting is effectively prevented. Using this estimate, the novel approach suggested in this paper is to adjust the penalty method by replacing $\hat{f}_{h_X}(x_j)$ with $\hat{f}_{h_X}^{-j}(x_j)$ and leaving out the penalty function A_j.

4 Empirical Study

When applying for higher education in Sweden, there is the possibility to either use the grades from high school or the results from a standardized test, the Swedish Scholastic Assessment Test (SweSAT). The latter is given as a paper and pencil test in the spring and fall each year. It contains a quantitative and a verbal section, both with 80 items that are equated separately. We have equated the quantitative section of two consecutive administrations. The sample from the fall administration consisted of 2,826 examinees, and the sample from the spring administration consisted of 2,783 examinees.

Based on the equating process used in practice for the SweSAT, we adopted a non-equivalent groups with anchor test design. The results presented here used post-stratification equating with a weight of 0.5 for the synthetic population. Both samples were presmoothed using log-linear models and with the AIC as the evaluation measure. The bandwidth methods were evaluated and compared using the difference in equated scores, SEED, SEE, and PRE. By letting $\mu_p(Y) = \sum_k (y_k)^p s_k$ and $\hat{\mu}_Y = \sum_j \left(\hat{\varphi}_Y(x_j)\right)^p r_j$, the PRE is calculated as $\mathrm{PRE}(p) = 100(\hat{\mu}_Y - \mu_p(Y))/\mu_p(Y)$. All analyses were performed in R, and the R-package (R Core Team 2017) kequate (Andersson et al. 2013) was used for the implementation of the penalty, SRT, and DS methods.

5 Results

The bandwidths obtained from each method are displayed in the upper section of Table 1. It is a big span between the largest and smallest bandwidth, where the smallest was selected by LCV and the largest by the SRT method. The bandwidths selected using the DS and penalty methods show great similarity.

The PRE for the first five moments of the five different KE transformations is displayed in the lower part of Table 1. None of the PREs for any of the moments are larger than 0.027%, meaning that the estimated distribution of the equated scores, regardless of bandwidth method, comes close to the distribution of the scores of the old test form. The KE transformations using the penalty and DS methods had the overall best performance in terms of PRE.

Figure 1 shows the differences of the SRT, DS, likelihood, and LCV methods compared to the penalty method in terms of the equated scores. Confidence bands are added to each plot in Fig. 1 showing ±2 × SEED. Bounds of ±0.5 are also

Table 1 The bandwidths for the Penalty, SRT, DS, Likelihood and LCV method, and the PRE for the first five moments

Bandwidth	Penalty	SRT	DS	Likelihood	LCV
h_X	0.73	2.25	0.75	1.85	0.34
h_Y	0.71	2.42	0.74	1.19	0.34
Moment					
1	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.001	−0.003
3	0.000	0.000	0.000	0.004	−0.009
4	0.000	0.001	0.000	0.008	−0.017
5	0.001	0.005	0.001	0.011	−0.027

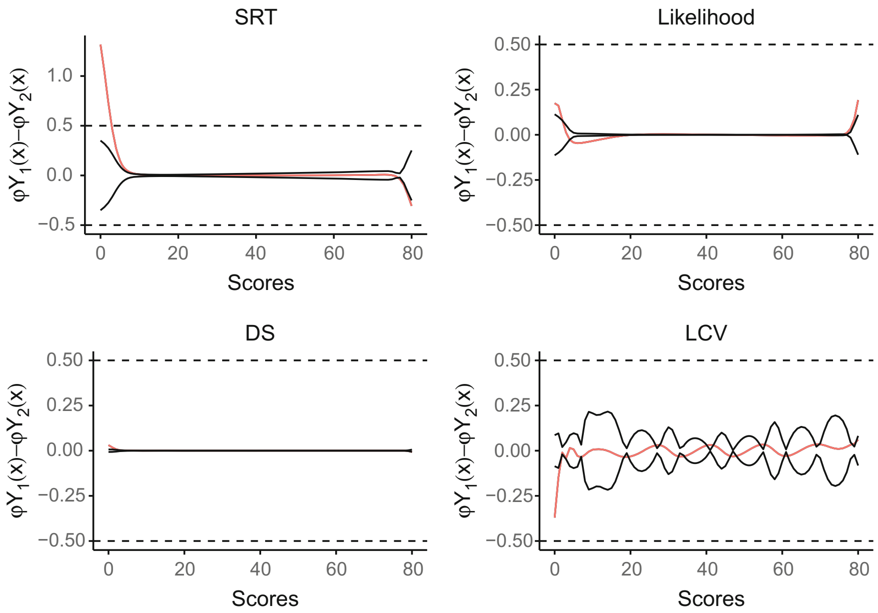


Fig. 1 The difference between the penalty method and the SRT, DS, Likelihood and LCV method, respectively, in terms of equated scores (red lines). The straight black lines are confidence bands of ±2 × SEED, and the dashed lines display the DTM

added to indicate the difference that matters (Dorans and Feigenbaum 1994). The differences in equated scores between each method and the penalty method are small for a large part of the score range, but every method have equated scores that fall outside of the confidence bands. Only the SRT method have scores that also fall outside the limits of the difference that matters.

The SEE was computed for each KE transformation and is presented in Fig. 2. All of the methods result in similar SEE, with a generally higher SEE for the low and high scores due to fewer observations in the tails. The peaks are the highest when using LCV.

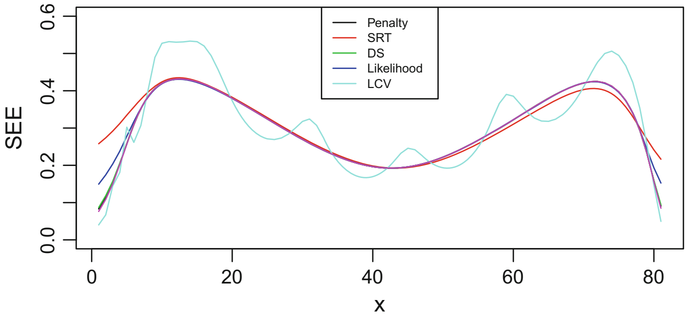


Fig. 2 The SEE for the KE transformation using the Penalty, DS, SRT, Likelihood and LCV method, respectively

6 Discussion

This paper has highlighted the continuization step of KE, and in particular the choice of bandwidth for the Gaussian kernel function used in the KE process. Every existing method have been evaluated and compared using real test data from the SweSAT. Furthermore, LCV for bandwidth selection has been implemented within the KE framework. The results indicate that LCV yields density estimates that are not as smooth as e.g. the penalty method, but that the resulting equated scores from using the two methods are very similar. The likelihood method also showed great resemblance to the penalty method in terms of equated scores, but it took a considerable time to compute. DS showed the greatest resemblance to the penalty method in terms of bandwidths, equated scores, SEE and PRE. Only the SRT deviated by more than 2 × SEED in the upper tail of the score scale, and all equated score differences stayed within the confidence bands for most of the score scale. Given the results of this study, there is a need to more rigorously investigate how the bandwidths affect the equating results and to determine if there is a possibility to identify certain test scenarios where each of the different bandwidth methods are particularly suitable. The SweSAT is a test with score distributions that are fairly symmetric and unimodal. We have also used a rather large number of examinees and items, so future research should try to vary these factors. This study has both confirmed and added to the results of previous studies by showing the small differences between the methods in terms of PRE and SEE. For most equated scores, there are only small differences between the methods that will not have any practical importance. However, the results have also shown important differences in the upper tail of the score scale where critical admission decisions are to be made. Thus the results have indicated that the choice of bandwidth is of importance for the equating transformation, and since there are five different methods available to select the bandwidth, it is critical to determine when each method is most appropriate.

Acknowledgements The research in this paper was funded by the Swedish Research Council grant: 2014-578.

References

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating using the R package kequate. *Journal of Statistical Software, 55*(6), 1–25.

Andersson, B., & von Davier, A. A. (2014). Improving the bandwidth selection in kernel equating. *Journal of Educational Measurement, 51,* 223–238.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. I. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

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). NewYork: Academic Press.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

González, J., & Wiberg, M. (2017). *Applying test equating methods using R*. Cham: Springer.

González, J., & von Davier, A. A. (2017). An illustration of the Epanechnikov and adaptive continuization methods in kernel equating. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C Wang (Eds.), *Quantitative psychology—81st annual meeting of the psychometric society, Asheville, North Carolina, 2016*. New York: Springer.

Häggström, J., & Wiberg, M. (2014). Optimal bandwidth selection in kernel equating. *Journal of Educational Measurement, 51,* 201–211.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association, 91,* 401–407.

Lee, Y.-H., & von Davier, A. A. (2011). Equating through alternative kernels. In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking* (pp. 159–173). New York, NY: Springer.

Liang, T., & von Davier, A. A. (2014). Cross validation: An alternative bandwidth selection method in kernel equating. *Applied Psychological Measurement, 38,* 281–295.

R Core Team. (2017). *R: A language and Environment for Statistical Computing*. R Foundation for statistical computing. Vienna, Austria. http://www.R-project.org/.

Scott, D. (1992). *Multivariate density estimation*. New York, NY: Wiley.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society Series B, 36,* 111–147.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics, 12*, 1285–1297

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika, 78*, 605–623.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.

Wasserman, L. (2006). *All of nonparametric statistics*. New York: Springer.

Evaluating Equating Transformations from Different Frameworks

Waldir Leôncio and Marie Wiberg

Abstract Test equating is used to ensure that test scores from different test forms can be used interchangeably. This paper aims to compare the statistical and computational properties from three equating frameworks: item response theory observed-score equating (IRTOSE), kernel equating and kernel IRTOSE. The real data applications suggest that IRT-based frameworks tend to provide more stable and accurate results than kernel equating. Nonetheless, kernel equating can provide satisfactory results if we can find a good model for the data, while also being much faster than the IRT-based frameworks. Our general recommendation is to try all methods and examine how much the equated scores change, always ensuring that the assumptions are met and that a good model for the data can be found.

Keywords Test equating · Item response theory · Kernel equating
Observed-score equating

1 Introduction

Test equating is used to ensure that scores from different test forms are comparable and can be used interchangeably (Kolen and Brennan 2014; González and Wiberg 2017). For instance, if we want to transform the test scores x from test form X to the scale of the test scores y from test form Y, we can define the general transformation between the cumulative distribution functions $F_{\mathrm{X}}(x)$ and $F_{\mathrm{Y}}(y)$ as

$$\varphi(x) = F_{\mathrm{Y}}^{-1}(F_{\mathrm{X}}(x)) \,. \tag{1}$$

W. Leôncio (✉)
Dipartimento di Scienze Statistiche, University of Padua, Via C. Battisti, 241,
35121 Padua, Italy
e-mail: waldir.leoncionetto@phd.unipd.it

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

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_9

This transformation is referred to as equipercentile equating, and all equating transformations can be considered special cases of this equation (Braun and Holland 1982). Many equating methods have been developed depending on the data collection design and the assumptions made, such as traditional equating methods (Kolen and Brennan 2014), observed-score kernel equating methods (von Davier et al. 2004), item response theory (IRT) methods (Lord 1980), local equating methods (van der Linden 2011), and mixtures of them as for example local kernel IRT observed-score equating (Wiberg et al. 2014). Equating methods with similar characteristics can be grouped into frameworks which are not necessarily mutually exclusive, such as the kernel equating framework (KE, von Davier et al. 2004) and the IRT observed-score equating framework (IRTOSE, Lord 1980), both of which can be further grouped into the observed-score equating framework.

The number of approaches to equating can be overwhelming, so it is important to have tools to evaluate the underlying transformation. One problem with many of the current evaluation tools is that they were created to compare methods within a specific framework and they are *equating-specific*, meaning they target specific parts of the equating process and thus aim to evaluate the equating based on different, isolated aspects. A different approach was proposed by Wiberg and González (2016), who showed how we could use a statistical approach to evaluate equating methods *within* a framework. Their work specifically used KE, although they also discussed how it could be applied within IRTOSE and local equating. A problem left for further research was how to evaluate equating transformations from different frameworks. This paper aims to explore a method that fills this gap. The focus will be on IRTOSE, KE and IRT observed-score kernel equating (IRTKE, Andersson and Wiberg 2017), and the evaluation will be performed with real data.

One challenge when comparing equating transformations from different frameworks is to set up a fair comparison which does not favor any particular method. Equating is a process where several decisions need to be made which could dramatically change the results. Ideally, simulations should be conducted on different scenarios that cover most approaches, like what Wiberg and González (2016) did within the KE framework.

The rest of this paper is structured as follows. The next section contains brief descriptions of the equating frameworks used. The third section describes the evaluation criteria used. The fourth section gives details on the real data study, and the results are given in the fifth section. The final section contains some concluding remarks as well as general recommendations.

2 Methodology

To equate test results from two independent populations of examinees P and Q, we use samples of size I_P and I_Q from these populations. Let test form X contain J_X items, test form Y contain J_Y items. Furthermore, let A be an anchor test form containing J_A items. In this study we consider the non-equivalent groups with anchor test design (NEAT, von Davier et al. 2004, Sect. 2.4), which means test $X^+ = \{X, A\}$ is administered to the sample of population P and test $Y^+ = \{Y, A\}$ is administered to the sample of population Q. In the following subsections, the equating frameworks used in this paper are briefly described.

2.1 IRT Observed-Score Equating (IRTOSE)

For each test form we observe dichotomous response data in a matrix with I rows and J columns. To perform IRTOSE, we start by fitting an IRT model to this data. A common model is the three-parameter logistic IRT model (3-PL) for binary data, which models the probability examinee i with ability $\theta \in \mathbb{R}$ correctly answering item j as

$$p_j = c_j + \frac{1 - c_j}{1 + \exp[-a_j(\theta - b_j)]} , \tag{2}$$

where a_j represents the discrimination, b_j the difficulty and c_j the pseudo-guessing probability of item j. These probabilities are then placed into a compound binomial model to generate the probability distribution of the number-correct scores for a given θ. Typically, an algorithm described in Lord and Wingersky (1984) is used for this, although other alternatives exist (see González et al. 2016). Then, the score distributions are cumulated over the population of θ to create aggregated distributions of the total scores for tests X^+ and Y^+ (Kolen and Brennan 2014, Sect. 6.6). As these probability distributions are discrete, they need to be made continuous before their percentiles can be compared. This is done by linear interpolation, after which equipercentile equating can be conducted.

2.2 Kernel Equating (KE)

KE is comprised of five steps. The first one, pre-smoothing, prepares the data for the estimation of score probabilities, which is the second step. Pre-smoothing is typically done by fitting a log-linear model to the observed scores, although the raw data can also be directly used in the second step. Then, the discrete scores distributions are made continuous using a kernel function (e.g., Gaussian, uniform, logistic). In the fourth step, equipercentile equating is performed. On a NEAT design, we can choose between chained equating (CE) and post-stratification equating (PSE) to equate the tests. In the final step, accuracy measures such as the standard error of equating (SEE) can be obtained.

2.3 IRT Observed-Score Kernel Equating (IRTKE)

IRTKE (Andersson and Wiberg 2017) also comprises the five steps described for KE, although it uses score probabilities derived from an IRT model as input for the kernel continuization. Alternatively, IRTKE can be seen as an IRTOSE method where a kernel function is used for the continuization instead of linear interpolation.

2.4 Chosen Methods

To perform equating in a particular framework, several choices must be made. A 3-PL model was fitted to the data when performing IRTOSE and IRTKE, with the Haebara (1980) method used to rescale the model parameters.

For KE, several log-linear models were considered, with the best fit being chosen by a stepwise procedure and the principle of parsimony. Both KE and IRTKE use PSE and a logistic kernel for continuization. The logistic kernel was chosen because its heavier tails (when compared with the more common Gaussian kernel) are a better match to the distribution of the observed data.

All procedures were performed in R (R Core Team 2017), with the `ltm` package (Rizopoulos 2006) being used to fit IRT models to the data and the `glm` function handling the log-linear models. IRTOSE was performed using `equateIRT` (Battauz 2015); KE and IRTKE were carried out using `kequate` (Andersson et al. 2013).

3 Evaluating Equating Transformations

The most common way to compare the performance of two equating transformations *within* a framework is through equating-specific evaluation measures (Wiberg and González 2016). In traditional equating methods, a commonly used measure is the "difference that matters" (DTM), originally defined as the difference between equated scores and scale scores that are larger than half of a reported score unit (Dorans and Feigenbaum 1994). In KE, a popular measure is the percent relative error (PRE), which compares the moments of the score distribution on the reference form to the score distributions of all the equated forms (Jiang et al. 2012; von Davier et al. 2004). Even though PRE was specifically developed for KE, it can be adapted to methods that use linear interpolation (Jiang et al. 2012).

Since equating transformations can be viewed as statistical estimators (Wiberg and González 2016), measures such as bias, mean squared error (MSE) and root mean squared error (RMSE) can also be calculated. Given a score x, its true equivalent score on test Y, $\varphi(x)$, and the estimated equivalent score $\hat{\varphi}(x)$, the bias and RMSE are defined as

$$\text{bias}[\hat{\varphi}(x)] = E\left[\hat{\varphi}(x) - \varphi(x)\right] \tag{3}$$

and

$$\text{RMSE}[\hat{\varphi}(x)] = \sqrt{E\left\{[\hat{\varphi}(x) - \varphi(x)]^2\right\}}. \tag{4}$$

A good estimator is expected to have low bias and RMSE, which means we could evaluate equating transformations from different frameworks simply by comparing their values on these measures. There are two issues with this approach: the true equated scores $\varphi(x)$ can vary from framework to framework, and $\varphi(x)$ is not even obtainable with real data. The former does not constitute a critical problem, since it only adds some algebraic complexity to the procedure, but the latter can be a major impediment to the application of this approach to real data unless a satisfiable proxy for $\varphi(x)$ can be found.

One way to circumvent the absence of $\varphi(x)$ is by defining an equating transformation to be the true and comparing the others to this benchmark. This is what Wiberg and González (2016) did within KE.

Another possibility, implemented in Lord (1977), is to let test forms X and Y actually be the same test, whilst having the computer procedure handle them as different. Under this scheme, called the "circular paradigm", we theoretically have $\hat{\varphi}(y) = \hat{\varphi}(x) = \varphi(x) = x$, meaning we should expect no difference between the equated scores and the raw ones. However, when using real data, the calculation of measures such as those in (3) and (4) under this simple approach presents one caveat: real-life test administrations rarely (if ever) have the same group of students take the same test multiple times, making this is a one-sample experiment. Under these circumstances, the equating estimates will contain an unknown amount of sampling error on top of any bias that particular equating framework already has. Nonetheless, given the overall homogeneity of the test subjects as well as the small variability between the test forms in the real data applications under study, we don't expect sampling error to be a critical issue. Hence, the circular paradigm seems to be an adequate approach for this paper.

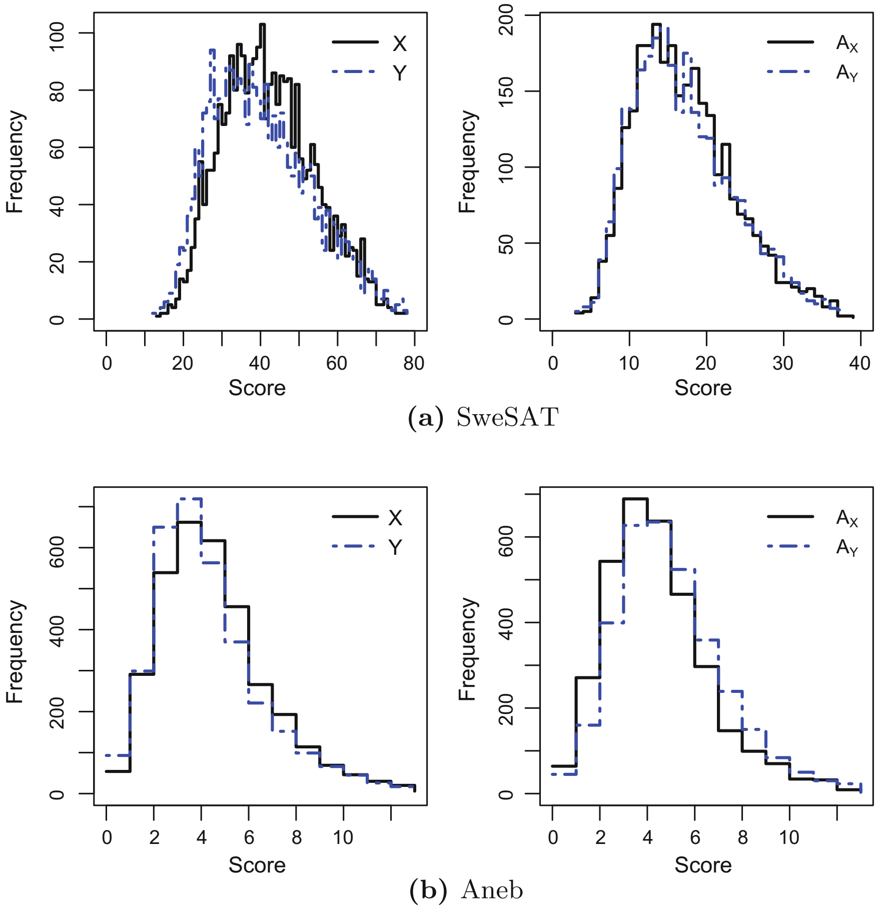
As pointed out by Harris and Crouse (1993), equating a test to itself solves the problem of having a true, known criteria. However, since the observed differences will be due not only to bias, but also to sample variability, we will refrain from referring to the calculated statistics as "bias" and will instead use the more comprehensive term "error". Moreover, considering that both real data cases represent only one sample, the RMSE would equal the absolute error. For this reason, only the errors were presented in this study.

As the different methods can take significantly different times to be executed, it is also interesting to compare the runtimes of the different equating procedures. This can be useful in estimating the feasibility of a certain method in circumstances where time is a major constraint.

4 Real Data Study

The real data application is composed of two test forms of the Swedish Scholastic Assessment Test (SweSAT) and the Brazilian National Assessment of Basic Education (Aneb). These tests provide adequate examples for our study, since they have quite different score distributions and number of items.

The SweSAT is a large-scale college admissions test given twice a year in Sweden. It is a multiple-choice test consisting of a quantitative and a verbal section with 80 items each. The two sections are equated separately using anchor tests containing 40 items each. In this study we used the quantitative section from the autumn 2014 and the spring 2015 administrations.



(a) SweSAT

(b) Aneb

Fig. 1 Distribution of the observed test scores

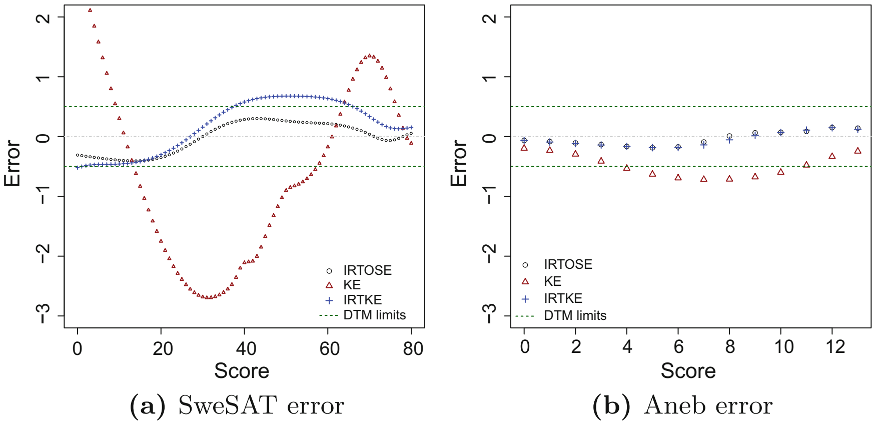
Aneb is a biennial large-scale assessment of the Brazilian school system. It is composed of a Math and a Language section. This study used the Math tests given to 12th graders of the 2015 administration and equated two booklets containing 13 unique items each and another 13 common items.

5 Results

The distribution of the observed scores can be seen in Fig. 1. In general, the SweSAT can be perceived as relatively more symmetric, with Aneb clearly presenting positive skewness. As a matter of fact, the skewness of the test forms on the SweSAT ranges from 0.35 to 0.66, whereas the test forms on Aneb had values of skewness between 0.73 and 1.07. In the Brazilian exam, test forms X and Y seem to have similar means, with averages ranging from 3.81 to 4.49 across all forms. On the SweSAT, the plots suggest Y has a slightly lower average than X, which is corroborated by their respective score averages of 39.89 against 41.68. The average scores of the anchor tests of the Swedish exam were even closer, with $\overline{A_X} \approx 16.71$ and $\overline{A_Y} \approx 16.64$.

Information on equating quality can be seen in Fig. 2. The DTM limits were drawn at error points −0.5 and +0.5. The plots show similar patterns for IRTOSE and IRTKE, which are completely contained within the DTM band for Aneb and only slightly trespass it at some points around the middle of the score range for the SweSAT. The behavior of KE is definitely more unstable, with higher errors for Aneb and even more for the SweSAT.

Table 1 complements the results from Fig. 2, allowing a numerical overview of the errors. Some scores were omitted for brevity, but were still included in the average and standard deviation. Once again, IRTOSE and IRTKE generally perform better



(a) SweSAT error

(b) Aneb error

Fig. 2 Error per score

Table 1 Bias per score

Score	SweSAT			Score	Aneb		
	IRTOSE	KE	IRTKE		IRTOSE	KE	IRTKE
0	−0.31	2.73	−0.52	0	−0.06	−0.20	−0.07
10	−0.40	0.30	−0.46	1	−0.08	−0.23	−0.09
20	−0.35	−1.75	−0.30	3	−0.13	−0.41	−0.14
30	0.00	−2.68	0.16	5	−0.19	−0.63	−0.19
40	0.29	−2.11	0.58	7	−0.09	−0.72	−0.14
50	0.26	−0.90	0.68	9	0.06	−0.68	0.02
60	0.22	−0.17	0.64	11	0.09	−0.48	0.11
70	0.04	1.35	0.34	12	0.15	−0.34	0.15
80	0.05	−0.11	0.15	13	0.14	−0.25	0.12
Average	−0.02	−0.59	0.17	Average	−0.03	−0.48	−0.05
Std. dev.	0.26	1.51	0.44	Std. dev.	0.12	0.19	0.12

than KE in both scenarios. While no framework presents the least error for all scores, IRTOSE offers the lowest average error for both the SweSAT and Aneb.

On an Intel i5-760 CPU with 8 GB of RAM, the respective runtimes for IRTOSE, KE and IRTKE were approximately 252, 66 and 253 s for the SweSAT. For the smaller Aneb test, they clocked around 21, 2 and 20 s. These results are within the expected values, with KE being by far the fastest procedure, whereas IRTOSE and IRTKE take roughly the same time. These differences are mostly due to the fact that IRTOSE and IRTKE process answers at the item level, whereas KE only uses the total scores.

6 Conclusion

The results show an overall advantage of IRTOSE and IRTKE as far as error and DTM are concerned, with KE prevalent when speed is a priority. KE can also be the best option when there is evidence that IRT models do not fit the data, which might be the case for the highly-skewed Aneb data (complete goodness-of-fit analyses of the chosen models were not performed on the datasets). These results are in line with Meng (2012), who found IRTOSE to be more accurate than KE under some of the conditions explored in their work. On the other hand, their work has also found KE to be more stable than IRTOSE, which contrasts with our findings. These differences can be due to the particularities of the scenarios chosen on each work, and more studies are encouraged to further explore the issue.

The model choice for pre-smoothing can drastically change the results of KE, including the reduction of the errors we observed. The assumption that a fair comparison would be completely objective is what led us to allow a stepwise procedure to choose a model instead of hand-picking one. In practical cases, however, we recommend supplementing an automated analysis with a manual one. Such care is especially important in high-stakes admissions tests, where choosing a particular method could lead to the approval or rejection of a candidate.

Even though a more careful analysis including simulated data with several replications should be conducted to allow more confident conclusions, the results of this paper support the formulation of the hypothesis that test size and score skewness have little effect on the quality of IRTOSE and IRTKE. If confirmed, this could mean that test developers using IRT equating would benefit from focusing on improving the quality of items rather than their quantity. For KE, the presence of low-frequency scores provides an extra challenge to fitting a proper pre-smoothing model. In such regions, careful modeling could make a big difference in equating quality. For IRT models, danger relies on the presence of exceptionally difficult or easy items, which can make these models fail to fit the data.

The equated scores are a function of the chosen parameters, so different decisions could inadvertently favor one method over another. Further studies could help the discussion on the best way to compare equating transformations on different frameworks. Further research should contain a simulation study which allows deeper analysis of the results. Simulated data can be generated multiple times, which allows for the reliable calculation of evaluating measures. Studies involving real data with other characteristics (three or more test forms, internal anchor items) or other equating frameworks such as those mentioned in Sect. 1 would also be a meaningful contribution.

This paper expands the work of Wiberg and González (2016) by developing methods to compare equating transformations from different frameworks. We welcome more studies to develop methods for evaluating equating transformations, since the circular paradigm, although very useful at solving the problem of finding a reference equating transformation, has at least two issues pointed out by Harris and Crouse (1993): the equating results may depend on the chosen base form, and it could favor frameworks which use one or two moments over more complex frameworks. The latter is not a problem encountered in this paper but might be an issue in other circumstances. In any case, it could be interesting to see future work addressing any or both of these issues.

Acknowledgements The research in this article was funded by the Swedish Research Council grant 2014-578 and by the Fondazione Cassa di Risparmio di Padova e Rovigo.

References

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika*, *82*(1), 48–66. https://doi.org/10.1007/s11336-016-9528-7.

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.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (Vol. 1, pp. 9–49). New York: Academic Press.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. Technical issues related to the introduction of the new SAT and PSAT/NMSQT (pp. 91–122).

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.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144–149.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*(3), 195–240.

Jiang, Y., von Davier, A. A., & Chen, H. (2012). Evaluating equating results: Percent relative error for chained kernel equating. *Journal of Educational Measurement*, *49*(1), 39–58.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

van der Linden, W. J. (2011). Local observed-score equating. In A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 201–223). New York: Springer.

Lord, F. M. (1977). Practical applications of item response theory. *Journal of Educational Measurement*, *14*(2), 177–138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement*, *8*(4), 453–461.

Meng, Y. (2012). Comparison of kernel equating and item response theory equating methods. Dissertation submitted to the graduate school of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of doctor of education, University of Massachusetts Amherst.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. http://www.jstatsoft.org/v17/i05/.

Wiberg, M., & González, J. (2016). Statistical assessment of estimated transformations in observed-score equating. *Journal of Educational Measurement*, *53*(1), 106–125.

Wiberg, M., van der Linden, W. J., & von Davier, A. A. (2014). Local observed-score kernel equating. *Journal of Educational Measurement*, *51*, 57–74.

An Alternative View on the NEAT Design in Test Equating

Jorge González and Ernesto San Martín

Abstract Assuming a "synthetic population" and imposing strong assumption to estimate score distributions has been the traditional practice when performing equating under the nonequivalent groups with anchor tests design (NEAT). In this paper, we use the concept of partial identification of probability distributions to offer an alternative to this traditional practice in NEAT equating. Under this approach, the score probability distributions used to obtain the equating transformation are bounded on a region where they are identified by the data. The advantages of this approach are twofold: first, there is no need to define a synthetic population and, second, no particular assumptions are needed to obtain bounds for the score probability distributions that are used to build the equating transformation. The results show that the uncertainty about the score probability distributions, reflected on the width of the bounds, can be very large, and can thus have a big impact on equating.

Keywords Test equating · NEAT design · Partial identifiability · Ignorability condition

1 Introduction

Test equating is used to make scores from different test forms comparable. An equating transformation function is used to map the scores on one scale into their equivalents on the other. Before this score transformation takes place, it is necessary to control for test takers ability differences, and different data collection designs have been described in the equating literature for such purpose (von Davier et al. 2004, Chap. 2; Kolen and Brennan 2014, Sect. 1.4 and González and Wiberg 2017, Sect. 1.3.1). These equating designs differ in that either common persons or common items are

J. González (✉) · E. San Martín
Faculty of Mathematics, Pontificia Universidad Católica de Chile,
Av. Vicuña Mackenna, 4860 Macul, Santiago, Chile
e-mail: jorge.gonzalez@mat.uc.cl

E. San Martín
e-mail: esanmart@mat.uc.cl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_10

used to perform the score transformation. In this paper we will focus the attention on the nonequivalent groups with anchor test design (NEAT).

The NEAT design is widely used in test equating. Under this design, two groups of test takers are administered separate test forms with each test form containing a common subset of items. Because test takers from different populations are assigned only one test form, missing score data emerge by design rendering some of the score distributions unavailable. The equating literature has treated this problem from different perspectives all of them making different assumptions in order to estimate the missing score distributions. In this paper, we offer an alternative view that is free of these types of assumptions to obtain the score distributions under a NEAT design.

We first argue that, rather than viewing the problem as one of missing data, there is an inherent identifiability problem underlying the NEAT design. Then, we further argue that the typical assumptions on the equality of conditional distributions are nothing more that identifiability restrictions. Because these assumptions might be too strong, and, moreover, are not empirically testable, we offer an alternative that does not make use of any assumption and show that the non identified score distributions are actually partially identified, deriving bounds for them on the partially identified region.

The rest of this paper is organized as follows. We first briefly revisit the current view on the NEAT design, including the definition of synthetic population and the assumptions commonly made to estimate score distributions. Then we introduce our view on the NEAT design as an identifiability problem and derive bounds where the non identified score distributions are partially identified. An illustration using an hypothetical data example appearing in the equating literature is presented. The paper ends with final remarks and ideas for future work.

2 NEAT Equating: The Current and an Alternative View

2.1 Notation and Preliminaries

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be the random variables representing test scores from tests forms X, Y. As mentioned before, the equating function $\varphi : \mathcal{X} \mapsto \mathcal{Y}$ defined as $\varphi(x) = F_Y^{-1}(F_X(x))$ maps the scores on the scale $\mathcal{X}$ into their equivalents on the $\mathcal{Y}$ scale (González and Wiberg 2017). This definition is established for φ defined on a common population where the equating is to be performed (Braun and Holland 1982). Accordingly, the score cumulative distribution functions used to build the equating transformation, should also be defined on a common population that will be denoted as T.

When single groups (SG), equivalent groups (EG) or counter balanced groups (CB) equating designs are considered, defining φ on a common population does not constitute a problem as samples of test takers are in fact taken from the same population. However, this is not the case for the NEAT design where samples of test takers come from two different populations, called here P and Q. As a consequence, score distributions of X and Y are defined in both P and Q and we denote these distributions here as $F_{XP}(x)$, $F_{YP}(y)$, $F_{XQ}(x)$, and $F_{YQ}(x)$, respectively.

2.2 NEAT Equating: The Current View

To solve the problem of defining the equating transformation on a common population, the equating literature has resorted in what is called a *synthetic population* (Braun and Holland 1982). This definition conceptualizes a common population as a weighted combination of P and Q in the form

$$T = wP + (1 - w)Q, \tag{1}$$

where w is a weight such that $0 \leq w \leq 1$. Using this definition, the corresponding score distributions used to build the equating transformation are obtained as

$$\begin{aligned} F_{XT}(x) &= wF_{XP}(x) + (1 - w)F_{XQ}(x) \\ F_{YT}(y) &= wF_{YP}(y) + (1 - w)F_{YQ}(y). \end{aligned} \tag{2}$$

A typical representation of the NEAT equating design is shown in Table 1. From the table, it can be seen that because test takers in P are only administered test X and those in Q are only administered Y, the corresponding score distributions F_{XQ} and F_{YP} needed to obtain F_{XT} and F_{YT} in (2) are said to be *missing*. Additional assumptions are thus needed to estimate them, and here is where the anchor test, A, has played a fundamental role. Most commonly, it is assumed that the conditional score distributions of X and Y given A are the same in both population: $F_{XP}(x \mid a) = F_{XQ}(x \mid a)$ and $F_{YP}(y \mid a) = F_{YQ}(y \mid a)$, with $A \in \mathcal{A}$. Using these assumptions, and the fact that marginal distributions of A are indeed observed in both populations, the score distributions of X and Y in T are obtained by marginalizing the joint distributions over A. The obtained score distributions are then used to build $\varphi(x) = F_{YT}^{-1}(F_{XT}(x))$.

2.3 NEAT Equating: An Alternative View

Rather than facing missing score distribution, what happens in reality is that the sampling process underlying the NEAT design does not give information on F_{YP} and F_{XQ}, and thus the target score distributions $F_{XT}(x)$ and $F_{YT}(y)$ are not identified.

Table 1 Schematic representation of the NEAT design

Population	Sample	X	Y	A
P	1	✓		✓
Q	2		✓	✓

Note X and Y are test forms. A is an anchor test

Moreover, the assumptions on equality of conditional score distributions are actually identification restrictions.

To introduce these ideas better, let us briefly revisit the definition of identifiability. If θ is a parameter indexing a family of distributions $\{f(x \mid \theta) : \theta \in \Theta\}$, then θ is said to be identified if distinct values of it lead to distinct probability distributions (Casella and Berger 2002). Equivalently, if the probability distribution can be uniquely determined by θ, then θ is identified. If the probability distribution cannot be uniquely determined (i.e., the model is not identified), putting certain restrictions on the parameter space can make the model identifiable.

In what follows, we show that the score distributions needed to build the equating transformation are identified on a bounded region. No assumptions or restrictions are needed for the derivation of these bounds.

2.3.1 Conditional Score Distributions with No Assumptions

Although the marginal score distributions are of main interest to build the equating transformation, we start analyzing the conditional score distributions as they are typically used in NEAT equating.

Let Z be a binary variable such that

$$Z = \begin{cases} 1, \text{ if test taker is administered X in } P; \\ 0, \text{ if test taker is administered Y in } Q. \end{cases} \tag{3}$$

Then, by the law of total probability (Kolmogorov 1950), it follows that

$$\begin{aligned} \text{(a) } P(X \le x \mid A) =& P(X \le x \mid A, Z = 1)P(Z = 1 \mid A) + \\ & P(X \le x \mid A, Z = 0)P(Z = 0 \mid A), \\ \text{(b) } P(Y \le y \mid A) =& P(Y \le y \mid A, Z = 1)P(Z = 1 \mid A) + \\ & P(Y \le y \mid A, Z = 0)P(Z = 0 \mid A). \end{aligned} \tag{4}$$

The statistical model underlying the NEAT design is accordingly parameterized by the parameters $\{P(X \le x \mid A = a), P(Y \le y \mid A = a)\}$. In order to show that these parameters are not identified, consider the following comments on (4):

1. $P(X \le x \mid A = a, Z = 1)$ is the conditional score probability of X given A for a test taker who actually answered form X (i.e., sampled from P) and scored $A = a$ on the anchor test.
2. $P(Z = 1 \mid A = a)$ corresponds to the proportion of test takers who were administered form X (or equivalently, proportion of people sampled from P) and scored $A = a$ on the anchor test.
3. $P(Z = 0 \mid A = a)$ corresponds to the proportion of test takers who were administered form Y (or equivalently, proportion of people sampled from Q) and scored $A = a$ on the anchor test.

4. $P(X \leq x \mid A = a, Z = 0)$ is the conditional score probability of X for a test taker who was actually administered form Y (or sampled from Q).

Consequently, $P(X \leq x \mid A = a)$ corresponds to the probability of scoring x on test form X *as if* all test takers with a score $A = a$ were administered test form X. However, this conditional probability is *not identified*. As a matter of fact, the data generating process that underlies the NEAT design only identifies $P(X \leq x \mid A = a, Z = 1)$ and $P(Z = z \mid A)$ for $z \in \{0, 1\}$. However, it does not provide any information about $P(X \leq x \mid A = a, Z = 0)$ and therefore the sampling process only reveals that

$$\begin{aligned} P(X \leq x \mid A = a) =& P(X \leq x \mid A = a, Z = 1) P(Z = 1 \mid A = a) + \\ & \gamma\, P(Z = 0 \mid A) \end{aligned}$$

for some *unknown* probability distribution γ. Therefore, $P(X \leq x \mid A = a)$ cannot be uniquely determined because γ can not be uniquely chosen. Consequently, $P(X \leq x \mid A = a)$ is not identified. Similar conclusions can be drawn for $P(Y \leq y \mid A)$.

In practice, $P(X \leq x \mid A)$ and $P(Y \leq y \mid A)$ are identified under an hypothesis of strong ignorability (e.g., Rosenbaum and Rubin 1983), namely

$$\begin{aligned} P(X \leq x \mid A, Z = 1) =& P(X \leq x \mid A, Z = 0) = P(X \leq x \mid A), \\ P(Y \leq y \mid A, Z = 1) =& P(Y \leq y \mid A, Z = 0) = P(Y \leq y \mid A), \end{aligned} \tag{5}$$

which, in the context of the current application can compactly be defined as

$$(X, Y) \perp\!\!\!\perp Z \mid A. \tag{6}$$

As a matter of fact, the strong ignorability condition essentially tells us that γ is not unknown, but it coincides with $P(X \leq x \mid A = a, Z = 1)$. This implies that $P(X \leq x \mid A = a)$ is uniquely determined, and thus identified. It is necessary to emphasize that the strong ignorability condition cannot empirically be refuted and, therefore, it should be justified in the context of an application.

2.3.2 Partially Identified Probability Distributions

The strong ignorability condition can be avoided if we find a region where the score probabilities are actually identified. In this section we show that such region indeed exists. As a matter of fact, because $P(X \leq x \mid A, Z = 0)$ is bounded between 0 and 1, from (4) it can easily be verified that

$$L_x \leq P(X \leq x \mid A) \leq U_x, \tag{7}$$

where

$$\begin{aligned} L_x =& P(X \leq x \mid A, Z = 1)P(Z = 1 \mid A) \\ U_x =& P(X \leq x \mid A, Z = 1)P(Z = 1 \mid A) + P(Z = 0 \mid A) \end{aligned} \tag{8}$$

Analogously for Y it can be verified that

$$L_y \leq P(Y \leq y \mid A) \leq U_y, \tag{9}$$

where

$$\begin{aligned} L_y =& P(Y \leq y \mid A, Z = 0)P(Z = 0 \mid A) \\ U_y =& P(Y \leq y \mid A, Z = 0)P(Z = 0 \mid A) + P(Z = 1 \mid A) \end{aligned} \tag{10}$$

Thus, the conditional score distributions are partially identified (Tamer 2010) on regions defined by the derived bounds. Note that the length of the intervals for $P(X \leq x \mid A)$ and $P(Y \leq y \mid A)$ are $P(Z = 0 \mid A)$ and $P(Z = 1 \mid A)$, respectively, and as mentioned before they correspond to the proportion of test takers in P and Q, respectively, for a given score A.

2.3.3 Marginal Distributions with No Assumptions

The equating transformation φ is built from marginal score distributions defined on a common population. It is thus of interest to examine if the preceding arguments are also valid when the conditional distributions are marginalized over the anchor scores. It is easy to see that marginalizing over A in (4) we obtain

$$P(X \leq x) = P(X \leq x \mid Z = 1)P(Z = 1) + P(X \leq x \mid Z = 0)P(Z = 0). \tag{11}$$

Note that the identifiability problem still remains in the marginal score distribution as $P(X \leq x \mid Z = 0)$ is non identified. However, because this probability is bounded between 0 and 1, we can show similarly as before that $P(X \leq x)$ can also be bounded. In fact,

$$L_x \leq P(X \leq x) \leq U_x, \tag{12}$$

where

$$\begin{aligned} L_x =& P(X \leq x \mid Z = 1)P(Z = 1) \\ U_x =& P(X \leq x \mid Z = 1)P(Z = 1) + P(Z = 0) \end{aligned} \tag{13}$$

Note that using the definition in (3), Eq. (11) can be rewritten as

$$F_X(x) = wF_{XP}(x) + (1 - w)F_{XQ}(x) \tag{14}$$

with $w = P(Z = 1)$. Interestingly, the right hand sides of Eqs. (2) and (14) are *visually* identical. This result would indicate that the weights in the definition of a synthetic population are actually related to the proportion of test takers in the populations and thus should not be arbitrarily chosen. Moreover, w corresponds to the length of the interval where the score distribution is partially identified. Analogous results as the ones shown in (11), (12), (13), and (14) can be derived for $P(Y \leq y)$.

A natural question at this stage is how $F_X(x)$ and $F_Y(y)$ compare to $F_{XT}(x)$ and $F_{YT}(y)$, respectively. Such comparison is not possible because the formers distributions are not identified and thus non observable. We have shown that they are however partially identified on a bounded region so that it is possible to evaluate the behavior of the bounds and how it relates to the target distributions traditionally obtained in NEAT equating using the definition of synthetic population and the ignorability condition. This is done in the following section.

Table 2 Bivariate score frequencies (X, A) and (Y, A)

X	A	Frequency	Y	A	Frequency
0	0	4	0	0	4
0	1	4	0	1	3
0	2	2	0	2	1
0	3	0	0	3	0
1	0	4	1	0	7
1	1	8	1	1	5
1	2	2	1	2	7
1	3	1	1	3	1
2	0	6	2	0	3
2	1	12	2	1	5
2	2	5	2	2	12
2	3	2	2	3	2
3	0	3	3	0	3
3	1	12	3	1	4
3	2	5	3	2	13
3	3	5	3	3	5
4	0	2	4	0	2
4	1	3	4	1	2
4	2	4	4	2	5
4	3	6	4	3	6
5	0	1	5	0	1
5	1	1	5	1	1
5	2	2	5	2	2
5	3	6	5	3	6

3 Illustrations

3.1 Data

We use data from an hypothetical example shown in Kolen and Brennan (Kolen and Brennan (2014), Sect. 5.1.3). In this example forms X and Y each contain 5 items and 3 common items. The data in Kolen and Brennan (2014) are originally displayed as joint probabilities $f_{XP}(x,a) = P(X = x, A = a)$ and $f_{YQ}(y,a) = P(Y = y, A = a)$ and we use this information to create raw data as displayed in Table 2. The table shows bivariate score frequencies for each test form. From the table, it can be seen that, for instance, 8 test takes scored $X = 1$ and $A = 1$, whereas 13 scored $Y = 3$ and $A = 2$, etc. For the information in the table (frequency), it follows that the sample size considered is 100 for both populations.

3.2 Results

Figure 1 shows a graphical representation of the bounds derived in (8) for the case when $A = 2$. From the figure, it can be seen that the bounds for the conditional distribution of X given A are wider than the ones for the conditional distribution of Y given A, when $A = 2$. Note, however, that this situation could change for other values of the anchor score. Moreover, the curves are *parallel* in the sense that the length of the intervals are constant for all values of scores on the scale, for a given value of A.

Fig. 1 Bounds for conditional score distributions $P(X \leq x \mid A = 2)$ and $P(Y \leq y \mid A = 2)$

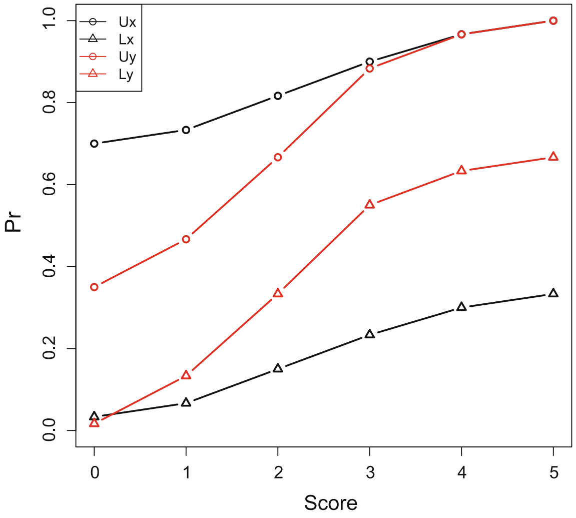


Table 3 Target cumulative distributions for Forms X and Y scores, and derived bounds

Score	F_{XT}	$[L_x, U_x]$	F_{YT}	$[L_y, U_y]$
0	0.100	[0.050; 0.550]	0.105	[0.040; 0.540]
1	0.250	[0.125; 0.625]	0.320	[0.140; 0.640]
2	0.500	[0.250; 0.750]	0.530	[0.250; 0.750]
3	0.750	[0.375; 0.875]	0.755	[0.375; 0.875]
4	0.900	[0.450; 0.950]	0.900	[0.450; 0.950]
5	1.000	[0.500; 1.000]	1.000	[0.500; 1.000]

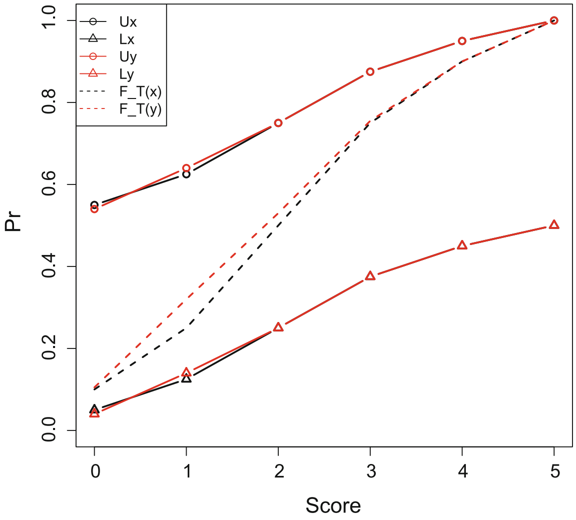


Fig. 2 Bounds for $F_X(x) = P(X \leq x)$ and $F_Y(y) = P(Y \leq y)$, and target score distributions $F_{XT}(x)$ and $F_{YT}(y)$ for the case $w = 1$

This is due to the fact that, as seen at the end of Sect. 2.3.2, the length of the intervals are defined by $P(X \leq x \mid A)$ and $P(Y \leq y \mid A)$.

Next, we calculated the bounds derived in Sect. 2.3.3 for each of the marginal score distributions. Because the real value of F_X and F_Y is unknown, we use the derived target cumulative distribution functions F_{XT} and F_{YT} as reference for comparison. The latter where obtained assuming that $w = 1$. Table 3 shows the target cumulative distributions and the corresponding bounds where the marginal score distributions are partially identified. Figure 2 shows a graphical representation of these results.

From Table 3 and Fig. 2, it can be seen that all the values of F_{XT} and F_{YT} lie in the intervals $[L_x, U_x]$ and $[L_y, U_y]$, respectively, as expected. Note also that the intervals have length equal to 0.5. This is because the sample sizes in both populations is exactly the same (100 in this case), so that $P(Z = 1) = P(Z = 0) = \frac{100}{200} = 0.5$ (see comments on Sect. 2.3.3 below Eq. (14)).

4 Concluding Remarks

In this paper, we have argued that there is an inherent identification problem underlying the NEAT equating design. The assumption on the equality of conditional score distributions, typically made in NEAT equating and called here an ignorability condition, has been shown to actually be an identification restriction. We offered an alternative to the ignorability condition and proposed to work with partially identified probability distributions.

The derived bounds on the partially identified region showed that there is huge uncertainty about the probability distributions that are to be used for equating. The actual impact of this method on equating is currently being investigated by the authors.

The exposition focused on poststratification equating under the NEAT design. However, the identifiability problem also arises for the case when chained equipercentile equating (e.g., Kolen and Brennan 2014) is used to equate score data collected under the NEAT design. In fact, different assumptions are needed to identify the target score distributions used to build the equating transformation (see, e.g., von Davier et al. 2004, Sect. 2.4.1). The derivation of bounds where the score distributions are partially identified for the case of chained equating is currently being investigated by the authors.

Acknowledgements Jorge González acknowledges partial support of grant Fondecyt 1150233. Ernesto San Martín acknowledges partial support of grant Fondecyt 1141030.

References

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). Academic Press.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). CA: Duxbury Pacific Grove.

von Davier, A., Holland, P., Thayer, D. (2004). *The kernel method of test equating*. Springer

González, J., & Wiberg, M. (2017). *Applying test equating methods, using R*. Springer International Publishing

Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

Kolmogorov AN. (1950). Foundations of the theory of probability. Chelsea Publishing Co.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Econometrics*, *2*(1), 167–195.

Simultaneous Equating of Multiple Forms

Michela Battauz

Abstract When test forms are calibrated separately, item response theory parameters are not comparable because they are expressed on different measurement scales. The equating process converts the item parameter estimates to a common scale and provides comparable test scores. Various statistical methods have been proposed to perform equating between two test forms. However, many testing programs use several forms of a test and require the comparability of the scores of each form. To this end, Haberman (ETS Res Rep Ser 2009(2):i–9, 2009) developed a regression procedure that generalizes the mean-geometric mean method to the case of multiple test forms. A generalization to multiple test forms of the mean-mean, the Haebara, and the Stocking-Lord methods was proposed in Battauz (Psychometrika 82:610–636, 2017b). In this paper, the methods proposed in the literature to equate multiple test forms are reviewed, and an application of these methods to data collected for the Trends in International Mathematics and Science Study will be presented.

Keywords Equating · Linking · Multiple forms

1 Introduction

Many testing programs use a large number of different forms of a test to assess the achievement levels. Two examples are given by the Program for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS). Both these testing programs involve a very large number of students from many different countries, and require the administration of a large number of items, which are organized in booklets. Of course, the raw scores are not directly comparable, thus requiring the use of equating procedures. One option is given by concurrent calibration, which estimates all item parameters in one run for all the forms. However, this approach is highly computationally demanding, as the data

M. Battauz (✉)
Department of Economics and Statistics, University of Udine, via Tomadini 30/A,
33100 Udine, Italy
e-mail: michela.battauz@uniud.it

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_11

matrix containing the responses of all the examinees to all the items is required. When a testing program continues over many different years, concurrent calibration becomes challenging if not unfeasible. A different approach is given by separate estimation of the item parameters for each form, followed by rescaling of the item parameter estimates in order to obtain values expressed on a common metric. The literature proposed various methods for the estimation of the equating coefficients, which are two constants used to convert the item parameters. These methods include the mean-mean (Loyd and Hoover 1980), the mean-geometric mean (Mislevy and Bock 1990), Haebara (1980) and Stocking and Lord (1983) methods. However, all these methods can be applied only to two forms with some items in common. The first proposal to handle the case of multiple forms was given by Haberman (2009), who developed a regression procedure that generalizes the mean-geometric mean method. A generalization of the mean-mean, Haebara and Stocking-Lord methods to the case of multiple forms is given in Battauz (2017b). This paper provides also a procedure for the computation of the standard errors of the equating coefficients estimated using all these methods.

Separate calibration is not only convenient from a computational point of view. Separate calibration allows for a better control of the accuracy of the equating process, which is indicated by the standard errors of the equating coefficients, and constitutes a suitable setting for monitoring item parameter drift.

In this paper, the methods proposed in the literature for the computation of the equating coefficients will be reviewed, and an application of these methods to data collected for TIMSS will be presented. The paper is structured as follows. Section 2 reviews the methods available for multiple equating, Sect. 3 provides a real data example, and Sect. 4 contains some concluding remarks.

2 Methods

In Item Response Theory (IRT), the probability of a correct response to item j is modeled as a function of the ability level, θ, and some item parameters

$$P(\theta; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp\{Da_j(\theta - b_j)\}}{1 + \exp\{Da_j(\theta - b_j)\}}, \qquad (1)$$

where a_j is the discrimination parameter, b_j is the difficulty parameter, c_j is the guessing parameter, and D is a known constant. This specification corresponds to the so called three-parameter logistic (3PL) model. The two-parameter logistic (2PL) model results when the guessing parameter is equal to zero, while the one-parameter logistic (1PL) model requires also that the discrimination parameter is set to one. IRT models are usually estimated by means of the marginal maximum likelihood method (Bock and Aitkin 1981), which treats the abilities as random variables. Due to identifiability issues in all IRT models (Reise and Revicki 2015, p. 45), the abilities are assumed to have zero mean and variance equal to one. For this reason, when the item parameters are estimated separately for different administrations of the test, the item parameter estimates are not comparable, as they are expressed on different measurement scales. The item parameters can be converted to a common metric using the following equations

$$a_j^* = \frac{a_{jt}}{A_t} \tag{2}$$

and

$$b_j^* = b_{jt} A_t + B_t, \tag{3}$$

where t is the index of the administrations, A_t and B_t are the equating coefficients related to administration t, a_j^* and b_j^* are the discrimination and difficulty parameters expressed on a common metric. In the following, the methods proposed in the literature for the estimation of the equating coefficients will be briefly described.

2.1 Multiple Mean-Geometric Mean

Haberman (2009) proposed to employ Eqs. (2) and (3) to specify the regression models

$$\log \hat{a}_{jt} = \log \hat{A}_t + \log \hat{a}_j^* + e_{1jt} \tag{4}$$

and

$$\hat{b}_{jt} \hat{A}_t = -\hat{B}_t + \hat{b}_j^* + e_{2jt}, \tag{5}$$

where e_{1jt} and e_{2jt} are the residuals that should be introduced because Eqs. (2) and (3) hold only approximately in samples. In the first stage, the estimates $\log \hat{A}_t$ and $\log \hat{a}_j^*$ are obtained using the least squares method. In the second stage, the estimates $\hat{A}_t = \exp(\log \hat{A}_t)$ are used to compute the responses $\hat{b}_{jt} \hat{A}_t$ of the regression model (5), and the estimates $\hat{B}_t$ and $\hat{b}_j^*$ are obtained by means of the least square method. The equating coefficients $\hat{A}_1$ and $\hat{B}_1$ are constrained to 1 and 0.

When this method is applied to two forms (i.e. $T = 2$) it can be shown that it corresponds to the mean-geometric mean method. For this reason, this method is called the multiple mean-geometric mean (MM-GM) method in this paper.

2.2 Multiple Mean-Mean

The mean-mean method for pairs of forms would estimate A_t using the following equation

$$\hat{A}_t = \frac{\sum_{j \in J_t} \hat{a}_{jt}}{\sum_{j \in J_t} \hat{a}_j^*}, \tag{6}$$

where J_t is the set of items administered in administration t. However, $\hat{a}_j^*$ is not available. The proposal in Battauz (2017b) is to replace $\hat{a}_j^*$ in (6) with

$$\hat{a}_j^* = \frac{\sum_{s \in U_j} \hat{a}_{js}}{\sum_{s \in U_j} \hat{A}_s}, \tag{7}$$

where U_j is the set of forms including item j. Substituting Eq. (7) in Eq. (6) yields

$$\hat{A}_t = \frac{\sum_{j \in J_t} \hat{a}_{jt}}{\sum_{j \in J_t} \frac{\sum_{s \in U_j} \hat{a}_{js}}{\sum_{s \in U_j} \hat{A}_s}}, \quad t = 2, \ldots, T. \tag{8}$$

The simultaneous solution of these $T - 1$ equations can be achieved by applying a numerical algorithm, setting $\hat{A}_1 = 1$. Once the estimates $\hat{A}_2, \ldots, \hat{A}_T$ are obtained, the estimates of the equating coefficients $B_2, \ldots, B_T$ can be obtained following the procedure of the MM-GM method.

When $T = 2$, this method is equivalent to the mean-mean method. For this reason, this method is called the multiple mean-mean (MM-M) method in this paper.

2.3 Multiple Item Response Function

The multiple item response function (MIRF) method requires the minimization of the following function with respect to all the equating coefficients

$$f_{IR}^* = \sum_{t=1}^{T} \int_{-\infty}^{\infty} \sum_{j \in J_t} \left(P_{jt} - P_{jt}^* \right)^2 h(\theta) d\theta, \tag{9}$$

where $h(\cdot)$ is the density of a standard normal distribution and

$$P_{jt} = P(\theta; \hat{a}_{jt}, \hat{b}_{jt}, \hat{c}_{jt}) \tag{10}$$

is the item response function computed using the item parameter estimates of administration t, while

$$P_{jt}^* = P(\theta; \hat{a}_{jt}^*, \hat{b}_{jt}^*, \hat{c}_{jt}), \tag{11}$$

is the item response function computed using the synthetic item parameters. The synthetic item parameters are obtain as a mean of the item parameters estimated in different administrations and converted to the common scale

$$\hat{a}_j^* = \frac{1}{u_j} \sum_{s \in U_j} \frac{\hat{a}_{js}}{\hat{A}_s} \quad \text{and} \quad \hat{b}_j^* = \frac{1}{u_j} \sum_{s \in U_j} (\hat{b}_{js} \hat{A}_s + \hat{B}_s), \tag{12}$$

and then converted back to the scale of administration t

$$\hat{a}^*_{jt} = \hat{a}^*_j \hat{A}_t \quad \text{and} \quad \hat{b}^*_{jt} = \frac{\hat{b}^*_j - \hat{B}_t}{\hat{A}_t}. \tag{13}$$

The integrals in Eq. (9) are approximated using Gaussian quadrature, and the minimization is performed numerically setting $\hat{A}_1 = 1$ and $\hat{B}_1 = 0$.

2.4 Multiple Test Response Function

The multiple test response function (MTRF) method takes its name from considering the quadratic difference between the test response functions, instead of the item response functions. This method is based on the minimization of the function

$$f^*_{TR} = \sum_{t=1}^{T} \int \left(\sum_{j \in J_t} P_{jt} - P^*_{jt} \right)^2 h(\theta) d\theta, \tag{14}$$

where P_{jt} and P^*_{jt} are defined as for the MIRF method. Similarly, Gaussian quadrature is used to approximate the integrals, and the minimization is performed numerically setting $\hat{A}_1 = 1$ and $\hat{B}_1 = 0$.

3 Real Data Example

As an example of a possible application, the multiple equating methods were applied to data collected for TIMSS 2011 (Foy et al. 2013). This example considers only achievement data in Mathematics at the fourth grade. Students were administered one of 14 forms (booklets). These forms present items in common as shown in Fig. 1. Only dichotomous items were considered for this analysis. An IRT model was fit to each form separately, using a 2PL specification for constructed response items and a 3PL specification for multiple choice items. All analyses were performed using the R statistical software (R Development Core Team 2017). The item parameter estimation was performed using the `tpm` function of the R package ltm (Rizopoulos 2006), constraining the guessing parameter to zero for the constructed response items. The R package equateMultiple (Battauz 2017a) implements all the methods illustrated in this paper and was used to estimate the equating coefficients and obtain the equated scores.

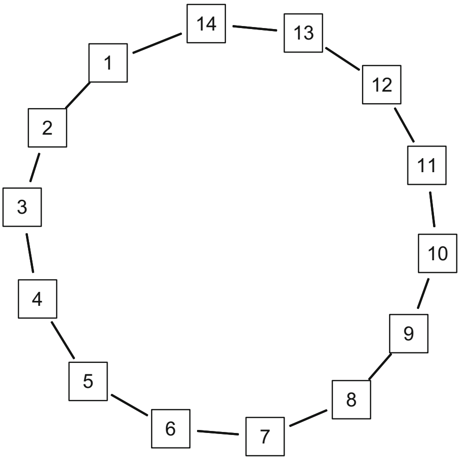


Fig. 1 Linkage plan

Table 1 shows the estimates and the standard errors of the equating coefficients. Form 1 was chosen as base form. It is possible to observe that the A equating coefficients are all around 1, while the B equating coefficients are all around 0, thus indicating that the populations who were administered the different forms do not differ much in the distribution of the abilities. Consistently with the simulation study presented in Battauz (2017b), the methods tend to give similar estimates of the equating coefficients, while the standard errors tend to be lower for the MIRF method. Observing the standard errors, it is possible to note that they are larger for forms that are linked through longer chains (see Battauz 2015).

After the conversion of the item parameters to the scale of Form 1 using the equating coefficients reported in Table 1, it is possible to obtain the equated scores. The literature proposes two main IRT methods to this end, which are true score equating and observed score equating (Kolen and Brennan 2014). Table 2 shows the equated scores obtained using the observed score equating method and the equating coefficients estimated with the MIRF method. Scores obtained using the true score equating method are not shown because very similar to those obtained with observed score equating. For example, a score of 15 in Form 1 is equivalent to a score of 19.2 in Form 2. Note that here the equivalent scores across the forms differ not only because of differences in item difficulties, but also because the number of items varies across the different forms. This value is reported in the last row of the table.

Table 1 Estimates (standard errors) of the equating coefficients for the TIMSS data

Form	MM-GM	MM-M	MIRF	MTRF
A equating coefficient				
2	1.054 (0.016)	1.053 (0.017)	1.056 (0.009)	1.069 (0.011)
3	1.043 (0.020)	1.045 (0.020)	1.023 (0.013)	1.055 (0.015)
4	1.087 (0.025)	1.097 (0.025)	1.059 (0.015)	1.122 (0.021)
5	1.117 (0.028)	1.130 (0.028)	1.054 (0.018)	1.141 (0.025)
6	1.152 (0.030)	1.163 (0.030)	1.067 (0.018)	1.179 (0.027)
7	1.123 (0.031)	1.137 (0.031)	1.025 (0.018)	1.133 (0.029)
8	1.095 (0.030)	1.108 (0.030)	1.073 (0.019)	1.146 (0.027)
9	1.045 (0.028)	1.051 (0.028)	1.042 (0.019)	1.105 (0.029)
10	1.024 (0.027)	1.032 (0.027)	1.038 (0.019)	1.088 (0.027)
11	1.066 (0.027)	1.074 (0.027)	1.008 (0.018)	1.079 (0.025)
12	1.098 (0.025)	1.095 (0.024)	1.088 (0.018)	1.113 (0.021)
13	0.999 (0.020)	0.995 (0.019)	0.971 (0.014)	0.981 (0.016)
14	0.946 (0.015)	0.943 (0.015)	0.961 (0.011)	0.947 (0.012)
B equating coefficient				
2	−0.138 (0.020)	−0.138 (0.019)	−0.103 (0.012)	−0.117 (0.015)
3	−0.127 (0.025)	−0.127 (0.024)	−0.090 (0.016)	−0.110 (0.019)
4	−0.161 (0.030)	−0.166 (0.029)	−0.116 (0.018)	−0.157 (0.022)
5	0.020 (0.031)	0.017 (0.031)	0.024 (0.020)	0.015 (0.023)
6	0.003 (0.033)	0.000 (0.033)	0.020 (0.021)	0.005 (0.025)
7	0.080 (0.036)	0.076 (0.036)	0.093 (0.021)	0.091 (0.025)
8	0.010 (0.038)	0.007 (0.038)	0.012 (0.022)	0.036 (0.026)
9	0.041 (0.037)	0.039 (0.036)	0.002 (0.021)	0.042 (0.027)
10	0.047 (0.034)	0.044 (0.034)	0.001 (0.020)	0.039 (0.025)
11	0.042 (0.032)	0.040 (0.031)	0.029 (0.018)	0.046 (0.022)
12	−0.011 (0.028)	−0.006 (0.028)	−0.026 (0.018)	0.011 (0.021)
13	0.059 (0.023)	0.062 (0.023)	0.026 (0.016)	0.057 (0.017)
14	0.070 (0.018)	0.072 (0.018)	0.034 (0.013)	0.060 (0.014)

4 Conclusions

The multiple equating methods presented in this paper constitute a good alternative to concurrent calibration to handle the case of multiple forms to be equated. The advantages of the equating methods based on separate calibration are related to less computational cost and a better control of the accuracy of the equating process. The latter can be based on the standard errors of the equating coefficients, which also affect the standard errors of the transformed ability values (see Battauz 2017b, Appendix 4).

Table 2 Equated scores for the TIMSS data

Forms	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Scores	0	0.9	0.8	0.7	0.9	1.3	1.1	1.4	1.6	1.2	0.9	1.1	0.8	0.1
	1	2.3	2.2	2.0	2.5	2.9	2.5	2.8	3.1	2.7	2.2	2.6	2.3	1.1
	2	3.7	3.6	3.2	4.0	4.4	3.6	4.0	4.6	4.1	3.4	4.0	3.8	2.1
	3	5.0	4.8	4.4	5.6	5.9	4.7	5.2	6.0	5.5	4.5	5.4	5.2	3.0
	4	6.4	6.1	5.5	7.2	7.4	5.8	6.3	7.3	6.8	5.6	6.8	6.7	3.9
	5	7.7	7.3	6.5	8.6	8.9	6.8	7.3	8.5	8.1	6.6	8.1	8.1	4.8
	6	9.0	8.4	7.5	9.9	10.2	7.9	8.3	9.6	9.4	7.6	9.4	9.4	5.7
	7	10.2	9.6	8.5	11.1	11.5	8.9	9.2	10.7	10.6	8.7	10.6	10.7	6.6
	8	11.4	10.7	9.5	12.3	12.7	9.9	10.2	11.8	11.8	9.7	11.8	12.0	7.5
	9	12.6	11.8	10.5	13.3	13.8	11.0	11.1	12.8	13.0	10.8	12.9	13.2	8.5
	10	13.7	12.9	11.4	14.3	14.9	12.1	12.1	13.9	14.3	12.0	14.0	14.5	9.5
	11	14.9	13.9	12.4	15.3	16.0	13.3	13.1	15.0	15.5	13.2	15.2	15.6	10.6
	12	16.0	15.0	13.4	16.2	17.0	14.4	14.2	16.1	16.8	14.4	16.2	16.8	11.7
	13	17.1	16.1	14.4	17.0	18.1	15.7	15.3	17.3	18.1	15.8	17.3	18.0	12.7
	14	18.1	17.2	15.4	17.9	19.1	16.9	16.4	18.6	19.4	17.1	18.4	19.1	13.8
	15	19.2	18.3	16.4	18.7	20.2	18.1	17.6	19.9	20.7	18.5	19.5	20.2	14.9
	16	20.2	19.3	17.4	19.5	21.2	19.4	18.9	21.3	22.2	20.0	20.6	21.2	15.9
	17	21.2	20.4	18.3	20.2	22.2	20.7	20.2	22.7	23.6	21.4	21.6	22.1	16.9
	18	22.2	21.3	19.3	20.9	23.2	22.1	21.6	24.2	25.1	22.9	22.7	23.1	18.0
	19	23.1	22.3	20.2	21.6	24.1	23.4	23.0	25.7	26.6	24.3	23.8	24.0	18.9
	20	24.1	23.2	21.1	22.3	25.1	24.6	24.4	27.3	28.2	25.6	24.9	25.0	19.9
Max	20	24	23	21	22	25	25	25	28	29	26	25	25	20

The application proposed in this paper considers TIMSS data of only one year. This should be regarded as an example of possible application of these methods, since considering more years can easily be implemented.

Hybrid strategies can also be considered and are straightforward to implement. For example, it is possible to use concurrent calibration for each year of assessment, and employ the equating methods presented here to achieve the comparability between different years.

The main limitations of the methods illustrated in this paper is that they treat the item parameter estimates as independent and homoscedastic variables. However, item parameter estimates are independent only between different administrations, while they are dependent within the same administration. Furthermore, the variability of the estimates is not constant over different items. The main factor that influences the variability of the item parameter estimates is the sample size. If the number of examinees taking the test does not vary largely over different administrations, the variability of the item parameter estimates tends to present similar values. On the contrary, different sample sizes result in a larger variability of the parameter estimates. In this case, it would be worth the development of a method for the simultaneous estimation of the equating coefficients between multiple forms that takes into account the different variability of the estimates. Accounting also for the dependence of the item parameter estimates deriving from the same administration, would further improve the efficiency of the estimators of the equating coefficients. It is worth noting that the main linking methods currently used to equate two form (i.e. mean-mean, the mean-geometric mean, Haebara and Stocking-Lord methods) do not account for the different variability of the item parameter estimates or their dependence. Thus, the development of new methods to equate simultaneously multiple forms that take into account the heteroscedasticity and the dependence of the item parameter estimates would also be of interest to equate two forms more efficiently. A new method that considers the item parameter estimates as dependent variables with different variances is currently under study by the author.

References

Battauz, M. (2015). Factors affecting the variability of IRT equating coefficients. *Statistica Neerlandica*, *69*, 85–101.

Battauz, M. (2017a). *equateMultiple: Equating of multiple forms*. R package version 0.0.0.

Battauz, M. (2017b). Multiple equating of separate IRT calibrations. *Psychometrika*, *82*, 610–636.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Foy, P., Arora, A., & Stanco, G. M. (2013). *TIMSS 2011 User Guide for the International Database*.

Haberman, S. J. (2009). Linking parameter estimates derived from an item response model through separate calibrations. *ETS Research Report Series*, *2009*(2), i–9.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179–193.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

R Development Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. P., & Revicki, D. A. (2015). *Handbook of item response theory modeling: Applications to typical performance assessment*. New York: Routledge.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

Incorporating Information Functions in IRT Scaling

Alexander Weissman

Abstract Item response theory (IRT) scaling via a set of items common to two test forms assumes that those item's parameters are invariant with respect to a linear transformation. Characteristic curve methods rely on this assumption; scale transformations are conducted by minimizing a loss function between item characteristic curves (ICCs), as in the case of Haebara (1980), or test characteristic curves (TCCs), as in the case of Stocking and Lord (1983). However, minimizing the loss function between characteristic curves does not guarantee that the same will hold for information functions. This study introduces two new scaling methodologies: one combines the ICC methodology of Haebara (1980) with item information functions (IIFs); the other combines the TCC methodology of Stocking and Lord (1983) with test information functions (TIFs). In a simulation experiment, Haebara's (1980) and Stocking and Lord's (1983) methodologies as well as the two new scaling methodologies were applied to simulated administrations of a fixed form under different latent trait distributions. Results suggest that IRT scaling by combining TCCs with TIFs yields some benefits over the existing characteristic curve methodologies; however, combining ICCs with IIFs did not perform as well as the other three scaling methodologies.

Keywords Item response theory · Scaling · Scale transformations
Characteristic curve methods · Information functions

1 Introduction

Consider two test forms where the item response theory (IRT) parameters for the items on each form have been estimated separately from one another. IRT scaling refers to the methodology for placing the item parameters from one form onto the scale of the other form. The need for IRT scaling comes about due to the inde-

A. Weissman (✉)
Law School Admission Council, 662 Penn Street, Newtown, PA, USA
e-mail: aweissman@lsac.org

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_12

terminacy of IRT latent trait scales; as Kolen and Brennan (1995) note, "[If] an IRT model fits a set of data, then any linear transformation of the [latent trait] scale also fits the set of data, provided that the item parameters are also transformed (p. 162)." Thus, IRT scaling seeks to find an optimal linear transformation of the parameters from one form onto the scale of another form.

When two test forms have items in common, the non-equivalent groups with anchor test design can be used in IRT scaling. In that case, only the items common to both forms are utilized in the IRT scaling method. There are two crucial assumptions with this design. First, the item parameters of the common items are assumed to be invariant with respect to a linear transformation; that is, any common item will function in the same way independent of the form on which it appears, and independent of the group to which it is administered. Second, a linear transformation of one form's item parameters to another is assumed to be valid for all items on that form, regardless of whether those items are in common with items on the other form; that is, the same linear transformation obtained from a set of common items can be extended to all items on a form.

IRT scaling may be conducted via methods that focus on the moments of the distribution of common item parameters only, such as the mean/sigma method (Marco 1977) and the mean/mean method (Loyd and Hoover 1980). Typically, however, IRT scaling is conducted by characteristic curve methods. In IRT, an item characteristic curve (ICC) is a function that relates the probability of correctly responding to an item with the level of the underlying latent trait. Likewise, a test characteristic curve (TCC) is a function that relates the expected number correct score on a set of items with the level of the underlying trait. Haebara (1980) introduced an IRT scaling methodology that utilizes item characteristic curves, while Stocking and Lord (1983) introduced a scaling methodology that utilizes test characteristic curves. In both cases, an optimal linear transformation for placing the item parameters of one form onto the scale of the other is obtained by minimizing an objective function defined in terms of a sum of squared differences. In the case of Haebara (1980), the squared differences are between the ICCs from each common item, whereas in the case of Stocking and Lord (1983), the squared differences are between the TCCs for all common items.

1.1 Impact of Item Parameter Transformations on Information Functions

When IRT scaling is conducted using characteristic curve methods, the optimal linear transformation is the one that minimizes a function of squared differences between characteristic curves, by definition. However, *information functions* in IRT are also affected by linear transformations of item parameters. (Note that this applies to any linear transformation, including those obtained by mean/sigma or mean/mean methods.) Like characteristic curves, information functions can be defined at the item- or test-level. Whereas characteristic curves relate expected number correct scores with the level of the underlying latent trait, information functions are related to the precision of measuring the level of the underlying latent trait itself.

Because information functions relate to the measurement precision of the latent trait, it is reasonable to expect that the information functions for a set of items common to two forms should be similar; otherwise, the measurement precision provided by the items on one form would differ from the measurement precision provided by the items on another form, even though the same set of items appear on both forms. An analogous argument can be made for characteristic curve methods; indeed, the minimization of squared differences between characteristic curves is a mathematical statement of the goal for having the item- or test characteristic curves between forms align as closely as possible.

The problem of obtaining an optimal linear transformation by characteristic curve methods is that it does not necessarily translate into minimizing the squared differences between information functions. That is, after transformation, characteristic curves may be similar, but information functions may not. An example is provided in Fig. 1 for IRT scaling using Stocking and Lord's (1983) methodology, where test characteristic curves are quite well aligned, but the test information functions are not.

The purpose of this study is to investigate how information functions can be incorporated into IRT scaling, and the effect of incorporating these functions on the item parameter transformations with respect to both characteristic curves and information functions. Two new scaling methodologies are introduced: one combines Haebara's (1980) methodology with item-level information functions; the other combines Stocking and Lord's (1983) methodology with test-level information functions. The performance of these two new methodologies is compared with the corresponding characteristic curve methodologies in a simulation study.

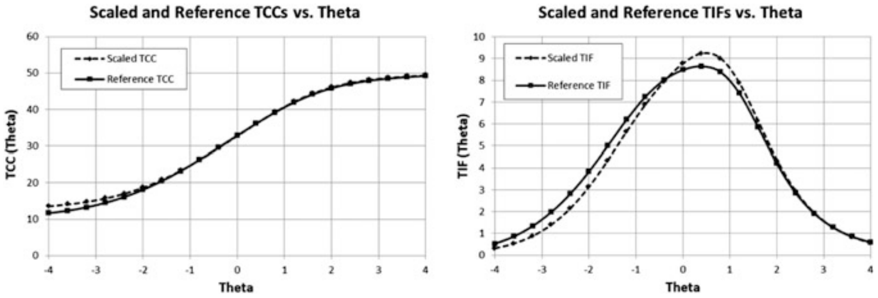


Fig. 1 Test characteristic curves (TCCs) and test information functions (TIFs) for two forms. IRT scaling transforms the item parameters from one form (labeled "scaled") onto the scale of another form (labeled "reference"). The latent trait level is indicated by "theta" on the horizontal axis

2 Methodology

As mentioned in the introduction, IRT scaling refers to the methodology for placing the item parameters from one form onto the scale of another form. Denote the form whose item parameters are being transformed as the *new form*, and the form whose item parameters are on the reference scale (and hence will not have its item parameters transformed) as the *reference form*. Further, denote the set of transformed new form item parameters as the *scaled* item parameters.

For the 3-parameter logistic (3PL) IRT model (Birnbaum 1968), a linear transformation of the new form item parameters to the scale of the reference form item parameters is defined as follows:

$$\begin{aligned} a_{j,scaled} &= \frac{a_{j,new}}{A} \\ b_{j,scaled} &= Ab_{j,new} + B \\ c_{j,scaled} &= c_{j,new} \end{aligned} \tag{1}$$

such that for item j on the new form, $a_{j,new}$ is the discrimination, or a, parameter, $b_{j,new}$ is the difficulty, or b, parameter, and $c_{j,new}$ is the pseudo-guessing, or c parameter. When linear transformation coefficients A and B are applied to the new form item parameters, the resulting set of scaled item parameters are denoted as $a_{j,scaled}$, $b_{j,scaled}$, and $c_{j,scaled}$ in (1). Note that the c parameter is not transformed, since it is not on the same metric as the latent trait scale. Further, the new form latent trait scale is transformed by applying the linear transformation as:

$$\theta_{i,scaled} = A\theta_{i,new} + B \tag{2}$$

where the subscript i indicates a latent trait level, $\theta_{i,new}$ is the latent trait level on the new form scale, and $\theta_{i,scaled}$ is the corresponding latent trait level on the reference form scale.

2.1 IRT Characteristic Curve Scaling as Optimization

IRT characteristic curve scaling is an optimization problem where nonlinear programming is used to minimize an objective function. The objective function may be defined in terms of item characteristic curves (ICCs) as in the case of Haebara (1980), or in terms of test characteristic curves (TCCs) as in the case of Stocking and Lord (1983). For the 3PL model, an ICC evaluated at latent trait level θ_k for item j is defined as:

$$ICC_j(\theta_k) = P\left(X_j = 1 \middle| \theta_k, a_j, b_j, c_j\right) \tag{3}$$

where $P\left(X_j = 1 \middle| \theta_k, a_j, b_j, c_j\right)$ is the probability of correctly responding to item j given θ_k and item parameters $\{a_j, b_j, c_j\}$, such that

$$P\left(X_j = 1 \middle| \theta_k, a_j, b_j, c_j\right) = c_j + \frac{1 - c_j}{1 + \exp\left[-Da_j(\theta_k - b_j)\right]} \tag{4}$$

where the constant D is usually chosen to be equal to 1.702.

Now consider a set V of items. For the entire set of items in V, a TCC evaluated at θ_k is defined as the summation of the ICCs for each of the items:

$$TCC(\theta_k) = \sum_{j \in V} ICC_j(\theta_k) \tag{5}$$

Suppose a set V of items is common to both the new form and the reference form. Then the optimization problem corresponding to Haebara's (1980) method may be written as:

$$\min_{A,B} \sum_{k=1}^{K} f(\theta_k) \sum_{j \in V} \left[ICC_{j,scaled}(\theta_k) - ICC_{j,ref}(\theta_k)\right]^2 \tag{6}$$

where

$$ICC_{j,scaled}(\theta_k) = P\left(X_j = 1 | \theta_k, \frac{a_{j,new}}{A}, Ab_{j,new} + B, c_{j,new}\right) \tag{7}$$

$$ICC_{j,ref}(\theta_k) = P\left(X_j = 1 | \theta_k, a_{j,ref}, b_{j,ref}, c_{j,ref}\right) \tag{8}$$

such that for any common item $j \in V$, $\{a_{j,new}, b_{j,new}, c_{j,new}\}$ are the item parameters for item j on the new form, and $\{a_{j,ref}, b_{j,ref}, c_{j,ref}\}$ are the item parameters for item j on the reference form. Note that the scaled item parameters for item j are implicit in (7); see (1) for the correspondences.

Because IRT scaling is conducted over a range of latent trait levels, the individual θ_k are indexed by $k = 1, 2, \ldots, K$ for the K latent trait levels being evaluated in (6). The function $f(\theta_k)$ serves the purpose of weighting the θ_k if so desired. Note that in Haebara's (1980) method, $f(\theta_k) = 1$ for all k.

The optimization problem corresponding to Stocking and Lord's (1983) method may be written in a similar fashion:

$$\min_{A,B} \sum_{k=1}^{K} f(\theta_k) \left[TCC_{scaled}(\theta_k) - TCC_{ref}(\theta_k)\right]^2 \tag{9}$$

where

$$TCC_{scaled}(\theta_k) = \sum_{j \in V} ICC_{scaled,j}(\theta_k) \tag{10}$$

$$TCC_{ref}(\theta_k) = \sum_{j \in V} ICC_{ref,j}(\theta_k) \tag{11}$$

such that the summation over common items $j \in V$ is incorporated in the definitions for the test characteristic curves in (10) and (11). Note also that in Stocking and Lord's (1983) method, $f(\theta_k) = 1$ for all k.

Incorporating Information Functions. The optimization problems in (6) and (9) corresponding to Haebara's (1980) and Stocking and Lord's (1983) methods can be extended to include item- or test-level information functions. Modifying the objective functions to include both characteristic curves and information functions yield two new scaling methodologies examined in this study. First, consider the item-level information function, or *item information function* (IIF), for a 3PL item j evaluated at θ_k:

$$IIF_j(\theta_k) = \frac{D^2 a_j^2 \left[1 - P_j(\theta_k)\right]}{P_j(\theta_k)} \left[\frac{P_j(\theta_k) - c_j}{1 - c_j}\right]^2 \tag{12}$$

where $P_j(\theta_k) \equiv P\left(X_j = 1 \middle| \theta_k, a_j, b_j, c_j\right)$ with item parameters $\left\{a_j, b_j, c_j\right\}$ as in (4). Next, consider again a set V of items. The test-level information function, or *test information function* (TIF) for the entire set of items in V evaluated at θ_k is defined as the summation of the IIFs for each of the items:

$$TIF(\theta_k) = \sum_{j \in V} IIF_j(\theta_k) \tag{13}$$

With these definitions for IIFs and TIFs, the optimization problems incorporating both characteristic curves and information functions can now be presented.

Combined Item-Level Methodology. This extension of Haebara's (1980) method is referred to here as the "Combined Item-Level" scaling methodology. Extending (6) to incorporate IIFs yields the following optimization problem:

$$\min_{A,B} \sum_{k=1}^{K} f(\theta_k) \sum_{j \in V} \left\{ \left[ICC_{j,scaled}(\theta_k) - ICC_{j,ref}(\theta_k)\right]^2 + \left[IIF_{j,scaled}(\theta_k) - IIF_{j,ref}(\theta_k)\right]^2 \right\} \tag{14}$$

where $ICC_{j,scaled}(\theta_k)$ and $ICC_{j,ref}(\theta_k)$ are as defined in (7) and (8), respectively, $IIF_{j,scaled}(\theta_k)$ is defined as:

$$IIF_{j,scaled}(\theta_k) = \frac{D^2 a_{j,new}^2}{A^2} \frac{\left[1 - P_{j,scaled}(\theta_k)\right]}{P_{j,scaled}(\theta_k)} \left[\frac{P_{j,scaled}(\theta_k) - c_{j,new}}{1 - c_{j,new}}\right]^2 \tag{15}$$

where

$$P_{j,scaled}(\theta_k) \equiv P\left(X_j = 1 | \theta_k, \frac{a_{j,new}}{A}, Ab_{j,new} + B, c_{j,new}\right) \tag{16}$$

and $IIF_{j,ref}(\theta_k)$ is defined as:

$$IIF_{j,ref}(\theta_k) = \frac{D^2 a_{j,ref}^2 \left[1 - P_{j,ref}(\theta_k)\right]}{P_{j,ref}(\theta_k)} \left[\frac{P_{j,ref}(\theta_k) - c_{j,ref}}{1 - c_{j,ref}}\right]^2 \tag{17}$$

where

$$P_{j,ref}(\theta_k) \equiv P\left(X_j = 1 | \theta_k, a_{j,ref}, b_{j,ref}, c_{j,ref}\right) \tag{18}$$

Combined Test-Level Methodology. This extension of Stocking and Lord's (1983) method is referred to here as the "Combined Test-Level" scaling methodology. Extending (9) to incorporate TIFs yields the following optimization problem:

$$\min_{A,B} \sum_{k=1}^{K} f(\theta_k) \left\{ \left[TCC_{scaled}(\theta_k) - TCC_{ref}(\theta_k)\right]^2 + \left[TIF_{scaled}(\theta_k) - TIF_{ref}(\theta_k)\right]^2 \right\} \tag{19}$$

where $TCC_{scaled}(\theta_k)$ and $TCC_{ref}(\theta_k)$ are as defined in (10) and (11), respectively, and

$$TIF_{scaled}(\theta_k) = \sum_{j \in V} IIF_{scaled,j}(\theta_k) \tag{20}$$

$$TIF_{ref}(\theta_k) = \sum_{j \in V} IIF_{ref,j}(\theta_k) \tag{21}$$

with the item information functions in (20) and (21) corresponding to (15) and (17), respectively.

3 Experimental Design

Four IRT scaling methodologies were investigated in this study: Haebara's (1980), Stocking and Lord's (1983), the Combined Item-Level, and the Combined Test-Level methodologies. In the following, Haebara's (1980) method will be referred to simply as "Haebara"; likewise, Stocking and Lord's (1983) method will be referred to as "Stocking-Lord."

A single reference form was used throughout the study. This form contained 50 3PL items and was assembled to match specified TCC and TIF targets. The TCC and TIF for the reference form are plotted in Fig. 1 and labeled "reference." For the simulation study, new form item parameters were obtained by simulated administrations of the reference form under different generating distributions for the latent trait. Thus, the reference and new forms had all 50 items in common.

The nine experimental conditions in the study involved three levels of means $\mu = \{-1, 0, 1\}$ crossed with three levels of standard deviations $\sigma = \{0.8, 1, 1.2\}$ for the latent trait generating distributions, which were all Normal; i.e., $\theta \sim N(\mu, \sigma)$. (Note that the mean and variance convention for moments of the Normal distribution is replaced here with mean and standard deviation.) Within each condition of generating θ distribution, 100 datasets were simulated using the reference form item parameters as generating parameters. Each of these replications contained the simulated responses of 5000 examinees to the 50 items. BILOG-MG 3 (Zimowski et al. 2003) was used to estimate the item parameters from each replication; the resulting item parameter estimates were taken as new form item parameters. Note that for each replication, the mean and standard deviation for the latent trait scale in BILOG-MG was set to 0 and 1, respectively. Then, for each replication, the new form item parameters were scaled to the reference form item parameters using the four IRT scaling methodologies discussed earlier: Haebara, Stocking-Lord, Combined Item-Level, and Combined Test-Level. All scalings were conducted using SAS (SAS Institute Inc. 2012); the OPTMODEL procedure (part of SAS/OR) was used for solving the optimization problems.

For each of the optimization problems (see (6), (9), (14), and (19)), $K = 21$ latent trait levels were chosen in equally spaced intervals between -4 and $+4$ (inclusive) such that $\theta_k = \{-4, -3.6, \ldots 0, \ldots, +3.6, +4\}$. A probability density function for $f(\theta_k)$ was assigned to correspond with the Normal probability density function. Specifically, for $\theta_1 = -4$ and $\theta_2 = -3.6, f(\theta_1) = \Phi\left(\frac{\theta_1 + \theta_2}{2}\right)$, where $\Phi(\cdot)$ is the cumulative normal distribution function. For θ_k where $\theta_1 < \theta_k < \theta_{K-1}$, the following was used:

$$f(\theta_k) = \Phi\left(\frac{\theta_k + \theta_{k+1}}{2}\right) - \Phi\left(\frac{\theta_{k-1} + \theta_k}{2}\right) \tag{22}$$

Once $f(\theta_k)$ was determined for all $k < K$, $f(\theta_K)$ was calculated as $1 - \sum_{k<K} f(\theta_k)$.

4 Results

Scaling results were evaluated by comparing the scaled and reference TCCs and TIFs. Differences between the scaled and reference TCCs for a given θ_k were calculated as:

$$TCCdiff(\theta_k) = \left[TCC_{scaled}(\theta_k) - TCC_{ref}(\theta_k)\right] \tag{23}$$

Since relative efficiency measures are commonly used to compare information functions between test forms, the following relative efficiency measure was used here:

$$RE(\theta_k) = \frac{TIF_{scaled}(\theta_k)}{TIF_{ref}(\theta_k)} \tag{24}$$

Thus, within an experimental condition, there were 100 $TCCdiff(\theta_k)$ values and 100 $RE(\theta_k)$ values for each θ_k.

Results for the TCC differences and relative efficiencies were plotted as box plots. In the following, the box plots were formatted such that: a box was plotted for each latent trait level; the midline of the box indicates the median; the length of the box indicates the interquartile range; and the upper and lower whiskers indicate the maximum and minimum values, respectively. In addition, a curve passing through the means across the latent trait levels was fit to each plot. Since the study included nine experimental conditions, four IRT scaling methodologies, and two outcome measures, a total of 72 plots were generated. Due to space constraints, only a subset of these plots is presented here.

4.1 Results for Baseline Condition

The experimental condition for which the latent trait generating distribution $\theta \sim N(0, 1)$ had the same mean and standard deviation as the reference form (latent trait) scale served as the baseline condition. The results for the TCC differences and relative efficiencies for this condition are shown in Figs. 2 and 3, respectively.

In Fig. 2, it is instructive to note: (1) the deviation of the fitted curve (i.e., the curve passing through the means of the box plots) with respect to the zero line on the vertical axis; and (2) the variability of the TCC differences. For all IRT scaling methodologies, the deviation of the fitted curve was greatest at lower latent trait levels, such as $\theta \leq -1.6$. In addition, the variability of the TCC differences for a given θ_k was also greatest at these levels. For all IRT scaling methodologies except Combined Item-Level, the deviation of the fitted curve as well as variabilities in the outcome measure were smallest in the neighborhood of $\theta = 0$; however, for $\theta \geq 1.2$, deviations of the fitted curve increased slightly, along with an increase in the

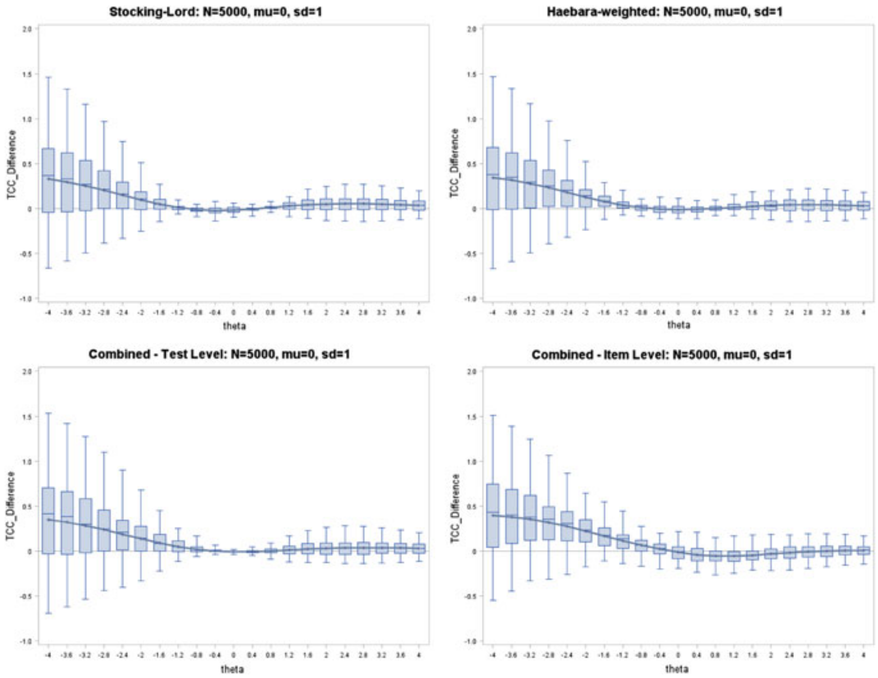


Fig. 2 TCC differences versus theta for baseline experimental condition $\theta \sim N(0, 1)$

variabilities of the outcome measure. The Combined Item-Level methodology yielded results quite different from the other three methodologies, with greater deviations and variabilities. Note that this pattern of TCC differences for Combined Item-Level persisted across experimental conditions.

In Fig. 3, where relative efficiency is the outcome measure, box plots were formatted similarly to Fig. 2, but with the vertical axis reference line indicated for a relative efficiency equal to 1. For all four IRT scaling methodologies, the deviations of the fitted curve from the reference line across all latent trait levels were not as substantial as those observed for the TCC differences; however, variabilities of the relative efficiency measure followed a similar pattern as those observed for TCC differences.

4.2 Results for Other Conditions

Across the nine experimental conditions, a relationship was observed between TCC differences and combinations of the mean and standard deviation of the latent trait generating distribution: the magnitude of the variability in TCC differences increased with increasing means, and the magnitude of this variability increased

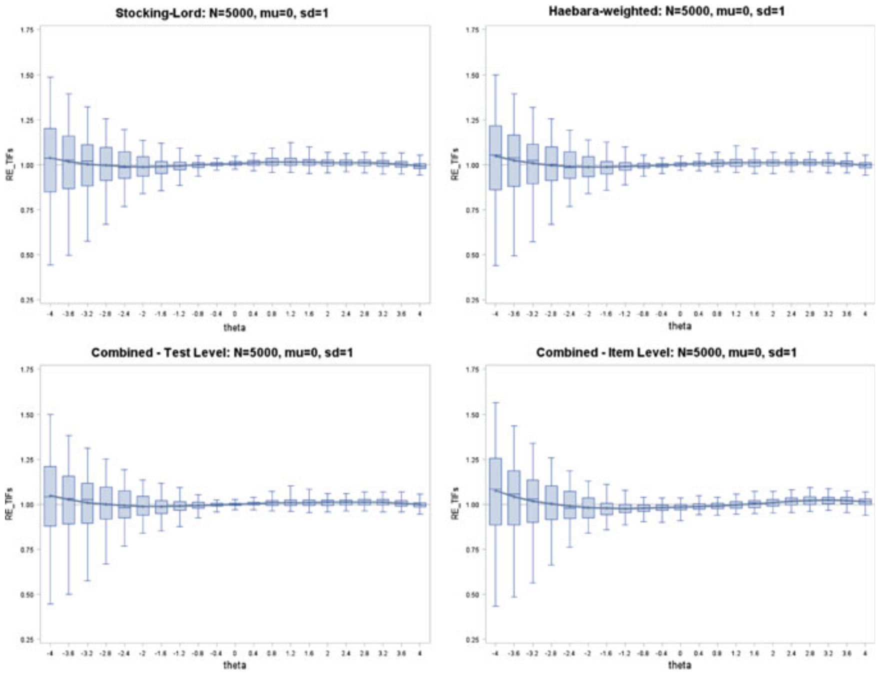


Fig. 3 Relative efficiency versus theta for baseline experimental condition $\theta \sim N(0, 1)$

with decreasing standard deviations. Thus, the experimental condition with the smallest range of TCC differences was where $\theta \sim N(-1,\ 1.2)$, and the condition with the largest range of TCC differences was where $\theta_k \sim N(1, 0.8)$. Due to space constraints, only the plots associated with these two conditions are presented.

Figures 4 and 5 present TCC differences and relative efficiencies, respectively, for the experimental condition $\theta \sim N(-1,\ 1.2)$. As shown in Fig. 4, the fitted curve passing through the mean TCC differences remained close to the zero line, even for $\theta \leq -1.6$, across all IRT scaling methodologies except Combined Item-Level. For the relative efficiencies shown in Fig. 5, the proximity of the fitted curves to the reference line was similar to that observed for the baseline condition, with somewhat greater deviations from the reference line for $\theta \leq -2.8$.

Figures 6 and 7 present TCC differences and relative efficiencies, respectively, for the experimental condition $\theta \sim N(1, 0.8)$. Under this condition, deviations of the fitted curves from their respective reference lines were substantially larger than those observed for the baseline and $\theta \sim N(-1,\ 1.2)$ conditions. These deviations were most striking for the relative efficiencies; unlike the baseline and $\theta \sim N(-1,\ 1.2)$ condition, where the fitted curve was in close proximity (or

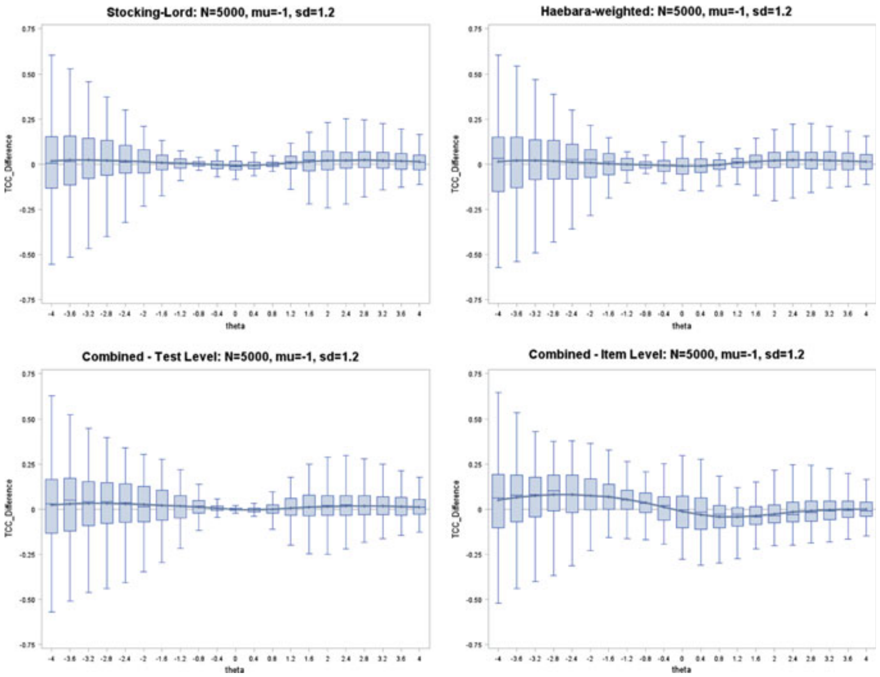


Fig. 4 TCC differences versus theta for experimental condition $\theta \sim N(-1,\ 1.2)$

overlapped) the reference line, in this condition the fitted curve wanders below and above the reference line, crossing it twice: once in the neighborhood of $\theta = 0$, and again in the region $2.8 \leq \theta \leq 4$.

When comparing the results of the four scaling methodologies across the three experimental conditions presented here, some interesting patterns and trends emerge. First, the results from the Combined Item-Level methodology were quite different from those for the other three methodologies, particularly for the TCC differences. Second, variability of the TCC differences and relative efficiencies were in general smallest in the neighborhood of $\theta = 0$ for Combined Test-Level as compared to the other scaling methodologies. Although not shown here, this pattern held across all experimental conditions for the Combined Test-Level methodology. Third, results for relative efficiencies from the Combined Test-Level methodology resembled those from Haebara more than Stocking-Lord, except possibly in the $\theta \sim N(-1,\ 1.2)$ condition, where these three methodologies yielded similar results. For TCC differences, the Combined Test-Level methodology resembled those from Stocking-Lord more than Haebara, except for the $\theta \sim N(1, 0.8)$ condition, where results from Combined Test-Level and Haebara were more similar.

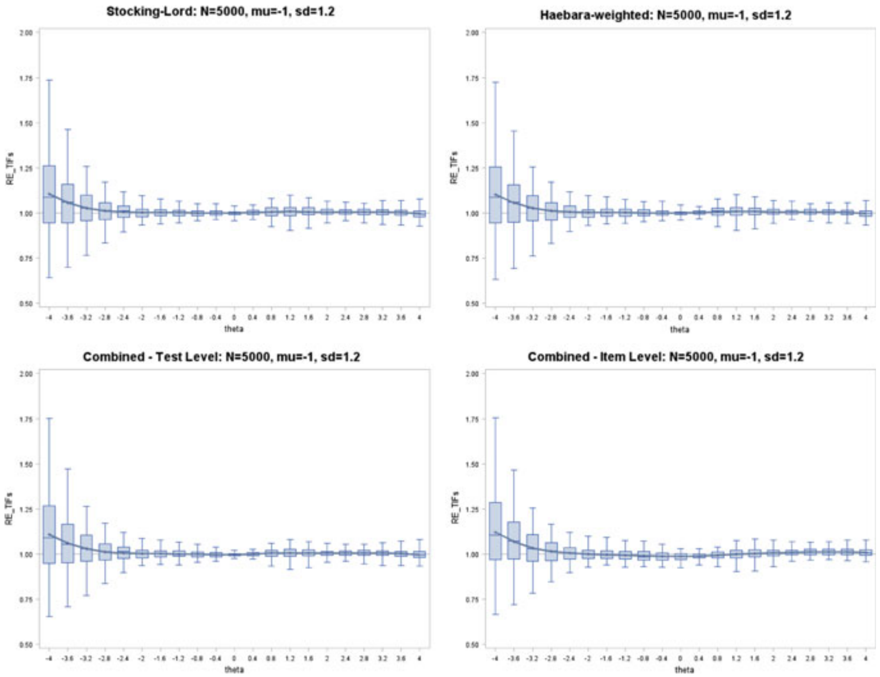


Fig. 5 Relative efficiency versus theta for experimental condition $\theta \sim N(-1, 1.2)$

5 Discussion

This study investigated how information functions could be incorporated into IRT characteristic curve scaling by introducing two new scaling methodologies, each based on the existing characteristic curve methodologies of Haebara (1980) or Stocking and Lord (1983). The Combined Item-Level methodology incorporated Haebara's (1980) method with item information functions; the Combined Test-Level methodology incorporated Stocking and Lord's (1983) method with test information functions.

Across the four scaling methodologies, the Combined Item-Level method yielded results that were noticeably different from the other scaling methodologies, particularly with respect to TCC differences. A possible explanation for this result is that minimizing the differences in item information functions came at the expense of minimizing differences in the item characteristic curves. Such an explanation would be supported by the comparatively more stable relative efficiency measures for this method.

One somewhat surprising result was that the variability in TCC differences observed for the Combined Test-Level methodology was consistently smaller than that observed for the other scaling methodologies in the neighborhood of $\theta = 0$. (Although not presented here, this pattern held across all nine experimental

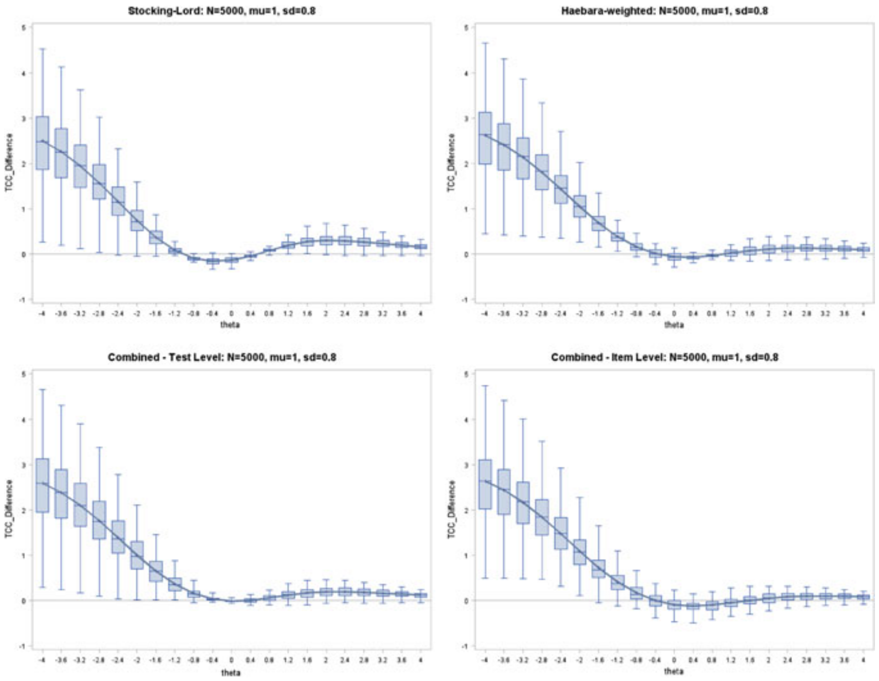


Fig. 6 TCC differences versus theta for experimental condition $\theta \sim N(1, 0.8)$

conditions.) One potential explanation concerns the density function $f(\theta_k)$ assigned to the latent trait levels evaluated in the optimizations (see (22)); this function was chosen to correspond to the Normal probability density function. Since the maximum of this function occurs at $\theta = 0$, more weight was assigned to that latent trait level in the optimizations (see (6), (9), (14), and (19)). However, this explanation is difficult to support for two reasons: (1) the same probability density function was assigned to all four scaling methodologies, yet this effect was observed only for Combined Test-Level; and (2) in follow-up studies, a uniform probability density function was chosen instead, and similar results were obtained. An alternative explanation is that since the test information reaches its maximum in the neighborhood of $\theta = 0$ (see Fig. 1), incorporating this information with the test characteristic curve contributes to the reduced variabilities in TCC differences observed in that region of the latent trait scale.

This study was designed to reduce sources of error. While this approach is advantageous from an experimental design perspective, it does present some limitations. For example, the new and reference forms contained identical items, so both forms had all items in common. Thus, parameter estimation of the new and reference form item parameters focused solely on the common items; in practice, it is rarely the case that new and reference forms will share all items in common. Further, all estimated new form item parameters were scaled to the reference form

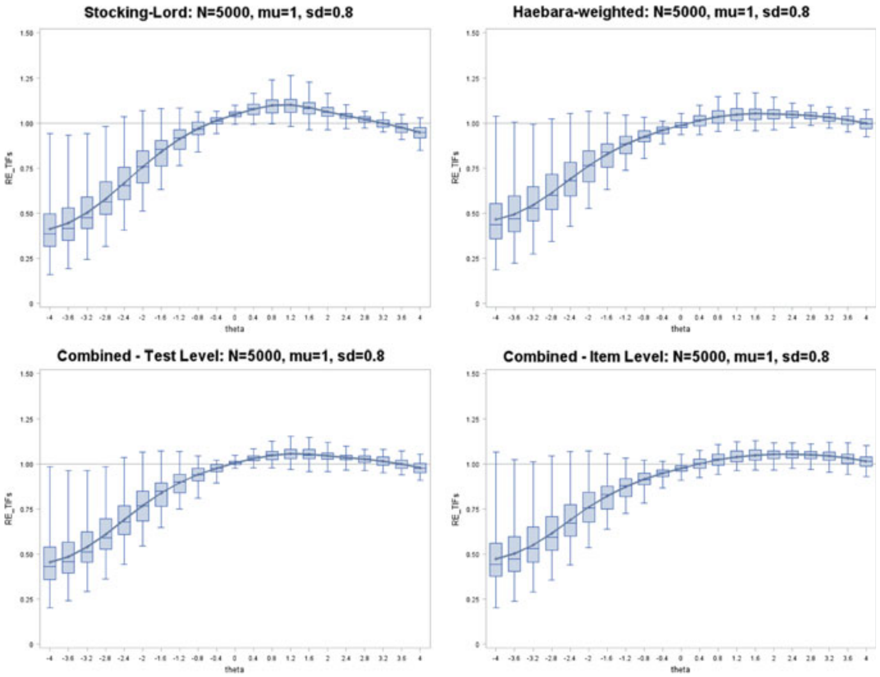


Fig. 7 Relative efficiency versus theta for experimental condition $\theta \sim N(1, 0.8)$

item parameters. This approach minimized scaling error, since the best set of reference form item parameters (the 'true' or generating parameters) were utilized. In practice, however, true item parameters are never known. To minimize parameter estimation error over replications, latent trait generating distributions across the experimental conditions were all selected to be normal, and sample sizes for each replication were fixed at 5000. Thus, sources of error in this study were mostly contained to replication error (from item response simulation) and item parameter estimation.

Future studies might expand upon the design of this study. Possibilities include: (1) assembling reference and new forms such that they do not share all items in common; (2) simulating item responses not only for new form administrations, but also for the reference form administrations, then estimating item parameters for both forms for each replication and conducting IRT scaling on those parameter estimates; (3) varying sample sizes for replications, particularly for smaller sample sizes that could be encountered in practice; (4) examining other IRT models, such as the 2PL model or polytomous IRT models; and (5) applying different methods for assigning latent trait levels or probability density functions in the IRT scaling optimizations.

References

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley Pub. Co.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: Springer.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

SAS Institute Inc. (2012). *SAS/STAT and SAS/OR [computer software]*. NC: Cary.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 26,* 261–271.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models [computer software]*. Chicago, IL: Scientific Software International Inc.

Reducing Conditional Error Variance Differences in IRT Scaling

Tammy J. Trierweiler, Charles Lewis and Robert L. Smith

Abstract The performance of a Hybrid scaling method that takes into account the differences between test characteristic curves as well as differences between conditional error variances when estimating transformation constants is proposed and evaluated. Results are evaluated and discussed in relation to the Stocking-Lord method. Findings using a Monte Carlo simulation approach suggest that when the two forms being scaled are parallel, the Hybrid method and the Stocking-Lord test characteristic curve method lead to similar results. However, when the forms being scaled have similar test characteristic curves but different conditional error variances, the Hybrid method does better near the mean of the ability distribution, especially for the test information function.

Keywords Item response theory (IRT) • Scaling • Equating
Test characteristic curve • Conditional error variance

1 Introduction

In item response theory (IRT), when item parameters for a given pair of forms (referred to here as the reference and new forms) whose items measure the same trait are independently estimated using data obtained from two different groups of test takers, the estimates will generally be on two different IRT scales. For the two

T. J. Trierweiler (✉)
Law School Admission Council, Newtown, PA 18940, USA
e-mail: ttrierweiler@lsac.org

C. Lewis
Fordham University, Bronx, NY 10458, USA
e-mail: clewis@fordham.edu

R. L. Smith
Smith Consulting, New Hope, PA 18938, USA
e-mail: rls1788rlews1@gmail.com

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_13

forms to be compared and used interchangeably, these estimates need to be placed on a common scale.

When working with the two-parameter logistic (2PL; Birnbaum 1968) IRT model for binary data, with the item response function written as

$$P(\theta; a, b) = \frac{\exp\,[Da(\theta - b)]}{1 + \exp\,[Da(\theta - b)]}, \tag{1}$$

(where D is a scaling constant equal to 1.702), Lord (1980) showed that the relationship between the scales of any two test calibrations is linear. Specifically,

$$\theta^* = A\theta + B, \tag{2}$$

where the slope A and intercept B are the linking coefficients of the linear function, θ is the trait value on the scale of the new form, and θ^* is the corresponding trait value on the scale of the reference form. Given the relationship in Eq. 2, the parameters for a given item from separate calibrations are linearly related as follows:

$$a^* = \frac{a}{A} \quad \text{and} \quad b^* = Ab + B\,, \tag{3}$$

where a^* and b^* are item parameters expressed on the reference form scale, and a and b are the corresponding item parameters expressed on the new form scale.

1.1 Functions Used in This Investigation

There are three functions that will be considered in this investigation, namely the test characteristic curve (TCC), the conditional error variance (CEV), and the test information function (TIF).[1] For a test consisting of J items, and working with the 2PL IRT model, they are defined as follows:

$$TCC(\theta) = \sum_{j=1}^{J} P\left(\theta; a_j, b_j\right), \tag{4}$$

$$CEV(\theta) = \sum_{j=1}^{J} \left\{P\left(\theta; a_j, b_j\right)\left[1 - P\left(\theta; a_j, b_j\right)\right]\right\}, \tag{5}$$

and

[1] These abbreviations are used to refer to both singular and plural forms of the names. Thus, TCC is used to refer to both "test characteristic curve" and "test characteristic curves," depending on the context.

$$TIF(\theta) = \sum_{j=1}^{J} \left\{ D^2 a_j^2 P(\theta; a_j, b_j) \left[1 - P(\theta; a_j, b_j) \right] \right\}. \tag{6}$$

Let X denote the number-correct observed score for the J-item test, with the classical test theory (CTT) expression giving the observed score as the sum of true and error scores:

$$X = T + E. \tag{7}$$

The TCC gives the number-correct true score T for the test and the CEV gives the number-correct $Var(E|T)$ for the test. These two functions thus provide a link between classical test theory and item response theory.

The TIF is a function that has no interpretation in CTT. It is strictly an IRT expression. If θ denotes the maximum likelihood estimate of θ, its (asymptotic) sampling variance is given by

$$Var(\hat{\theta}|\theta) = \frac{1}{TIF(\theta)}. \tag{8}$$

It is instructive to compare the CEV and TIF for the special case where all $a_j = a$ (i.e., all item slopes are equal). In this case the relationship between two functions may be written as

$$TIF(\theta) = (Da)^2 CEV(\theta). \tag{9}$$

This expression says that the location on the trait scale where the CEV is the greatest is also the location where the test information is the greatest, or where the sampling variance of θ is the smallest. Expressed in terms of error variances, Eq. 9 may be rewritten as

$$\frac{1}{Var(\hat{\theta}|\theta)} = (Da)^2 Var(X|\theta). \tag{10}$$

This emphasizes the point that errors of measurement can function very differently depending on the scale being used for the measurement (true score compared to latent trait).

Finally, note that Eqs. 9 and 10 only apply to the case where all the a-parameters are equal. For the general 2PL model, the TIF and CEV functions will be similar in shape but not strictly proportional. Moreover, as will be seen in our studies, the two functions show differential sensitivity to changes in the choice of scaling criterion, so both functions are of interest.

1.2 Stocking-Lord TCC Scaling

A standard approach for placing new form item parameter estimates onto a reference form scale uses a NEAT (non-equivalent groups anchor test) design. Specifically, the two sets of item parameter estimates for the common (anchor) items obtained from calibrating the two forms are used to estimate the A and B transformation constants. Once these transformation constants are estimated using the common items, the estimated constants are then used to place the parameter estimates for the remaining items in the new form onto the reference form scale.

When a test is scored using IRT ability estimates, there is no need to establish a further relationship between the two forms as the abilities for the reference and new form are now considered comparable. However, if a raw score (number-right or formula score) is needed for reporting, the transformed ability value can be used to find corresponding true scores for both forms through the TCC (IRT true score equating).

There are a number of methods that can be used to estimate the A and B transformation constants used in the linear scaling process (e.g., Haebara 1980; Loyd and Hoover 1980; Marco 1977; Stocking and Lord 1983). The most popular of these methods is the Stocking-Lord TCC method (Stocking and Lord 1983).

Suppose there are J common items on the reference and new forms. Denote the reference form item parameter estimates (for the 2PL model) by (a_{1j}, b_{1j}) and the corresponding new form item parameter estimates by (a_{2j}, b_{2j}) for the common items. Also, denote the new form item parameter estimates after transformation using the expressions in Eq. 3 by (a^*_{2j}, b^*_{2j}). The TCC functions for the reference and new forms (after transformation) may be written as $TCC_1(\theta)$ and $TCC^*_2(\theta)$, respectively.

With the Stocking-Lord TCC method, the A and B transformation constants are chosen so that the squared difference between the TCC functions after transformation, averaged over the values for a suitable group of N test takers with reference scale trait values θ_i, $i = 1, \ldots, N$, is as small as possible. The objective function that is minimized may be written as (Stocking and Lord 1983, Eq. 6):

$$F_{SL} = \frac{1}{N} \sum_{i=1}^{N} \left[TCC_1(\theta_i) - TCC^*_2(\theta_i) \right]^2. \tag{11}$$

Rather than averaging over a finite group of test takers to represent a corresponding population, in the current investigation a discrete approximation to a specified population distribution (in this case, the standard Normal) for the latent trait is used, with quadrature points θ_q and weights w_q for $q = 1, \ldots, Q$. This modification results in a new objective function to be minimized for the Stocking-Lord TCC method:

$$\text{modified}\,F_{SL} = \sum_{q=1}^{Q} \left\{ w_q \left[TCC_1(\theta_q) - TCC_2^*(\theta_q) \right]^2 \right\}. \tag{12}$$

1.3 Illustration of a Scaling Issue

The two primary target criteria used in test construction are form difficulty and reliability. In an IRT context the TCC represents form difficulty and the CEV and TIF represent reliability. The Stocking-Lord method only uses the TCC. Consequently, there may be systematic variation that cannot be removed with a linear transformation.

As a simple illustration of this point, Table 1 presents item parameters for a 2PL model for both a reference and new form, each composed of two test items. Figure 1 presents the corresponding TCC and CEV functions for the two tests. In this case, the *a*- and *b*-parameters for the two forms result in similar TCC but different CEV.

Since the *b*-parameters for the new form have a greater range than those from the reference form, to match the two TCC, the resulting slope transformation constant (*A*) would need to be equal to 0.91. However, since the *a*-parameters of the new form are steeper than those of the reference form, a slope transformation constant of 1.05 would be required to match them. In this illustration, there is no linear transformation that would transform the new form onto the reference form scale perfectly to match both TCC and CEV functions.

Table 1 Item parameters for two 2-item tests

	Reference form		New form	
	a	*b*	*a*	*b*
Item 1	1.00	−1.00	1.05	−1.10
Item 2	1.00	1.00	1.05	1.10

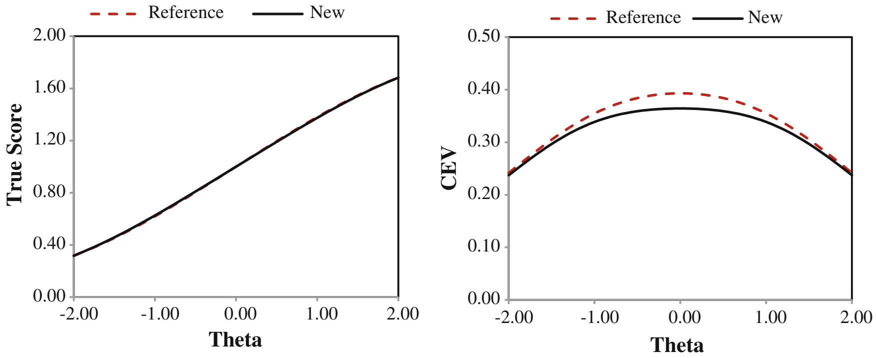


Fig. 1 TCC and CEV functions for reference and new form illustration

1.4 Proposed Hybrid and Weighted Hybrid Scaling Methods

Theoretically, assuming the IRT model holds, the reference form and new form item response functions for a specific common item after transformation should be identical. However, in practice, there will always be systematic variation that cannot be removed through a linear transformation. By including the CEV function in the transformation process, more information can be accounted for if there are systematic variations between forms in either the TCC or the CEV.

As can be seen from Eqs. 11 and 12, the Stocking-Lord method does not consider any information in the scaling process other than the TCC. Extending the work of Stocking and Lord (1983), we present a Hybrid method that takes into account both the TCC and the precision of the two forms being scaled.

As was done above for the TCC functions, the CEV functions for the reference and new forms (after transformation) may be written as $CEV_1(\theta)$ and $CEV_2^*(\theta)$, respectively. Using both the true score information (TCC) and the precision information (CEV), we may write a Hybrid function to be minimized to find the A and B transformation constants as

$$F_{Hybrid} = \sum_{q=1}^{Q} w_q \left\{ \left[TCC_1(\theta_q) - TCC_2^*(\theta_q) \right]^2 + \left[CEV_1(\theta_q) - CEV_2^*(\theta_q) \right]^2 \right\}. \quad (13)$$

The A and B transformation constants that minimize the quantity given in Eq. 13 would include information associated with both the TCC and the CEV or, more specifically, would take into account both the true score and the reliability of the tests.

An important feature of Eq. 13 is that it can be modified to incorporate weights that may be specified for each component:

$$F_{WtHybrid} = \sum_{q=1}^{Q} w_q \left\{ \lambda_T \left[TCC_1(\theta_q) - TCC_2^*(\theta_q) \right]^2 + \lambda_C \left[CEV_1(\theta_q) - CEV_2^*(\theta_q) \right]^2 \right\} \quad (14)$$

where λ_T is the weight is associated with the TCC component, and λ_C is the weight associated with the CEV component.

1.5 Current Investigation

Three studies are presented below. In the first study, the Hybrid scaling method is evaluated against the Stocking-Lord method under the assumption that the reference form and new form are parallel. In the second study, the Hybrid scaling method and the Stocking-Lord method are evaluated against each other when the two forms have similar TCC but different CEV. In the third study, the Weighted Hybrid model is evaluated, varying the weights on the two components (i.e., TCC and CEV) included in the Hybrid scaling method.

2 Method

For each of the three studies presented below, the general methodology is as follows:

1. Reference form item parameters are fixed for all three studies and are considered to be known true parameters when performing the scaling for each study;
2. New form item parameters (given the specific study design presented below) are used to generate item responses for 3,000 test takers across 200 replications using a generating distribution of ability taken to be $N(0, 1)$;
3. For each replication, item response data for the new form are calibrated using the 2PL model in BILOG-MG (Zimowski et al. 1996);
4. Scaling is performed for each replication, where the new form item parameter estimates are scaled to the reference form item parameters. All scaling methods are written and conducted in SAS/OR® 9.4 (SAS Institute, Cary NC).

2.1 Item Parameters

For the studies presented below, the reference form item parameters are taken from a 50 item multiple-choice test, calibrated using the 2PL model. These item parameters are considered the 'generating item parameters' for the reference form. Twenty-three of the items are considered common items and are used in the scaling process as required by the NEAT design.

2.2 Evaluation of Results

Evaluation of results for each of the three studies is done several ways. First, an unweighted mean squared deviation (MSD) statistic is computed to quantify the closeness between the reference form generating parameters and the transformed new form estimated item parameters averaged across replications. The value of the MSD statistic averaged across replications is given by

$$MSD_{\kappa} = \frac{1}{RQ}\sum_{r=1}^{R}\sum_{q=1}^{Q}\left[\kappa_{2,r}\left(\theta_q\right) - \kappa_{1,r}\left(\theta_q\right)\right]^2. \tag{15}$$

This MSD statistic is computed for the TCC, the CEV, and the TIF, and is an unweighted mean measure of the difference between the functions at each quadrature point (θ_q). In Eq. 15, κ can represent the TCC, CEV or TIF, R is the number of replications and Q is the number of quadrature points at which each of the functions is evaluated.

Additionally, the TCC, CEV and TIF functions are graphically presented to illustrate visually the differences found between these functions and each corresponding scaling method.

3 Study 1

3.1 Methods

As described above, the reference form item parameters are used to specify the reference form. In this study, these same reference form parameters are used to generate data for the "new" form over replications. For each replication, item response data are calibrated using the 2PL logistic model and each of the scaling method analyses is conducted.

3.2 Results

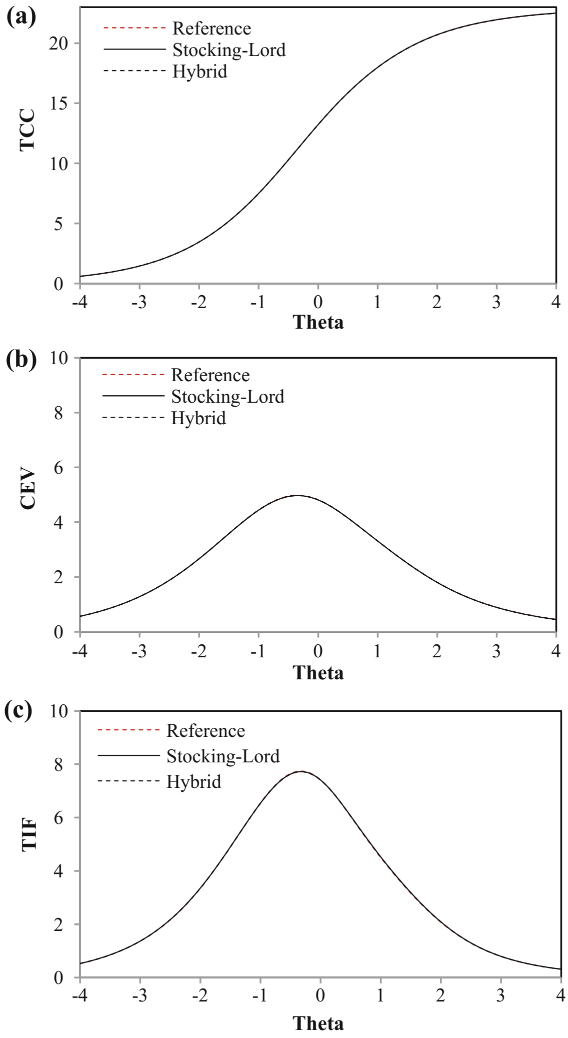
Table 2 presents the average mean squared deviations averaged across replications between the reference and new form TCC, CEV, and TIF functions for each scaling method. Both scaling methods appear to capture the transformed parameters almost perfectly.

Figure 2a–c graphically present the TCC, CEV, and TIF for the reference and for the transformed new form to illustrate how similar the scaling methods are under the condition when the reference and new forms are parallel. Note that the three curves in each figure essentially lie on top of each other and cannot be distinguished.

Table 2 Average mean squared deviations for the TCC, CEV and TIF functions between the reference and new form for each scaling method

Scaling method	TCC	CEV	TIF
Stocking-Lord	0.001	0.000	0.003
Hybrid	0.001	0.000	0.003

Fig. 2 **a** TCC, **b** CEV and **c** TIF functions, averaged across replications, for each scaling method



4 Study 2

4.1 Methods

Unlike Study 1 where the reference form parameters were used to simulate new form item responses, new form item parameters for Study 2 were created to produce similar TCC but different CEV.

To achieve this result, new form item parameters were created using nonlinear optimization in Excel Solver, where the method applied to find the new item parameters included constraints so that the resulting new form parameters would be

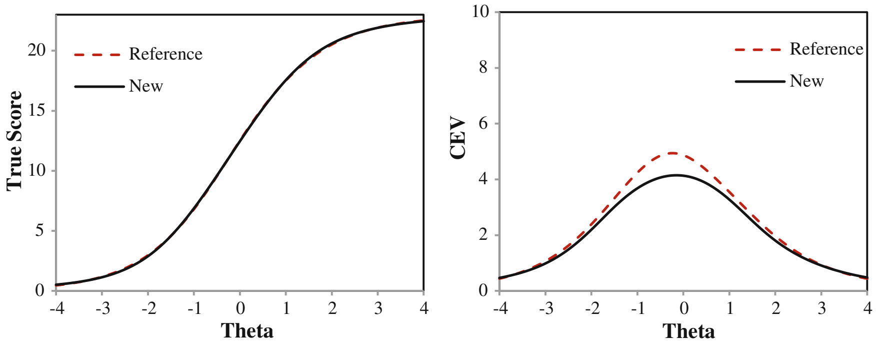


Fig. 3 Reference and new form target TCC and CEV functions

within ±0.25 from the original reference form parameters. These new form parameters created using this optimization process are used as the generating parameters for the new form in Study 2.

The reference and new form TCC and CEV for the common item generating parameters (for both the new and reference forms) are shown in Fig. 3, where the reference form curves are indicated by the dashed lines and the new form curves are indicated by the solid lines.

Calibration and scaling in Study 2 are conducted following the same procedures used in Study 1.

4.2 *Results*

Table 3 presents the average mean squared deviations across replications between the reference and new form TCC, CEV, and TIF for each scaling method. Unlike Study 1, where the mean squared deviations are similar (and close to zero) regardless of scaling method, results for Study 2 show that there are some distinct differences between these methods. First, the Stocking-Lord and Hybrid methods appear to capture the TCC very similarly. This is not totally unexpected as the generating item parameters for the reference and new form were created to produce similar TCC. However, results show that the Hybrid model appears to capture the reference form slightly better than the Stocking-Lord method for the CEV and somewhat better than the Stocking-Lord method for the TIF as well.

Table 3 Average mean squared deviations across replications between the reference form and the new form TCC, CEV, and TIF functions for each scaling method

Scaling method	TCC	CEV	TIF
Stocking-Lord	0.005	0.127	1.010
Hybrid	0.006	0.117	0.723

Figure 4a–c graphically present the reference and transformed new form TCC, CEV, and TIF to further illustrate the performance of each of the scaling methods.

The findings regarding the TCC are consistent with the results in Table 3. The TCC in Fig. 4a show that the Stocking-Lord and the Hybrid methods both appear to capture the reference TCC very well. What can be seen from Fig. 4b is that the Stocking-Lord and the Hybrid scaling methods are very similar to each other in capturing the average CEV across replications. Although the two methods produce similar results, they both differ from the reference form CEV. This reflects the fact that the new form was created to have this discrepancy from the reference form in the CEV. However, as can be seen in Fig. 4c, the Hybrid method captures

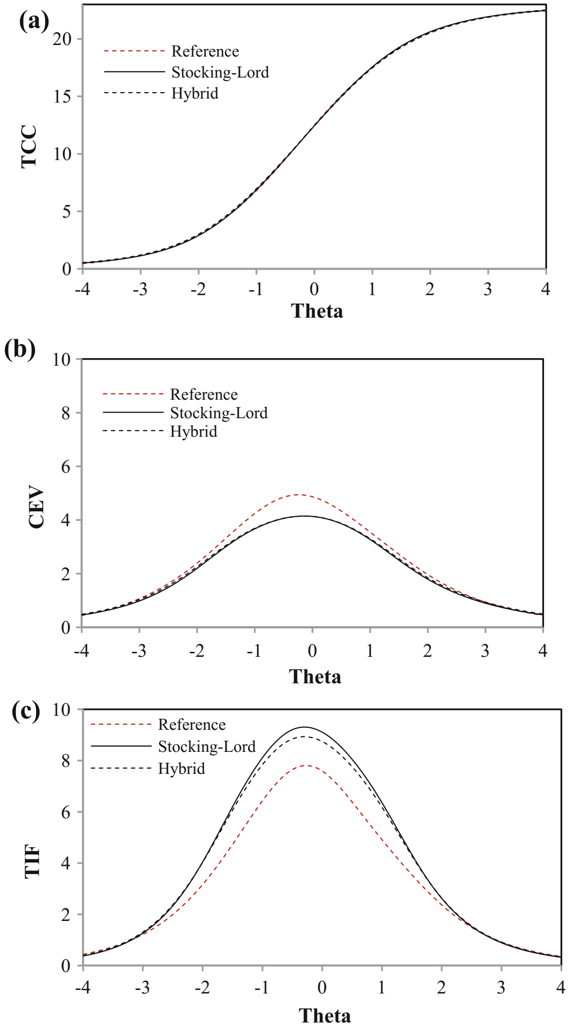


Fig. 4 **a** TCC, **b** CEV and **c** TIF functions, averaged across replications, for each scaling method

the TIF somewhat better than the TCC method near the mean of the ability distribution.

5 Study 3

5.1 Methods

The reference form and new form generating parameters are those used in Study 2, and new form item response data are simulated and calibrated in the same way as for the previous studies.

Different weights are systematically applied to the TCC and CEV components in the scaling process, where the item parameter estimates for the new form are scaled back to the reference form generating parameters for each combination of the 6 pairs of λ weights presented in Table 4 for each replication.

5.2 Results

Table 5 presents the average mean squared deviations across replications between the reference and new form TCC, CEV, and TIF for each of the Weighted Hybrid scaling methods. As might be expected, when more weight is given to one of the

Table 4 Weights for the TCC and CEV components in the Weighted Hybrid model

	λ_T	λ_C
More weight to the TCC component	0.80	0.20
	0.70	0.30
	0.60	0.40
More weight to the CEV component	0.40	0.60
	0.30	0.70
	0.20	0.80

Table 5 Average mean squared deviations across replications between the reference and new form TCC, CEV, and TIF functions for each set of weights

	λ_T	λ_C	TCC	CEV	TIF
More weight to the TCC component	0.80	0.20	0.004	0.124	0.923
	0.70	0.30	0.004	0.122	0.868
	0.60	0.40	0.005	0.119	0.803
More weight to the CEV component	0.40	0.60	0.010	0.114	0.625
	0.30	0.70	0.019	0.110	0.503
	0.20	0.80	0.041	0.107	0.353

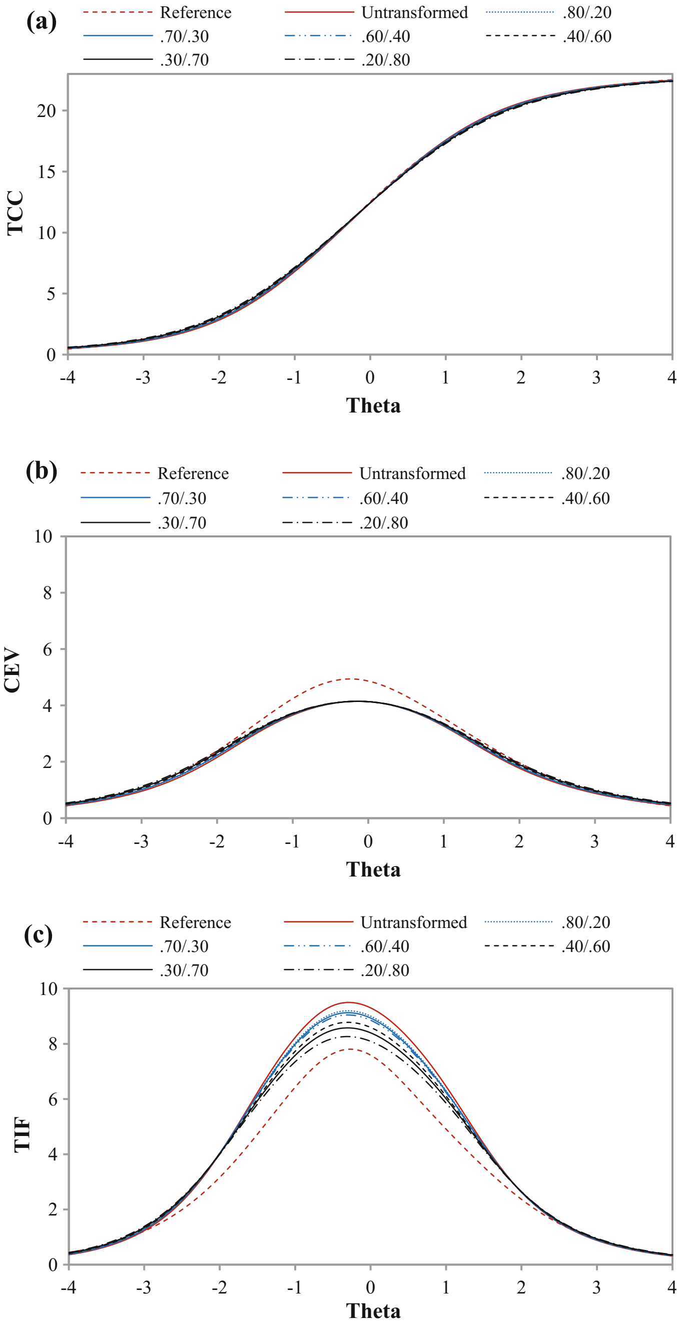


Fig. 5 **a** TCC, **b** CEV and **c** TIF functions, averaged across replications, for each Weighted Hybrid method

components (CEV or TCC), the corresponding MSD becomes smaller. Also, when the weight is increased on the CEV component, the mean squared deviation for the TIF between forms decreases substantially.

Figure 5a–c graphically present the reference and transformed new form TCC, CEV and TIF functions to illustrate the similarities and differences in performance for each of the Weighted Hybrid scaling methods evaluated in Study 3.

As can be seen from Fig. 5a, regardless of the weights applied to the TCC or CEV, the TCC functions for the new and reference form are very similar. Even when weighting the CEV component 80% in the scaling process, the differences between the TCCs are very small.

Figure 5b shows that there is an impact of the different weighting methods on the CEV component, especially in the in the −2.5 to −1.0 and 1.0 to 2.5 ability ranges. In these ranges, the resulting transformed CEV becomes closer to the reference form CEV under the conditions when more weight is applied to the CEV component in the scaling.

The impact of the weighting is even more evident when looking at the TIF function. When more weight is applied to the CEV component, as can be seen in Fig. 5c, the TIF function becomes more similar to the reference form TIF function.

6 General Discussion

Results evaluating a Hybrid scaling method proposed in this study against the popular Stocking-Lord scaling method indicate that, when forms are parallel, the scaling methods perform almost identically. However, when the TCC for the forms are similar but the CEV are different, the Hybrid scaling method does somewhat better at capturing both the CEV and the TIF across the ability distribution.

A Weighted Hybrid scaling method is also evaluated. Results indicate that in general, giving more weight to the CEV component in the scaling process results in a better transformation of both the CEV and TIF when compared to the reference form. Overall, both the CEV and the TIF appear to be relatively sensitive to the weights applied in the scaling process. However, the TCC is quite robust to the weighting of the components, especially in the middle of the ability distribution.

Results presented here suggest that it may be advantageous in the scaling process to use information about the precision as well as the true scores of the forms being scaled. Further research should evaluate the impact of using the Hybrid and Weighted Hybrid scaling methods on overall equating results as well as the performance of these methods under different distributional conditions.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Haebara, T. (1980). *Equating logistic ability scales by a weighted least squares method* (Iowa Testing Programs Occasional Papers, No. ITPOP27). Iowa City, IA: University of Iowa, Iowa Testing Programs.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

SAS Institute Inc. (2014). *SAS/OR® 13.2 user's guide: Mathematical programming*. Cary, NC: SAS Institute Inc.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer program]. Chicago, IL: Scientific Software International.

An IRT Analysis of the Growth Mindset Scale

Brooke Midkiff, Michelle Langer, Cynthia Demetriou and A. T. Panter

Abstract Growth mindset has gained popularity in the fields of psychology and education, yet there is surprisingly little research on the psychometric properties of the Growth Mindset Scale. This research presents an item response theory analysis of the Growth Mindset Scale when used among college students in the United States. Growth Mindset is the belief that success comes through hard work and effort rather than fixed intelligence. Having a growth mindset is believed to be important for academic success among historically marginalized groups; therefore it is important to know if the Growth Mindset Scale functions well among first generation college students. The sample consists of 1260 individuals who completed the Growth Mindset Scale on one of 5 surveys. The Growth Mindset Scale consists of 8 items, with responses ranging from *strongly disagree* (1) to *strongly agree* (5). IRT analysis is used to assess item fit, scale dimensionality, local dependence, and differential item functioning (DIF). Due to local dependence within the 8-item scale, the final IRT model fit 4 items to a unidimensional model. The 4-item scale did not exhibit any local dependence or DIF among known groups within the sample. The 4-item scale also had high marginal reliability (0.90) and high total information. Cronbach's alpha for the 4-item scale was $\alpha = 0.89$. Discussion of the local dependence issues within the 8-item scale is provided.

Keywords Local dependence • First generation college students
FGCSs • Implicit theories of intelligence • Personality

B. Midkiff (✉) · A. T. Panter
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: bmidkiff@email.unc.edu

A. T. Panter
e-mail: panter@email.unc.edu

M. Langer
American Institutes for Research, Chapel Hill, NC, USA
e-mail: michelle.marie.langer@gmail.com

C. Demetriou
University of Arizona, Tucson, AZ, USA
e-mail: cpd@email.arizona.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_14

1 Introduction

The aim of this research is to provide a first item response theory (IRT) examination of the 8-item Growth Mindset Scale. While some evidence of reliability is available for shorter versions of the scale, little psychometric research has been done on the more commonly used form.

The extant literature on the reliability and validity of the Growth Mindset Scale is varied, largely due to the variety of versions of the scale that have been used in research. In some of Dweck's earliest work (1995), what is now commonly known as growth mindset was measured by three items that compromised a scale for the measurement of implicit theories of intelligence. Dweck et al. (1995, p. 269) report that only three items were used because "implicit theory is a construct with a unitary theme, and repeatedly rephrasing the same idea may lead to confusion and boredom on the part of the respondents." They report data from six validation studies showing high internal reliability ($\alpha = 0.94$–0.98), as well as a test-retest reliability of 0.80 over a 2-week interval. Using factor analysis, Dweck et al. (1995, p. 269) demonstrate that the implicit theory of intelligence is a separate construct from other implicit theories (they also tested implicit morality and world theories), and that endorsement of the implicit theory items does not constitute an acquiescence set. Lastly, Dweck et al. (1995) present evidence that the 3-item measure of implicit theories of intelligence is unrelated to cognitive ability, confidence in intellectual ability, self-esteem, optimism or confidence in other people and the world, social-political attitudes such as authoritarianism, or political conservatism or liberalism. The three items on this early instrument are: (a) "You have a certain amount of intelligence and you really can't do much to change it"; (b) "Your intelligence is something about you that you can't change very much"; and (c) "You can learn new things, but you can't really change your basic intelligence."

Dweck (1999) includes an 8-item scale that uses the 3 items from the original Implicit Theories of Intelligence scale, published shortly after the previous study. Of the 8 items on the new scale, 4 are marked to indicate that they can be used separately; these 4 items include the original 3 with the addition of, "To be honest, you can't really change how intelligent you are" (Dweck, 1999, p. 178). Dweck cites the earlier article along with Levy et al. (1998), Levy and Dweck (1999), Erdley and Dweck (1993), and Erdley et al. (1997) as evidence of the reliability and validity of the scale provided in the book. However, evidence of reliability and validity within these studies is varied.

First, Levy et al. (1998) report a high reliability ($\alpha = 0.93$), but used the domain general measure of implicit theories rather than the domain-specific measure of implicit theory of intelligence. Next, Levy and Dweck (1999) use a newly created measure that includes both entity and incremental items and is designed for use with children; they report reliability of $\alpha = 0.62$ and a test-retest reliability of $r = 0.70$ over a 1 week period. Erdley and Dweck (1993) report reliability at $\alpha = 0.71$ with test-retest reliability at $r = 0.64$ over a 1 week period, but they use the Implicit Personality Theory Questionnaire—Others Form for children. Lastly, Erdley et al.

(1997) report reliability at α = 0.75, but use the children's form of the 3-item Implicit Personality Theory Questionnaire—Self-form.

More recent work includes a study by Degol et al. (2017) that finds a significant relationship between growth mindset and task values in mathematics. However, the study uses a single item to measure growth mindset, taken from the U.S. Educational Longitudinal Study of 2002 (ELS) database. The item is, "Most people can learn to be good at math."

Karwowski (2014) developed a 10-item Creative Mindset Scale adapting the items from previous growth mindset scales. Karwowski (2014) demonstrates through exploratory factor analysis, confirmatory factor analysis, and an IRT Rasch model the psychometric properties of the Creative Mindset Scale. The psychometric evidence shows that the scale represents two separate factors—fixed and malleable mindsets—rather than one factor conceptualized as two ends of a continuum. However, IRT parameters are not provided in the published study.

2 Methods

This research used the 8-item version of the Growth Mindset Scale, originally published by Dweck (1999) as the "Theories of Intelligence Scale—Self Form For Adults" (p. 178). Response options differed slightly from the published scale which ranges from (1) Strongly Agree to (6) Strongly Disagree. Response options used in this study were (1) Strongly Disagree, (2) Disagree, (3) Neither Agree nor Disagree, (4) Agree, and (5) Strongly Agree. The response options used in this research were consistent across the five surveys from which the sample are drawn.

2.1 Sample

The sample consists of 1,260 college students who completed the Growth Mindset Scale on one of five surveys administered in research projects within The Finish Line Project ("Finish Line Overview", 2017). Of the 1,260 participants, 691 were first in their family to attend college (first generation college students; FGCSs), 549 were non-FGCSs, 273 were currently enrolled, and 987 were recent college graduates.

2.2 Reliability Analysis

Corrected item-total correlations and Cronbach's alpha (1951) were examined to assess the reliability of the scale. These analyses were repeated iteratively after fitting IRT models. Internal consistency was evaluated by Cronbach's alpha using the software Mplus (Muthén and Muthén 1998). Alpha values of 0.70 or greater was used as an acceptable minimum for group-level assessment (Cronbach 1951).

2.3 IRT

IRT analysis was conducted using Samejima's (2010) graded response model (GRM) for polytomous items using IRTPRO3 (Cai et al. 2011). Based on the original literature around the psychometric properties of the scale—that it represents a single latent construct of growth mindset (Dweck 1999)—a unidimensional model for all 8-items was first fit. Subsequent IRT models included a bifactor model and a unidimensional model on a subset of items identified through bifactor IRT analysis and exploratory factor analysis (EFA). The subset of items identified through bifactor and EFA analysis were items 3, 5, 7, and 8 (see Table 3). The EFA used maximum likelihood estimation with orthogonal, varimax rotation, with explained common variance (ECV) > 0.85 indicating unidimensionality. The analyses presented here used only the portion of the sample with no missing data on any of the scale items ($N = 1129$). IRT model fit based on $SS\text{–}X^2$ (Orlando and Thissen 2000, 2003) was assessed examining root mean square error of approximation (RMSEA), wherein adequate fit is 0.05 or less. IRT model fit was also fit through comparisons of -2 log likelihood Akaike Information Criteria (AIC) (Akaike 1974) and Bayesian Information Criteria (BIC) (Schwarz 1978), with lower scores indicating better model fit for both statistics relative to AIC and BIC for compared IRT models.

2.4 Local Dependence

Local dependence was assessed based on the Chen and Thissen (1997) local dependence indices, wherein LD χ^2 values greater than ten suggest significant local dependence. Item wording was also assessed to further investigate underlying possible causes of local dependence.

2.5 Differential Item Functioning

Differential item functioning (DIF) was assessed using the IRT-based Wald test (Langer 2008). DIF was assessed between known groups within the sample including FGCS status, gender, underrepresented minority (URM) status, and current students versus recent graduates.

2.6 Known Groups Validity

The validity of the Growth Mindset Scale was examined by assessing the extent to which it could discriminate between several known groups: FGCS status, gender, URM status, current students versus recent graduates, and URM interacted with gender and FGCSs. We compared summed scores on the final 4-item scale across all groups using a one-way analysis of analysis of variance (ANOVA). Statistical significance was defined at the 0.05 alpha level for evaluation of known groups validity.

2.7 Discriminant Validity

Participants who completed the Growth Mindset Scale also completed the five item Guilt-Proneness Scale (Cohen et al. 2014). To assess discriminant validity, the correlation between the mean item score on the Growth Mindset Scale and the mean item score of the Guilt Proneness Scale was computed. Guilt Proneness response items were (1) "Extremely Unlikely," (2) "Unlikely," (3) "Neither Likely nor Unlikely," (4) "Likely," and (5) "Extremely Likely," and were scored such that higher mean scores reflect more guilt proneness. Statistical significance was defined at the 0.05 alpha level for evaluation of discriminant validity.

3 Results

3.1 Descriptive Statistics

Of the total sample ($N = 1{,}250$), 1,148 participants answered at least one growth mindset scale item; 1.7% (19 students) skipped one or two items. Analysis was conducted only for responses with no missingness. Mean item scores for those with no items missing ($N = 1129$) range from 3.27 to 3.82, suggesting that, on average, participants tended to be neutral or slightly agree with all items on the scale. Descriptive statistics for scale items are given in Table 1. The mean summed score for the 1129 respondents with no missing is 28.69 (SD = 6.56); summed scores presented a negatively skewed distribution.

The demographic information for the sample with no missing scale items ($N = 1129$) is given in Table 2. Known group totals vary due to missing demographic data within the overall sample of non-missing responses on scale items ($N = 1129$) (e.g. there is no missing data within the scale items but some missing data within demographic data).

Table 1 Mean item scores and item-total correlations for growth mindset scale

Item #	Scale item	Mean	SD	Item-total correlation
1	You have a certain amount of intelligence, and you can't really do much to change it	3.75	0.98	0.78
2	Your intelligence is something about you that you can't change very much	3.78	1.01	0.78
3	No matter who you are, you can significantly change your intelligence level	3.63	1.02	0.73
4	To be honest, you can't really change how intelligent you are	3.82	0.98	0.79
5	You can always substantially change how intelligent you are	3.43	1.03	0.72
6	You can learn new things, but you can't really change your basic intelligence	3.27	1.12	0.71
7	No matter how much intelligence you have, you can always change it quite a bit	3.57	0.94	0.75
8	You can change even your basic intelligence level considerably	3.44	1.01	0.70

Table 2 Demographic makeup of sample

Group	*N*	Group	*N*	Totals
First generation	629	Continuing generation	495	1124
Current student	245	Recent graduate	884	1129
Men	358	Women	687	1045
Underrepresented minority	239	Non-underrepresented minority	801	1040

3.2 Reliability

Relevant items were reverse coded so that higher scores reflected more growth mindset. Cronbach's alpha was high (α = 0.93) for the 8-item scale at. Corrected item-total correlations were strong, ranging from 0.70 to 0.79. Item-total correlations are given previously in Table 1. The final 4-item scale also showed strong corrected item-total correlations (given in Table 3), ranging from 0.71–0.79, higher than the 8-item scale. Cronbach's alpha remained high (α = 0.89) for the 4-item scale.

3.3 IRT

GRMs were fit within unidimensional and bifactor frameworks to assess scale dimensionality. The first model fit a unidimensional GRM with all eight items

Table 3 Corrected item-total correlations for 4-item growth mindset scale

Item #	Scale item	Item-total correlation
3	No matter who you are, you can significantly change your intelligence level	0.75
5	You can always substantially change how intelligent you are	0.77
7	No matter how much intelligence you have, you can always change it quite a bit	0.79
8	You can change even your basic intelligence level considerably	0.71

based on the literature that purports the scale measures 1 latent construct. The *SS* $-\chi^2$ item fit was significant for all items in the first unidimensional model that included all 8 items, and local dependence was detected for the following item pairs: 1&2, 1&4, 3&5, 3&7, 3&8, 5&7, 5&8, 6&8, 7&8. The model RMSEA was 0.25, indicating poor fit.

3.4 Local Dependence

In the unidimensional IRT model for the 8-item scale, local dependence was detected for multiple item pairs. Inspection of the wording of the items reveals that items 1, 2, and 4 simply state that you can't change intelligence, whereas the other five items qualify changing intelligence with words such as significantly, considerably, and so on.

3.5 Bifactor GRM

In light of the preponderance of local dependence and poor model fit of the unidimensional GRM, we fit a bifactor GRM model to the 8-items, shown in Table 4. Items 1, 2, and 4 loaded onto one specific factor and items 3, 5, 6, 7, and 8 loaded onto the second specific factor; all items loaded on the overall factor. However, the bifactor model overall model fit was poor: RMSEA = 0.60, SS$-\chi^2$ item fit was significant for all items, and local dependence was detected for the item pair 6 & 8.

The factor loadings suggest two factors, with an ECV of 0.79 (ECV > 0.85 typically indicates unidimensionality). Based on these findings, we conducted an EFA to further investigate the factor structure of the 8-item scale because the first two models fit so poorly and extant literature suggested the scale measured only one latent construct.

The EFA used oblique-varimax rotation and showed 2 factors with each item showing high loadings onto one or the other factor, with the exception of item 6,

Table 4 Factor loadings from bifactor model of 8-item growth mindset scale

Item #	Item	Factor loadings		
		Overall	Specific	Specific
1	You have a certain amount of intelligence, and you can't really do much to change it	0.86	0.44	–
2	Your intelligence is something about you that you can't change very much	0.87	0.42	–
3	No matter who you are, you can significantly change your intelligence level	0.77	–	0.47
4	To be honest, you can't really change how intelligent you are	0.90	0.26	–
5	You can always substantially change how intelligent you are	0.73	–	0.54
6	You can learn new things, but you can't really change your basic intelligence	0.86	–	0.03
7	No matter how much intelligence you have, you can always change it quite a bit	0.75	–	0.54
8	You can change even your basic intelligence level considerably	0.74	–	0.43

"You can learn new things, but you can't really change your basic intelligence." The factor loadings indicate there are both positively-worded and negatively-worded items. Item#6 likely cross-loads ($\lambda_1 = -0.34$, $\lambda_2 = 0.56$) because it has *both* positive ("can learn new things") *and* negative ("can't change basic intelligence") wording. The cross-loading is also evident in the previous bifactor model, which detected high loading for item 6 on the overall factor, but only $\lambda = 0.03$ on one of the specific factors, suggesting that item 6 measures mindset but does not fit with either growth or fixed mindset as a separate construct.

3.6 *Final IRT Model*

Growth mindset is a positive construct, and good measurement practice is to have items worded positively. Therefore, we proceeded with IRT analysis using only items 3, 5, 7, and 8, all of which are positively worded and are statements in congruence with a growth mindset rather than a fixed mindset. This analytic strategy was chosen in light of the previous IRT models and EFA, and the understanding that the scale ultimately *should* measure only one construct. A unidimensional GRM was fit using the items 3, 5, 7, and 8. IRT parameters of this model are given in Table 5.

The final unidimensional IRT model for the 4-item scale, like the 8-item unidimensional and bifactor IRT models, still did not fit well ($SS{-}\chi^2$ showed poor item

Table 5 IRT parameters: 4-item growth mindset scale

Item #	Scale item	*a*	b_1	b_2	b_3	b_4
3	No matter who you are, you can significantly. Change your intelligence level	3.42	−2.21	−1.05	−0.41	0.96
5	You can always substantially change how intelligent you are	3.71	−2.10	−0.88	−0.12	1.16
7	No matter how much intelligence you have, you can always change it quite a bit	4.24	−2.25	−1.07	−0.29	1.20
8	You can change even your basic intelligence level considerably	2.83	−2.14	−1.00	−0.15	1.32

fit and RMSEA = 0.37). However, the 4-item scale did not exhibit any local dependence, making it preferable for research use to the 8-item scale. The shorter, 4-item scale also had high marginal reliability (0.90) and high total information. Test information is high from −2.5 SDs below the mean to 1.5 SDs above the mean. The total information curve is given in Fig. 1.

Differential Item Functioning. DIF analysis of the 4-item scale showed no DIF between the known groups in the sample—FGCSs and non-FGCSs, gender, URM status, and current students versus recent graduates.

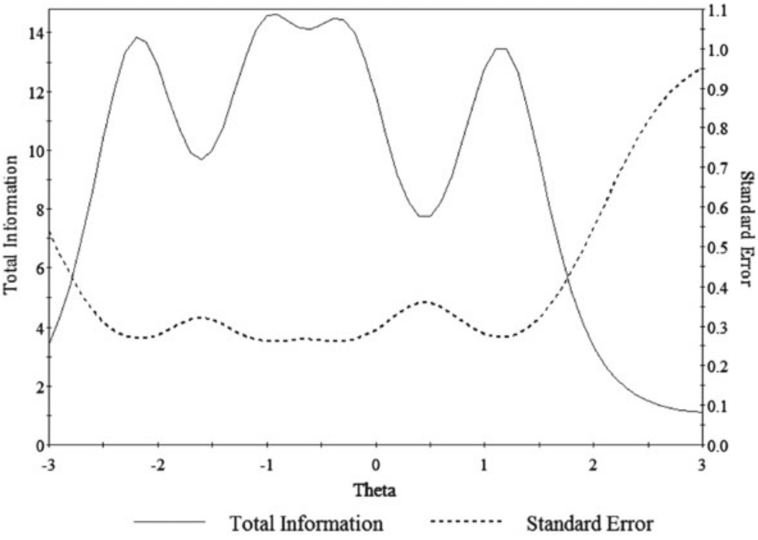


Fig. 1 Total information curve: 4-item growth mindset scale

3.7 Known Groups Validity

One-way ANOVA using summed scores from the 4-item scale was used to assess known groups validity. Significant differences were found between URM students and their majority peers as well as between FGCSs and non-FGCSs. ANOVA results are given in Table 6.

3.8 Discriminant Validity

Of the sample with no missing items on the Growth Mindset Scale, 1020 also completed the Guilt Proneness Scale with no missing items from either scale. The mean item score on the Guilt Proneness Scale was 4.09 (SD = 0.74); the mean item score on the 8-item Growth Mindset Scale was 3.57 (SD = 0.83); the mean item score on the 4-item Growth Mindset Scale was 3.49 (SD = 0.87). The Guilt Proneness Scale was significantly correlated with the 8-item Growth Mindset Scale, but with a small magnitude ($r = 0.14$, $p < 0.05$). The Guilt Proneness Scale was also significantly correlated with the 4-item Growth Mindset Scale, but also with a small magnitude ($r = 0.15$, $p < 0.05$).

3.9 Measuring Fixed Versus Growth Mindset

Interestingly, the remaining items not used in the final IRT model (items 1, 2, 4, and 6) exhibit slightly stronger reliability and higher average inter-item covariance than the 4-item scale containing items 3, 5, 7, and 8, shown in Table 7. Because of the wording of the items, each set of 4 items may be conceptualized as a subscale measuring two factors—growth mindset and fixed mindset.

In fact, in the original published scale, Dweck notes that items 1, 2, 4, and 6 "can be used alone" (Dweck 1999, p. 178). The mean item-scores of both subscales are correlated at $r = 0.72$, $p < 0.01$.

Lastly, the version of the Growth Mindset Scale that is currently used on www.mindsetworks.com ("What's My Mindset?" 2017), founded by Dweck in 2007, uses completely different items with the exception of item 7. On www.

Table 6 ANOVA Results for 4-item growth mindset scale

Groups	Df	Mean square	F	Pr > F
URM	1	218.77	18.27	0.00
FGCS	1	186.66	15.54	0.00
Gender	1	0.00	0.00	0.99
Current student status	1	6.88	0.56	0.45

Table 7 Growth and fixed mindset subscales interitem covariance and reliability

	Growth mindset subscale	Fixed mindset subscale
	(Items 3, 5, 7, and 8)	*(Items 1, 2, 4, and 6)*
Average interitem covariance:	0.68	0.77
Number of items in the scale:	4	4
Scale reliability coefficient:	0.89	0.92

mindsetonline.com ("Test Your Mindset," 2017), another website run by Dweck, the scale contains 16 items, incorporating new items that contain wording around talent as well as intelligence.

4 Discussion

The IRT findings presented here suggest that the use of all 8-items may not be the most efficient way to measure the latent construct of growth mindset. A subset of four items (3, 5, 7, and 8) shows no local dependence or DIF among the known groups in the sample, however, a unidimensional GRM still fit poorly. The items in the final IRT model exhibit good test practices in the use of positive wording. Additional IRT analysis is needed to determine if a unidimensional GRM for the fixed mindset subscale fits better than the 4-item growth mindset subscale. IRT analysis of the two new scales available online to assess their psychometric properties in comparison to the 4-item growth mindset is also recommended. One limitation of this study is that all participants were college students; it is possible that the 8-item scale performs differently in the larger, adult population. Until additional IRT analyses are available, researchers are advised to use items 3, 5, 7, and 8 for the measurement of growth mindset as a single latent construct.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Cai, L., Thissen, D., & DuToit, S. H. C. (2011). *ITPRO 3 (Version 3) [Windows]*. Lincolnwood, IL: Scientific Software International.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Cohen, T. R., Kim, Y., & Panter, A. T. (2014). *The five-item guilt proneness scale (GP-5)* (p. 1). Pittsburgh, PA: Carnegie Mellon University.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555.

Degol, J. L., Wang, M.-T., Zhang, Y., & Allerton, J. (2017). Do growth mindsets in math benefit females? Identifying pathways between gender, mindset, and motivation. *Journal of Youth and Adolescence*. https://doi.org/10.1007/s10964-017-0739-8.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.

Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry, 6*(4), 267–285.

Erdley, C. A., & Dweck, C. S. (1993). Children's implicit personality theories as predictors of their social judgments. *Child Development, 64*(3), 863–878.

Erdley, C. A., Loomis, C. C., Cain, K. M., & Dumas-Hines, F. (1997). Relations among children's social goals, implicit personality theories, and responses to social failure. *Developmental Psychology, 33*(2), 263.

Finish Line Overview. (2017, February 17). Retrieved from http://studentsuccess.unc.edu/finish-line-overview/.

Karwowski, M. (2014). Creative mindsets: Measurement, correlates, consequences. *Psychology of Aesthetics, Creativity, and the Arts, 8*(1), 62.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation (Doctoral dissertation)*. Chapel Hill, NC: University of North Carolina at Chapel Hill.

Levy, S. R., & Dweck, C. S. (1999). The impact of children's static versus dynamic conceptions of people on stereotype formation. *Child Development, 70*(5), 1163–1180. https://doi.org/10.1111/1467-8624.00085.

Levy, S. R., Stroessner, S. J., & Dweck, C. S. (1998). Stereotype formation and endorsement: The role of implicit theories. *Journal of Personality and Social Psychology, 74*(6), 1421.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (Version 7th edition). Los Angeles, CA

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298.

Samejima, F. (2010). The general graded response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 77–108). New York, NY: Routledge.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Test Your Mindset. (2017, December 8). Retrieved from http://mindsetonline.com/testyourmindset/step1.php.

What's My Mindset? (2017, September 30). Retrieved from http://blog.mindsetworks.com/what-s-my-mindset.

Considering Local Dependencies: Person Parameter Estimation for IRT Models of Forced-Choice Data

Safir Yousfi

Abstract The Thurstonian IRT model of Brown and Maydeu-Olivares (Educ Psychol Meas 71:460–502, 2011) was a breakthrough in estimating the structural parameters of IRT models for forced-choice data of arbitrary block size. However, local dependencies of pairwise comparisons within blocks of more than two items are only considered for item parameter estimates, but are explicitly ignored by the proposed methods of person parameter estimation. A general analysis of the likelihood function of binary response indicators (used Brown and Maydeu-Olivares) for arbitrary IRT models of forced-choice questionnaires is presented that reveals that Fisher Information is overestimated by Brown and Maydeu-Olivares' approach of person parameter estimation. Increasing block size beyond 3 leads only to a slight increase measurement precision. Finally, an approach that considers local dependencies within blocks adequately is outlined. It allows for Maximum-Likelihood and Bayesian Modal Estimation and numerical computation of observed Fisher information.

Keywords Forced-choice · Thurstonian IRT model · IRT
Person parameter estimation · Fisher information

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 method (e.g. sum scores), i.e. construct validities, criterion-related valdities and reliabilities are distorted (Brown and Maydeu-Olivares 2013). Recently, Maydeu-Olivares and Brown (2010) proposed an IRT approach to modeling and analyzing forced choice data that effectively overcomes these problems by binary coding and considering local dependencies of

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

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_15

the binary response indicators in the process of estimating the structural model parameters. However, the proposed methods of person parameter estimation explicitly neglect local dependencies of the binary response indicators within a block. Consequently, the respective estimates of person parameters might be flawed. Fisher information might be affected, too. Consequently, recommendations derived from properties of the Fisher information matrix are also called into question.

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.

A (•) indicates that all rows or columns are extracted.

2 Binary Coding of Forced Choice Data

Let $\mathbf{y_b}$ 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, $\mathbf{y_b} = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$ 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 full forced choice questionnaire of K blocks can be described by a sequence of K rankings $\mathbf{Y} := (\mathbf{y}_1, \ldots, \mathbf{y_b}, \ldots, \mathbf{y}_\mathrm{K})$

Let $\mathbf{Y_b}$ 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:

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

Maydeu-Olivares and Brown (2010) referred only to the entries above the diagonal of $\mathbf{Y_b}$ which results in a full description of the data as $y_{pq} = 1 - y_{qp}$.

3 The Likelihood of the Response $\mathbf{y_b}$ to Block b

Let Ω be the set of structural parameters of an arbitrary model for forced-choice questionnaire data (e.g. the item parameters and the parameter of the latent trait distribution of the Thurstonian IRT model established be Maydeu-Olivares and Brown 2010). Let $\boldsymbol{\theta}$ be vector of incidental person parameters (i.e. latent trait values). $l_{\Omega;\boldsymbol{\theta}}(\mathbf{Y})$ denotes the likelihood of the response pattern $\mathbf{Y}$ as a function of Ω and $\boldsymbol{\theta}$. If local (i.e. conditional on $\boldsymbol{\theta}$) stochastic independence of the response to different blocks holds, then $l_{\Omega;\boldsymbol{\theta}}(\mathbf{Y})$ can be decomposed to the product of the likelihood of the responses to the blocks:

$$l_{\Omega;\boldsymbol{\theta}}(\mathbf{Y}) = \prod_{b=1}^{K} l_{\Omega;\boldsymbol{\theta}}(\mathbf{y}_b) = \prod_{b=1}^{K} l_{\Omega;\boldsymbol{\theta}}(\mathbf{Y}_b) \tag{2}$$

For person parameter estimation Maydeu-Olivares and Brown (2010) explicitly neglected local dependencies of the binary response indicator within blocks which results in:

$$l_{\Omega;\boldsymbol{\theta}}(\mathbf{Y}_b) \cong l_{\mathbf{U}_{\boldsymbol{b}}\perp} := \prod_{(p,q)\in \mathrm{U}_{\boldsymbol{b}}} l_{\hat{\Omega};\boldsymbol{\theta}}\left(\mathbf{Y}_{\mathbf{b}(p,q)}\right) \tag{3}$$

whereby $\mathrm{U}_{\boldsymbol{b}} \; : = \{(p,q)|p,q \in \mathbb{N}, p < q \leq n_{\mathbf{b}}\}$. However, if we consider the subset N_b of U_b that refers to items with neighbored ranks (with respect to the response pattern under consideration), i.e.

$$\mathrm{N}_b := \left\{(p,q)|(p,q) \in \mathrm{U}_{\boldsymbol{b}} \wedge \left|\mathbf{y}_{\boldsymbol{b}(p)} - \mathbf{y}_{\mathbf{b}(q)}\right| = 1\right\} \tag{4}$$

it becomes obvious, that $l_{\mathbf{U}_b\perp}$ is expected to be smaller than $l_{\Omega;\boldsymbol{\theta}}(\mathbf{Y}_b)$, because the conditional likelihood of the binary comparisons of items with non-neighbored rank, given the values of all binary comparisons of items with neighbored ranks is always 1, i.e.

$$l_{\hat{\Omega};\boldsymbol{\theta}}\left(\mathbf{Y}_{\mathbf{b}(\mathrm{U}_b - \mathrm{N}_b)} \middle| \mathbf{Y}_{\mathbf{b}(\mathrm{N}_b)}\right) = 1 \tag{5}$$

because the binary comparisons of items with neighbored ranks imply the values of the remaining binary response indicators (for the response pattern under consideration).

This implies that

$$l_{\Omega;\boldsymbol{\theta}}(\mathbf{Y}_{\boldsymbol{b}}) = l_{\hat{\Omega};\boldsymbol{\theta}}\left(\mathbf{Y}_{\mathbf{b}(\mathrm{N}_b)}\right) l_{\hat{\Omega};\boldsymbol{\theta}}\left(\mathbf{Y}_{\mathbf{b}(\mathrm{U}_b - \mathrm{N}_b)} \middle| \mathbf{Y}_{\mathbf{b}(\mathrm{N}_b)}\right) = l_{\hat{\Omega};\boldsymbol{\theta}}\left(\mathbf{Y}_{\mathbf{b}(\mathrm{N}_b)}\right) \tag{6}$$

Consequently, the likelihood of the response to the forced choice block equals the joint likelihood of all binary comparisons that refer to neighbored ranks (for the response pattern under consideration).

Relaxing the assumption on local independence within a forced choice block by neglecting only local dependencies of items with neighbored ranks results to

$$l_{\Omega;\boldsymbol{\theta}}(\mathbf{Y}_{\boldsymbol{b}}) \cong l_{\mathrm{N}_b\perp} := \prod_{(p,q)\in \mathrm{N}_b} l_{\hat{\Omega};\boldsymbol{\theta}}\left(\mathbf{Y}_{\mathbf{b}(p,q)}\right) \tag{7}$$

and consequently

$$l_{\mathrm{N}_b\perp} \cdot \left(\prod_{(p,q)\in \mathrm{U}_{\boldsymbol{b}} - \mathrm{N}_{\boldsymbol{b}}} l_{\hat{\Omega};\boldsymbol{\theta}}\left(\mathbf{Y}_{\mathbf{b}(p,q)}\right) \right) = l_{\mathbf{U}_{\boldsymbol{b}}\perp} \tag{8}$$

The term in the brackets might be a good approximation of the size of underestimating $l_{\Omega;\boldsymbol{\theta}}(\mathbf{Y}_{\boldsymbol{b}})$ by using $l_{\mathbf{U}_{\boldsymbol{b}}\perp}$. Considering the respective log-likelihoods

$$logL_{\mathrm{N}_b\perp} + \sum_{(p,q)\in \mathrm{U}_{\boldsymbol{b}} - \mathrm{N}_b} logL_{\hat{\Omega};\boldsymbol{\theta}}\left(\mathbf{Y}_{\mathbf{b}(p,q)}\right) = logL_{\mathrm{U}_{\boldsymbol{b}}\perp} \tag{9}$$

it becomes obvious that the curvature of the log-likelihood and consequently the observed Fisher information is expected to be overestimated by including the terms that refer to binary comparisons of items with non-neighbored ranks.

If all binary comparison within a forced-choice would contribute independently to the Fisher information then the amount of information would increase dramatically with block size as there are $n_{\mathbf{b}}(n_{\mathbf{b}}-1)$ binary comparisons. Figure 1 shows how these expectation must be revised if local independence is only assumed for items with neighbored ranks which leads to $n_{\mathbf{b}}-1$ binary comparisons. Figure 1 includes direct (Likert-type response) item responses (to all the $n_{\mathbf{b}}$ items of the respective block) as benchmark (whereby it was assumed that a direct response is as informative as a binary comparison within a forced-choice block). It is obvious that the assumption of local independence of all binary item comparisons in a block leads to expectations which are by far too optimistic. In particular, the forced choice method is expected to outperform direct item responses under this scenario for block sizes greater than 3. The relative efficiency of the forced choice method (compared to Likert-type responses) would be expected to increase linearly with block size. In contrast, if the assumption of local independence is restricted to binary comparisons without algebraic dependencies the relative efficiency of the forced choice method (compared to Likert-type responses) is also expected increase monotonically. However, the positive effects of increasing block size are substantial for very small blocks only. Forced-choice tests would not be expected to outperform Likert-type tests for any block size. These considerations are not tied to the Thurstonian IRT model but apply to any arbitrary IRT measurement model of forced-choice data. In the remainder it is outlined how local dependencies of comparisons of items with neighbored ranks can be considered adequately in the framework of the Thurstonian IRT model .

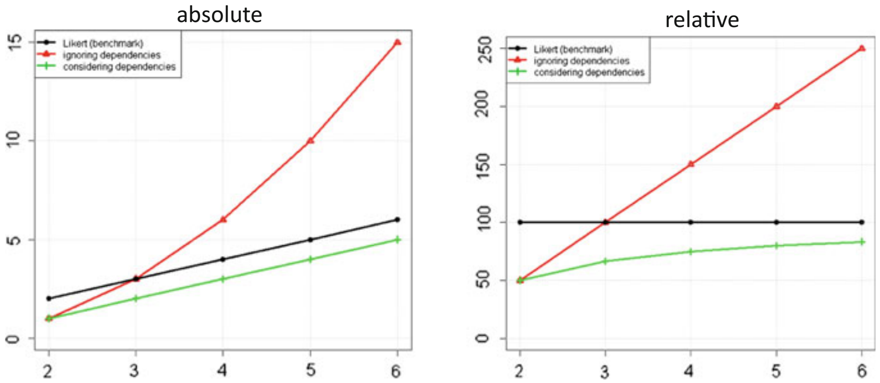


Fig. 1 Relative efficiency of the forced-choice method as function of block size and assumptions with regard to local dependencies

4 Thurstonian MIRT Model of Forced Choice Data

Thurstone's law of comparative judgment states, that the observed binary comparisons of the items $\mathbf{Y_b}$ are determined by a vector latent utilities $\mathbf{t_b}$ (of length $n_\mathbf{b}$) in the following way:

$$\mathbf{Y}_{\mathbf{b}(p,\,q)} = \begin{cases} 1 & \text{if} \quad \mathbf{t}_{\mathbf{b}(p)} - \mathbf{t}_{\mathbf{b}(q)} \geq 0 \\ 0 & \text{if} \quad \mathbf{t}_{\mathbf{b}(p)} - \mathbf{t}_{\mathbf{b}(q)} < 0 \end{cases} \tag{10}$$

The entries in $\mathbf{t_b}$ are assumed to be multivariate normally distributed:

$$\mathbf{t_b} \sim N(\boldsymbol{\mu}_\mathbf{b} + \boldsymbol{\Lambda}_\mathbf{b}\boldsymbol{\theta}, \boldsymbol{\Psi}_\mathbf{b}) \tag{11}$$

$\boldsymbol{\mu}_\mathbf{b}$ refers to the intercepts and $\boldsymbol{\Lambda}_\mathbf{b}$ refers to the rows of matrix $\boldsymbol{\Lambda}$ (of factor loadings), that correspond to the items of block $\mathbf{b}$, respectively. $f_{\mathbf{t_b}}$ is the probability density function of $\mathbf{t_b}$.

5 The Likehood Function of a Response Under the Thurstonian MIRT Model

The likelihood of $\mathbf{y_b}$ (the response to block $\mathbf{b}$) is given by:

$$l_{\hat{\Omega};\,\boldsymbol{\theta}}(\mathbf{y_b}) = l_{\hat{\Omega};\,\boldsymbol{\theta}}(\mathbf{Y_b}) = \int\limits_{S_\mathbf{b}} f_{\mathbf{t_b}}(\mathbf{x})d\mathbf{x} \tag{12}$$

whereby $S_{\mathbf{b}}$ refers to the region of $\mathbb{R}^{n_{\mathbf{b}}}$ where the following system of $n_{\mathbf{b}} - 1$ inequalities holds true:

$$\mathbf{C_b t}_{\mathbf{b}\langle \mathbf{y_b} \rangle} \geq 0^{n_{\mathbf{b}} - 1} \tag{13}$$

whereby $0^{n_{\mathbf{b}} - 1} \in \mathbb{R}^{n_{\mathbf{b}} - 1}$ is a vector of $n_{\mathbf{b}} - 1$ entries of 0 and $\mathbf{C_b}$ is a matrix with $n_{\mathbf{b}} - 1$ rows and $n_{\mathbf{b}}$ columns with

$$\mathbf{C}_{\mathbf{b}(p,q)} = \begin{cases} 1 & \text{if} \quad q = p \\ -1 & \text{if} \quad q = p + 1 \\ 0 & \text{if} \quad \text{otherwise} \end{cases} \tag{14}$$

Geometrically, $S_{\mathbf{b}}$ is a $n_{\mathbf{b}}$-dimensional parallelotope with two infinite ends on one of its dimensions and one infinite end on the remaining dimensions. Integration of $f_{\mathbf{t_b}}$ over $S_{\mathbf{b}}$ can done by the methods developed by Genz (2004) which allows computing the likelihood of any response to a forced-choice block. Computing the log-likelihood across all blocks is straightforward and maximizing the respective likelihood function to get Maximum-likelihood estimates of the latent trait can be done by standard optimizing procedures. The observed Fisher information can be computed numerically as well. The properties of the estimation procedure will be dealt with by Yousfi (in prep.).

6 Conclusions

Recommendations for assembling forced-choice questionnaires of Brown and Maydeu-Olivares (2011) that aim at measurement precision and trait recovery rely to some degree on properties of the Fisher information matrix. However, ignoring local dependencies might result in misleading conclusions. Brown and Maydeu-Olivares (2011) warned that forced-choice blocks should not contain more than four items in order to avoid cognitive overload, but established the expectation that increasing block size should be an effective way to enhance measurement precision. The considerations in this paper suggest that it usually won't pay off to increase block size until the limit of cognitive capacity is reached. The outlined approach of computing the likelihood and Fisher information might contribute to a solid psychometrical basis for recommendations with regard to the assembly of forced choice questionnaires.

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.

Genz, A. (2004). Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing, 14,* 251–260.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45,* 935–974.

Yousfi, S. (in preparation). Considering local item dependencies in the estimation of person parameters in the Thurstonian IRT model of forced-choice questionnaires.

Elimination Scoring Versus Correction for Guessing: A Simulation Study

Qian Wu, Tinne De Laet and Rianne Janssen

Abstract Administering multiple-choice questions with correction for guessing fails to take into account partial knowledge and may introduce a bias as examinees may differ in risk-taking to guess the correct answer when not having full knowledge. In the latter case, elimination scoring gives examinees the opportunity to express their partial knowledge as this alternative scoring procedure requires examinees to eliminate all the response alternatives they consider to be incorrect. The current simulation study investigates how these two scoring procedures affect response behaviors of examinees who differ not only in ability but also in their attitude toward risk. Combining a psychometric model accounting for ability and item difficulty with the decision theory accounting for individual differences in risk aversion, a two-step response-generating model is proposed to predict the expected answering patterns on given multiple-choice questions. The results of the simulations show that overall there are no substantial differences in the answering patterns for examinees at both ends of the ability continuum under two scoring procedures, suggesting that ability has a predominant effect on the response patterns. Compared to correction for guessing, elimination scoring leads to fewer full score response and more demonstration of partial knowledge, especially for examinees with intermediate success probabilities on the items. Only for those examinees, risk aversion has a decisive impact on the expected answering patterns.

Keywords Multiple-choice questions • Elimination scoring • Correction for guessing • IRT • Prospect theory

Q. Wu (✉) · R. Janssen
Faculty of Psychology and Educational Sciences, KU Leuven, Dekenstraat 2,
Box 3773, 3000 Leuven, Belgium
e-mail: qian.wu@kuleuven.be

R. Janssen
e-mail: rianne.janssen@kuleuven.be

T. De Laet
Faculty of Engineering Science, KU Leuven, Dekenstraat 2,
Box 3773, 3000 Leuven, Belgium
e-mail: tinne.delaet@kuleuven.be

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_16

1 Introduction

Multiple-choice (MC) questions are often scored dichotomously with a penalty for wrong answers. This scoring method, known as "correction for guessing", is based on the assumption that examinees either possess the full knowledge to know the answer to the question, or they do not know and guess randomly among all response alternatives. However, this assumption fails to take partial knowledge into account. Examinees can always use their partial knowledge to make an informed or profitable guess rather than a random guess (Frary 1988). According to Lindquist and Hoover (2015), what can only be implied from responses to MC questions is that some examinees may answer the question on the basis of more certain or complete knowledge, having a more accurate memory, or a sounder reasoning than others. Therefore, responses to MC questions should be considered to be on a continuum rather than a discrete know–don't know dichotomy.

Moreover, in situations of partial knowledge, examinees may differ in their willingness to guess—some are more daring to take the risk of receiving a penalty in case they guess incorrectly, while others may take the more conservative answering strategy of omission to avoid getting a penalty. Consequently, correction for guessing may introduce a bias with regard to risk aversion (Bereby-Meyer et al. 2002; Lesage et al. 2013). This alleged bias has led some universities and testing institutions to abandon correction for guessing on MC tests (De Laet et al. 2015; SAT).

As an alternative to correction for guessing, elimination scoring (Coombs et al. 1956) gives examinees the opportunity to express their knowledge level on the MC question by instructing examinees to eliminate all the response alternatives that they consider to be incorrect. For a MC question with four alternatives, this leads to 15 possible answer patterns that can be classified into five different knowledge levels (see Table 1). Note that the response of eliminating all alternatives (XXXX) is considered to be irrational as one of them is known to be correct. Under the scoring rule proposed by Arnold and Arnold (1970), partial credit is given to each correct elimination of a distractor and a penalty to the elimination of the correct answer so that the expected score of random elimination is zero (see the last column in Table 1). It has been shown that by rewarding partial knowledge, elimination scoring increases student performance and test satisfaction, while reduces test anxiety (Bond et al., 2013; De Laet et al. 2016).

The current study aims to compare how these two scoring procedures affect response behaviors of examinees who differ not only in ability but also in their attitude toward risk. In a simulation study, a psychometric model of item response theory (IRT) that takes into account ability of a person and item characteristics is combined with a behavioral decision model of prospect theory that accounts for individual differences in attitude toward risk and losses. The rationale of our study is in line with the approach by Budescu and Bo (2015) who simulated the effect of ability and risk aversion on response omissions on MC items under correction for guessing using a model combining IRT and decision theory. In the present study,

Table 1 Response patterns and scores under correction for guessing and elimination scoring for a multiple-choice item with four alternatives of which the first one is correct

		Score	
Knowledge level	Response pattern	Correction for guessing	Elimination scoring
Full knowledge	OXXX	1	1
Partial knowledge 2	OXXO, OXOX, OOXX	–	1/3
Partial knowledge 1	OXOO, OOXO, OOOX	–	1/9
No knowledge	OOOO	0	0
Misconception	XXXO, XXOX, XOXX	−1/3	−1/3
	XXOO, XOXO, XOOX	–	−1/3
	XOOO	–	−1/3

Note X = elimination; O = non-elimination; –: not applicable; −1/3: the penalty for wrong responses and eliminating the correct answer in correction for guessing and elimination scoring, respectively

we expand their approach by simulating which particular response pattern examinees of various levels of ability and risk aversion will give to a MC question under the two scoring procedures. Although examinees in vivo may behave differently from what a fully rational model predicts, the results of this in vitro study may reveal the differential impact of ability and risk aversion under the two scoring procedures and may give an indication of the size of the alleged willingness to guess bias in MC questions.

2 Method

2.1 The Response-Generating Model

The model consists of two steps. In the first step, the Rasch model is used to model the subjective probabilities of knowing the correct response to each of the alternatives of a MC item, given an examinee's ability and the alternatives' difficulties. In the second step, prospect theory is used to predict how examinees make a decision on how to answer the MC item given the obtainable scores for all possible response patterns and their probabilities to receive those scores by taking into account individual differences in risk aversion. A MC question with four response alternatives A, B, C and D, of which A is the correct answer, is used as an example.

Step 1: Modeling the probability of knowing a correct response to each alternative. Consider a MC question as a testlet with the four alternatives as its binary sub-items. First, assume that those four sub-items can be viewed as fully independent from each other. Then the probability of giving a correct response to each alternative is modeled using the Rasch model with the examinee's ability and the sub-item's difficulty as parameters, i.e., $P(A_{correct})$, $P(B_{correct})$, $P(C_{correct})$, and

$P(D_{correct})$, where the latter three probabilities refer to the probability of considering that alternative as a wrong answer. Thus, a high ability not only increases the probability of recognizing the correct answer, but also increases the probability of identifying distractors as incorrect. The rationale behind using the Rasch model for the individual sub-items is to consider the modeled probability on each alternative as the probability of the examinee giving that answer indicated on the alternative as would be the case that the MC question were posed in an open response format. Hence, there would be a universe of all possible responses. Since the response alternatives of the MC question are assumed to be only a sample of the universe of all possible responses, the modeled probabilities on the given alternatives can be used without the constraint of summing up to 1 at each level of ability.

Step 2: Modeling the decision-making process. The goal of a rational test-taker is to maximize the expected scores on the test. As shown in Table 1, there are five possible response patterns under correction for guessing (selecting one of the four alternatives or omission), and 15 under elimination scoring (each alternative can be eliminated or not). Each response pattern is associated with certain points set out by the scoring rules. Given all the possible answering patterns and scores, prospect theory (Kahneman and Tversky 1979) is used to predict which response will be given by examinees.

Prospect theory is a behavioral economic theory that describes how people make decisions between probabilistic alternatives under uncertainty and risk. The theory states that people make decisions based on the potential values of gains and losses rather than the objective outcomes. When faced with a number of actions, each of which gives rise to more than one possible outcomes x_k $(k = 1, 2, \ldots, n)$ with different (objective) probabilities p_k, a person will make a decision that optimizes the expected utility U, depending on the (a) subjective probability $\pi(p_k)$ of each possible outcome, (b) the personal value function $v(x_k)$ of potential losses and gains, and (c) the loss aversion parameter λ, as:

$$U = \sum_{k=1}^{n} \pi(p_k) * v(x_k) \tag{1}$$

where

$$v(x_k) = \begin{cases} x^a, & when\, x \geq 0 \\ -\lambda * (-x)^{\beta}, & when\, x < 0 \end{cases} \tag{2}$$

and a, β are diminishing sensitivity parameters,[1] which were fixed to 0.75 in the present study.

[1] a and β can take different values, but in many studies they are often set to be equal (see Budescu and Bo 2015).

According to the theory, each response pattern is a prospect. Since the correct answer is not known to examinees, for each response pattern there are four possible outcomes, namely, either A, B, C, or D being correct. The subjective probability of such an occurrence is derived as the multiplication of the four possibilities on the alternatives calculated in Step 1. That is, a subjective probability of considering A being the correct answer, when it is indeed the correct response, implies giving correct responses to all four sub-items, i.e., $\pi_A = P(A_{correct})*P(B_{correct})*P(C_{correct})*P(D_{correct})$. A subjective probability of B being the correct answer implies that two incorrect responses are made: A is considered not as the correct answer and B not a distractor, and hence, $\pi_B = [1-P(A_{correct})]*[1-P(B_{correct})]*P(C_{correct})*P(D_{correct})$. The subjective certainty of the other two alternatives follows the same logic, $\pi_C = [1-P(A_{correct})]*P(B_{correct})*[1-P(C_{correct})]*P(D_{correct})$ and $\pi_D = [1-P(A_{correct})]*P(B_{correct})*P(C_{correct})*[1-P(D_{correct})]$.

Using these subjective probabilities, the expected utility of a response pattern is then calculated as the sum of the personal values of obtainable scores weighted by the probabilities of each outcome, taking into account the loss and risk aversion. Table 2 gives the calculations of expected utilities for all response patterns. The response pattern of omission has an expected utility of zero, as no points are to be gained or lost. The answering pattern with the maximum utility is expected to be chosen as the final response given by examinees to the MC question.

Table 2 Calculation of expected utilities of all possible answering patterns

Answering pattern	Possible outcome: if the correct answer is …			
	A	B	C	D
OXXX	π_A*1^a+	$\pi_B*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_C*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_D*-\lambda*\left(\frac{1}{3}\right)^a$
XOXX	$\pi_A*-\lambda*\left(\frac{1}{3}\right)^a+$	π_B*1^a+	$\pi_C*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_D*-\lambda*\left(\frac{1}{3}\right)^a$
XXOX	$\pi_A*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_B*-\lambda*\left(\frac{1}{3}\right)^a+$	π_C*1^a+	$\pi_D*-\lambda*\left(\frac{1}{3}\right)^a$
XXXO	$\pi_A*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_B*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_C*-\lambda*\left(\frac{1}{3}\right)^a+$	π_D*1^a
OOXX	$\pi_A*\left(\frac{1}{3}\right)^a+$	$\pi_B*\left(\frac{1}{3}\right)^a+$	$\pi_C*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_C*-\lambda*\left(\frac{1}{3}\right)^a$
OXOX	$\pi_A*\left(\frac{1}{3}\right)^a+$	$\pi_B*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_C*\left(\frac{1}{3}\right)^a+$	$\pi_D*-\lambda*\left(\frac{1}{3}\right)^a$
OXXO	$\pi_A*\left(\frac{1}{3}\right)^a+$	$\pi_B*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_C*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_D*\left(\frac{1}{3}\right)^a$
XOOX	$\pi_A*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_B*\left(\frac{1}{3}\right)^a+$	$\pi_C*\left(\frac{1}{3}\right)^a+$	$\pi_D*-\lambda*\left(\frac{1}{3}\right)^a$
XOXO	$\pi_A*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_B*\left(\frac{1}{3}\right)^a+$	$\pi_C*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_D*\left(\frac{1}{3}\right)^a$
XXOO	$\pi_A*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_B*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_C*\left(\frac{1}{3}\right)^a+$	$\pi_D*\left(\frac{1}{3}\right)^a$
XOOO	$\pi_A*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_B*\left(\frac{1}{9}\right)^a+$	$\pi_C*\left(\frac{1}{9}\right)^a+$	$\pi_D*\left(\frac{1}{9}\right)^a$
OXOO	$\pi_A*\left(\frac{1}{9}\right)^a+$	$\pi_B*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_C*\left(\frac{1}{9}\right)^a+$	$\pi_D*\left(\frac{1}{9}\right)^a$
OOXO	$\pi_A*\left(\frac{1}{9}\right)^a+$	$\pi_B*\left(\frac{1}{9}\right)^a+$	$\pi_C*-\lambda*\left(\frac{1}{3}\right)^a+$	$\pi_D*\left(\frac{1}{9}\right)^a$
OOOX	$\pi_A*\left(\frac{1}{9}\right)^a+$	$\pi_B*\left(\frac{1}{9}\right)^a+$	$\pi_C*\left(\frac{1}{9}\right)^a+$	$\pi_D*-\lambda*\left(\frac{1}{3}\right)^a$
OOOO	0			

Table 3 Difficulty parameters of the response alternatives of multiple-choice items used in the simulations

	Response alternative			
Item	A*	B	C	D
1	0	−1	−2	−3
2	1	0	−1	−2
3	2	1	0	−1
4	0	1	−1	−2
5	1	0.5	0	−0.5

Note *Alternative A is the correct answer

2.2 *The Choice of the Person and Item Parameters*

A sample of 2000 examinees was simulated with the ability parameters θ_i from a normal distribution of $N(0,1)$. The level of risk aversion is modeled using the loss-aversion parameter λ_i. The prospect theory states that typically losses hurt more than gains feel good (Kahneman and Tversky 1979), and hence the value of λ_i is usually equal to or larger than one. Given that there has been no empirical evidence in the literature about the association between ability θ_i and risk aversion λ_i, these two parameters are assumed to be uncorrelated in the current study, and the risk aversion parameters λ_i is generated from a uniform distribution of $U[1,6]$. Table 3 gives the different sets of difficulty parameters of the response alternatives of the MC items that were used in the simulation. Items 1–3 are conventional items where the correct alternative A has the highest difficulty. They represent items of low, medium, and high difficulties, respectively. Item 4 is a so-called unconventional or tricky item where one of the distractors is the most difficult, and Item 5 is a conventional item of intermediate difficulty but with smaller difficulty differences between response alternatives.

3 Results

Figure 1 presents the plots of the expected answering patterns on Items 1–5 as a function of ability and risk aversion using the response generating model under the two scoring procedures. Note that because the difficulties of the four alternatives are arranged in an increasing order, only seven possible response patterns are observed. Looking at the plots in general, it can be seen that there is a main effect of both ability and risk aversion on the expected response patterns. However, as the separations between the adjacent response patterns in each of the plots are tilted, there is also an interaction between both variables. The effect of risk-aversion depends on the level of ability and vice versa. In case of two independent main effects, the separations between the adjacent response patterns should follow straight vertical or horizontal lines.

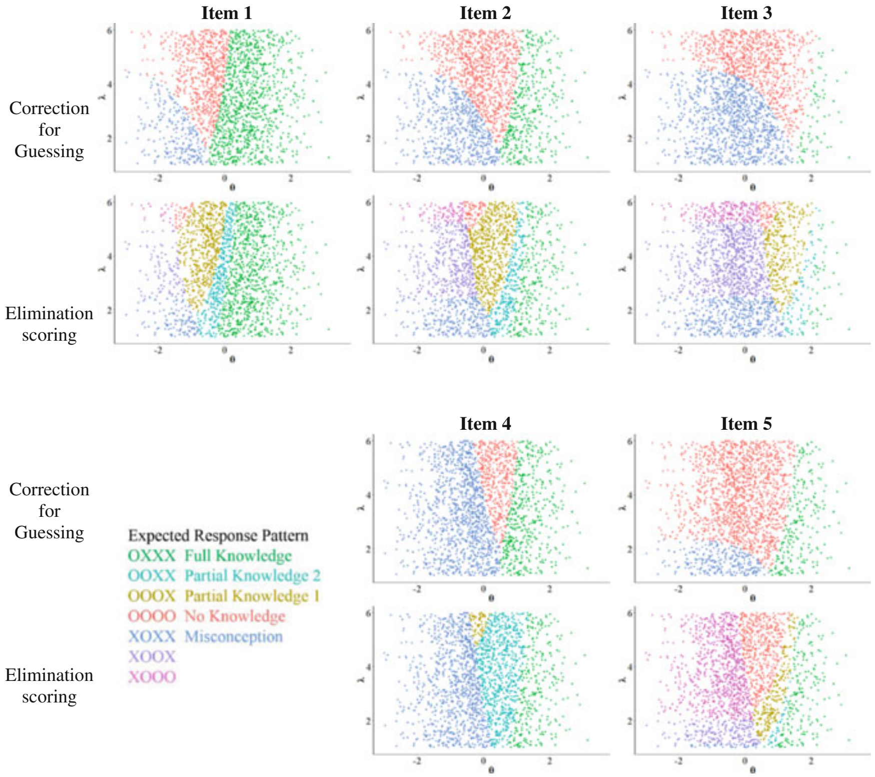


Fig. 1 Plots of the expected answering patterns on Items 1–5 under correction for guessing and elimination scoring

The upper three sets of plots are the results from the three conventional items with easy, medium, and high difficulties, respectively. Within each scoring procedure, there is a clear shift of the expected response patterns to the higher ability end, expressing the change in success probabilities along the ability scale when item difficulty increases.

To compare the response patterns under two scoring procedures, take a closer look at Item 2 with the medium difficulty. For examinees with no risk aversion $(\lambda_i = 1)$, there is a clear switch point on the ability scale between incorrect and correct responses under correction for guessing (upper panel). Under elimination scoring (lower panel), on the other hand, there is a small proportion of partial knowledge observed around the switch point in correction for guessing, resulting in a smaller amount of full score response. This suggests that at least part of the full score response in correction for guessing is due to (informed) guessing. When given the opportunity under elimination scoring, those examinees with partial knowledge choose to express their doubt by leaving some alternatives open.

When risk aversion increases, the proportion of omission responses increases under correction for guessing, whereas under elimination scoring, omission is rather scarce. Instead, examinees choose to show their partial knowledge. Note that this effect is asymmetrical. It is more prominent for examinees with lower abilities. For both scoring procedures, examinees with higher abilities $(\theta_i > 1)$ show no effect of risk aversion—they all receive full scores; and examinees with lower abilities $(\theta_i < 1)$ barely show the effect of risk aversion—they all give an incorrect response receiving a penalty. The strongest effect of risk aversion is observed in the middle range of the ability scale. These examinees only have partial knowledge of the item, and hence, are in doubt. There is a fair probability of gaining the point, but also a fair probability of receiving a penalty. Therefore, the factor of risk aversion has a bigger impact for them. It is also interesting to notice that the separation between misconception and omission under correction for guessing is more tilted than that between misconception and the adjacent response categories under elimination scoring, suggesting that to some extent elimination scoring diminishes the effect of risk aversion, although not entirely.

The bottom left panel of Fig. 1 shows the expected answering patterns on an unconventional item (Item 4). Under correction for guessing, the unconventional item seems to confuse examinees with lower abilities and leads to more incorrect responses compared with Item 2. In contrary, under elimination scoring the distributions of the knowledge levels on the two items are rather similar. There is no increase in the amount of incorrect responses on the unconventional item. Lower ability examinees still show misconception and higher ability examinees obtain a full score. Examinees with partial knowledge obtain a higher score in elimination scoring than in correction for guessing.

The comparison of the expected answering patterns on Item 2 and Item 5 (bottom right) indicates that when the differences between alternatives become smaller, omission becomes the dominant strategy for most of the examinees who are in doubt, given both scoring procedures. Nevertheless, examinees with relatively higher abilities can still receive partial credit by demonstrating partial knowledge under elimination scoring, while examinees with lower abilities tend to lose points by showing misconception.

4 Discussion

4.1 Conclusions

The results of the simulation study show that overall ability has a predominant effect on the response patterns. There are no substantial differences in the answer patterns for examinees at both ends of the ability continuum under two scoring procedures. Risk aversion only has a stronger impact for examinees with intermediate success probabilities on the items, but elimination scoring diminishes this effect to some extent. When examinees are in doubt, they benefit from elimination scoring by leaving the alternatives they feel uncertain about open and obtaining partial credit on the correct elimination(s) they can make. Although the majority of the variation in responses is captured by ability, supporting the validity of administering MC tests with correction for guessing, elimination scoring may further improve MC tests by reducing the effect of risk aversion on examinees who are in doubt. Note that the impact of risk aversion depends on the relative difficulty of the item for the examinee. When a test is composed of items of a wide range of difficulties, risk aversion affects examinees in several ranges on the ability scale, but not necessarily on all the items. Moreover, in case risk aversion and ability are in reality correlated, the former variable may also indirectly affect the response probabilities on each of the alternatives and hence have a higher impact on the responses.

By rewarding partial knowledge and allowing expression of doubt, elimination scoring offers examinees the opportunity to express their uncertainty when they do not have full knowledge, and consequently reduces the need and the amount of guessing the correct answer to the question. It is also useful in providing both examinees and examiners with more differentiated feedback on what kind of misconception or problems examinees have, and facilitate remedial instructions for future learning. Ben-Simon et al. (1997) concluded in their comparative study of several scoring methods in MC tests that no response method was uniformly best across criteria and content domains, but the current study shows that elimination scoring can be a more neutral (with respect to risk aversion), and hence, a viable alternative to correction for guessing in MC tests.

4.2 Limitations

The conclusions of this study were drawn given a theoretical model set up to predict examinees' purely rational decisions based on ability and risk aversion. The response-generating model used in the study may be a simplified representation of real response behaviors to MC items in the following ways.

First, the alternatives of the MC items are treated as independent sub-items in the first psychometrical step when estimating the response probabilities, and then are considered simultaneously in the second decision-theoretical step when choosing the most optimal answering pattern. Another possibility is to model the probability of being in a certain knowledge level for the MC item as a whole and then using prospect theory to make a response decision, as was done by De Laet et al. (2016). Despite its different approach, the latter study yielded similar conclusions as the present one.

Second, the subjective probabilities of each possible outcome used in the calculation of the expected utility in the second step were set equal to the objective (true) probabilities derived from the Rasch model in the first step. However, according to the prospect theory, these two probabilities do not necessarily match perfectly, because people tend to mis-calibrate extreme probabilities, e.g., overconfidence or underestimation (Kahneman and Tversky 1979). A potential link function between the subjective and objective probabilities as proposed by Budescu and Bo (2015) may be a useful addition to the model.

Finally, it is hard to see how the risk aversion parameter in prospect theory links to the reality of MC items, given that a high value of λ_i may correspond to a perceived value of the penalty that is much higher in absolute value than the value of one point an item has. In sum, although useful theoretical results were obtained on the comparison of the two scoring procedures, the present in vitro study should definitely be supplemented with empirical in vivo studies on examinees' response behaviors.

References

Arnold, J. C., & Arnold, P. L. (1970). On scoring multiple choice exams allowing for partial knowledge. *The Journal of Experimental Education, 39,* 8–13. https://doi.org/10.1080/00220973.1970.11011223.

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21,* 65–88. https://doi.org/10.1177/0146621697211006.

Bereby-Meyer, Y., Meyer, J., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making, 15,* 313–327. https://doi.org/10.1002/bdm.417.

Bond, A. E., Bodger, O., Skibinski, D. O. F., Jones, D. H., Restall, C. J., Dudley, E., et al. (2013). Negatively-marked MCQ assessments that reward partial knowledge do not introduce gender bias yet increase student performance and satisfaction and reduce anxiety. *PLoS ONE*, *8*. https://doi.org/10.1371/journal.pone.0055956.

Budescu, D. V., & Bo, Y. (2015). Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika, 80,* 1105–1122. https://doi.org/10.1007/s11336-014-9425-x.

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16,* 13–37. https://doi.org/10.1177/001316445601600102.

De Laet, T., Vanderoost, J., Callens, R., & Janssen, R. (September 2016). *Assessing engineering students with multiple choice exams: Theoretical and empirical analysis of scoring methods*. Paper presented at the 44th annual SEFI Conference. Tampere, Finland.

De Laet, T., Vanderoost, J., Callens, R., & Vandewalle, J. (June 2015). *How to remove the gender bias in multiple choice assessments in engineering education?* Paper presented at the 43rd annual SEFI conference. Orléans, France.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice, 7,* 33–38. https://doi.org/10.1111/j.1745-3992.1988.tb00434.x.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–292. https://doi.org/10.2307/1914185.

Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education–Is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation, 39,* 188–193. https://doi.org/10.1016/j.stueduc.2013.07.001.

Lindquist, E. F., & Hoover, H. D. (2015). Some notes on corrections for guessing and related problems, *34*, 15–19.

SAT Suit of Assessments. (n.d.). How SAT is scored. Retrieved from https://collegereadiness.collegeboard.org/sat/scores/how-sat-is-scored.

Three-Way Generalized Structured Component Analysis

Ji Yeh Choi, Seungmi Yang, Arthur Tenenhaus and Heungsun Hwang

Abstract Generalized structured component analysis (GSCA) is a component-based approach to structural equation modeling, where components of observed variables are used as proxies for latent variables. GSCA has thus far focused on analyzing two-way (e.g., subjects by variables) data. In this paper, GSCA is extended to deal with three-way data that contain three different types of entities (e.g., subjects, variables, and occasions) simultaneously. The proposed method, called three-way GSCA, permits each latent variable to be loaded on two types of entities, such as variables and occasions, in the measurement model. This enables to investigate how these entities are associated with the latent variable. The method aims to minimize a single least squares criterion to estimate parameters. An alternating least squares algorithm is developed to minimize this criterion. We conduct a simulation study to evaluate the performance of three-way GSCA. We also apply three-way GSCA to real data to demonstrate its empirical usefulness.

Keywords Generalized structured component analysis • Three-way data
Structural equation modeling • Alternating least squares

J. Y. Choi
National University of Singapore, 9 Arts Link, Singapore, Singapore
e-mail: psycjy@nus.edu.sg

S. Yang
McGill University, 1020 Pine Ave West, Montreal, QC, Canada
e-mail: seungmi.yang@mcgill.ca

A. Tenenhaus
Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec-CNRS-Université Paris-Sud, Université Paris-Saclay, 3, rue Joliot Curie, 91192 Gif-sur-Yvette, France
e-mail: arthur.tenenhaus@centralesupelec.fr

H. Hwang (✉)
McGill University, 2001 McGill College, Montreal, QC, Canada
e-mail: heungsun.hwang@mcgill.ca

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_17

1 Introduction

Generalized structured component analysis (GSCA; Hwang and Takane 2004, 2014) is a component-based approach to structural equation modeling, in which weighted composites or components of observed variables are used as proxies for conceptual or latent variables. GSCA involves three sub-models to specify a general structural equation model: weighted relation, measurement, and structural models. The weighted relation model is used to define a latent variable as a weighted composite or component of observed variables; the measurement model is to specify the relationships between latent variables and their observed variables; and the structural model is to express the relationships between latent variables. In GSCA, these three sub-models are combined into a single model formulation, which in turn facilitates the derivation of a global optimization criterion for parameter estimation. An alternating least squares algorithm (de Leeuw et al. 1976) was developed to minimize this criterion.

Various extensions of GSCA have been developed to enhance its data-analytic scope and flexibility. For instance, Hwang et al. (2007) proposed fuzzy clusterwise GSCA, which integrated GSCA and fuzzy clustering in a unified framework to uncover subgroups of observations, each of which may involve different path-analytic relationships between observed and latent variables. Another extension focused on accommodating various interaction terms of latent variables (Hwang et al. 2010). Moreover, Hwang et al. (2007) developed multilevel GSCA to take into account hierarchically structured data, where cases at a lower-level unit are grouped within those at a higher-level unit, e.g., students nested within classrooms. Refer to Hwang and Takane (2014) for a comprehensive discussion of a wide range of the extensions.

To date, GSCA and all its extensions have been geared for the analysis of two-way data, which contain two different types of entities (e.g., subjects and variables). Nonetheless, in practice, the same subject can often be measured on a set of variables over another type of entities, for example, times or situations. This gives rise to so-called three-way data consisting of three different types of entities concurrently, each of which is called a mode. A few examples of three-way data include neuroimaging data as an array of subjects by brain locations by time points/scans (e.g., Cox 1996; Germond et al. 2000; Thirion and Faugeras 2003), fluorescence spectroscopy data as an array of samples by emission spectra by excitation wavelengths (e.g., Andersen and Bro 2003; Bro 1997; Christensen et al. 2005), and multivariate longitudinal data as an array of subjects by variables by occasions (e.g., Kroonenberg 1987; Kuze et al. 1985; Oort 2001). Although there exist various forms of three-way data, as mentioned above, we shall assume an array of subjects by variables by occasions as the standard data structure hereafter unless otherwise specified.

Having collected data in a three-way array structure, researchers in many areas of psychology, such as neuropsychology, developmental and cognitive psychology, are interested in addressing at least two different types of questions (e.g., De Roover et al. 2012; Ferrer and McArdle 2010; Kroonenberg 2008). The first main objective is aimed at revealing relationships among latent and observed variables, and the second is at describing how the modelled psychological process unfolds over different occasions. While the first objective itself can be addressed by aggregating or pooling three-way data over the third mode and then applying two-way data analysis, to answer both types of questions simultaneously, researchers should be able to exploit the three-way data as they are collected.

In this paper, we propose an extension of GSCA to the analysis of three-way data, called three-way GSCA. Three-way GSCA is free from any aggregating or pooling procedure and applies the analysis directly to the three-way data. As in GSCA, three-way GSCA consists of the same sub-models. However, it extends the measurement model to relate each latent variable to entities in both second and third modes (i.e., variables and occasions, respectively), so that it provides the estimates of loadings for both second and third modes simultaneously. As a result, three-way GSCA enables researchers to investigate how variables and occasions are associated with a latent variable.

Three-way GSCA differs from other approaches to three-way data, multilinear partial least squares (M-PLS; Bro 1996) and multiway regularized generalized canonical correlation analysis (MGCCA; Tenenhaus et al. 2015). M-PLS extracts components of a set of three-way data via parallel factor analysis (PARAFAC; Harshman 1970) and then investigates the effects of the components on endogenous observed variables. M-PLS involves two sequential steps of estimating model parameters (i.e., one step for estimating components and another for estimating regression coefficients). In contrast, three-way GSCA involves a single estimation procedure to estimate all parameters simultaneously. Also importantly, three-way GSCA is more general than M-PLS, because the former can contemplate multiple sets of exogenous and endogenous three-way data, whereas the latter concerns multiple sets of exogenous three-way data only. Recently, RGCCA was proposed to analyze multiple sets of three-way data. MGCCA extracts components from each set of three-way data in such a way that they are mutually orthogonal to each other within the same set, but maximally correlated across different sets. However, MGCCA focuses on investigating non-directional associations (i.e., correlations) among multiple sets of three-way data. Conversely, three-way GSCA aims to examine directional (path-analytic) relationships among latent variables as well as observed entities in three-way data.

The paper is organized as follows. In Sect. 2, we provide technical accounts of three-way GSCA. It describes model specification and parameter estimation. In Sect. 3, we conduct a Monte Carlo simulation study to evaluate the performance of three-way GSCA in parameter recovery. In Sect. 4, we illustrate the empirical feasibility through the analysis of a real data set. In the final section, we summarize the implications of three-way GSCA and discuss directions for future research.

2 Method

Let $\underline{\mathbf{X}}_p$ denote the pth three-way data set ($p = 1, \ldots, P$), arranged in a block of I by J_p by K_p, where I is the number of entities (e.g., subjects) in the first mode, which is assumed to remain the same across all P data sets, J_p is the number of entities (e.g., variables) in the second mode, and K_p is the number of entities (e.g., occasions) in the third mode. Let $\mathbf{X}_p$ denote an I by J_pK_p matrix constructed by aligning each I by J_p frontal matrix of $\underline{\mathbf{X}}_p$ K_p times next to one another. Let $\underline{\mathbf{X}}_p$ have been centered across the first mode and normalized within the second mode. This centering is done by subtracting the column mean from every value in each column of $\mathbf{X}_p$ (Bro 1997). Also, we assume that $\underline{\mathbf{X}}_p$ are normalized within the variables' mode to adjust their values measured in different units to a common scale. It is performed by dividing each element in data by a square root of sum of squares of all element associated with the j_pth variable (Bro 2003). Let $\boldsymbol{\gamma}_p$ denote an I by 1 column vector of the pth latent variable scores. Let $\mathbf{w}_p = \left[w_{p11}, \ldots, w_{pJ_P1}, w_{p12}, \ldots, w_{pJ_P2}, \ldots, w_{p1K_P}, \ldots, w_{1J_PK_P}\right]'$ denote a J_pK_p by 1 column vector of component weights for the pth latent variable. Let $\mathbf{c}_p^{\mathrm{J}} = \left[c_{p1}^{\mathrm{J}}, \ldots, c_{pJ_P}^{\mathrm{J}}\right]'$ and $\mathbf{c}_p^{\mathrm{K}} = \left[c_{p1}^{\mathrm{K}}, \ldots, c_{pK_P}^{\mathrm{K}}\right]'$ denote a J_p by 1 and K_p by 1 column vectors of loadings relating γ_p to the second and third modes of the pth data set, respectively.

As stated earlier, three-way GSCA involves three sub-models. The weighted relation model defines a latent variable as a weighted composite or component of the first mode, as follows:

$$\boldsymbol{\gamma}_p = \mathbf{X}_p\mathbf{w}_p. \tag{1}$$

This is similar to (two-way) GSCA. On the other hand, the measurement model specifies the relationship between entities in the second and third modes and its latent variable, as follows:

$$\mathbf{X}_p = \boldsymbol{\gamma}_p\left(\mathbf{c}_p^{\mathrm{K}} \otimes \mathbf{c}_p^{\mathrm{J}}\right)' + \mathbf{E}_{1p}, \tag{2}$$

where $\mathbf{E}_{1p}$ is the residual for $\mathbf{X}_p$, and $\otimes$ indicates the Kronecker product. More generally, let $\mathbf{X}^* = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_P]$, $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_P]$ and $\mathbf{E}_1 = [\mathbf{E}_{11}, \mathbf{E}_{12}, \ldots, \mathbf{E}_{1P}]$. The measurement model can be then re-expressed in matrix notation as follows:

$$\mathbf{X}^* = \boldsymbol{\Gamma}\mathbf{C} + \mathbf{E}_1, \tag{3}$$

where $\mathbf{C} = \begin{bmatrix} \left(\mathbf{c}_1^{\mathrm{K}} \otimes \mathbf{c}_1^{\mathrm{J}}\right)' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \left(\mathbf{c}_P^{\mathrm{K}} \otimes \mathbf{c}_P^{\mathrm{J}}\right)' \end{bmatrix}$.

The structural model expresses hypothesized path-analytic relationships among latent variables as follows:

$$\boldsymbol{\gamma}_p = \sum_{q=1}^{P} \boldsymbol{\gamma}_q b_{qp} + \boldsymbol{\varepsilon}_{2p}, \tag{4}$$

where b_{qp} indicates a path coefficient of $\boldsymbol{\gamma}_q$ on $\boldsymbol{\gamma}_p$, and $\boldsymbol{\varepsilon}_{2p}$ an I by 1 vector of residuals. In matrix notation, the structural model is re-expressed as

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\mathbf{B} + \mathbf{E}_2, \tag{5}$$

where $\mathbf{B}$ is the matrix whose (q, p)th element is b_{qp} and $\mathbf{E}_2 = \left[\boldsymbol{\varepsilon}_{21}, \ \boldsymbol{\varepsilon}_{22}, \ \ldots, \ \boldsymbol{\varepsilon}_{2p}\right]$.

Figure 1 shows a hypothetical example of a three-way GSCA model. This model involves two three-way data sets ($P = 2$), each of which consists of three entities in the second ($J_1 = J_2 = 3$) and two entities in the third mode ($K_1 = K_2 = 2$). In this example, the weighted relation model specifies two latent variables $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ as follows.

$$\boldsymbol{\gamma}_1 = \mathbf{X}_1\mathbf{w}_1 = \mathbf{X}_1 \begin{bmatrix} w_{111} \\ w_{121} \\ w_{131} \\ w_{112} \\ w_{122} \\ w_{132} \end{bmatrix}, \quad \boldsymbol{\gamma}_2 = \mathbf{X}_2\mathbf{w}_2 = \mathbf{X}_2 \begin{bmatrix} w_{211} \\ w_{221} \\ w_{231} \\ w_{212} \\ w_{222} \\ w_{232} \end{bmatrix}, \tag{6}$$

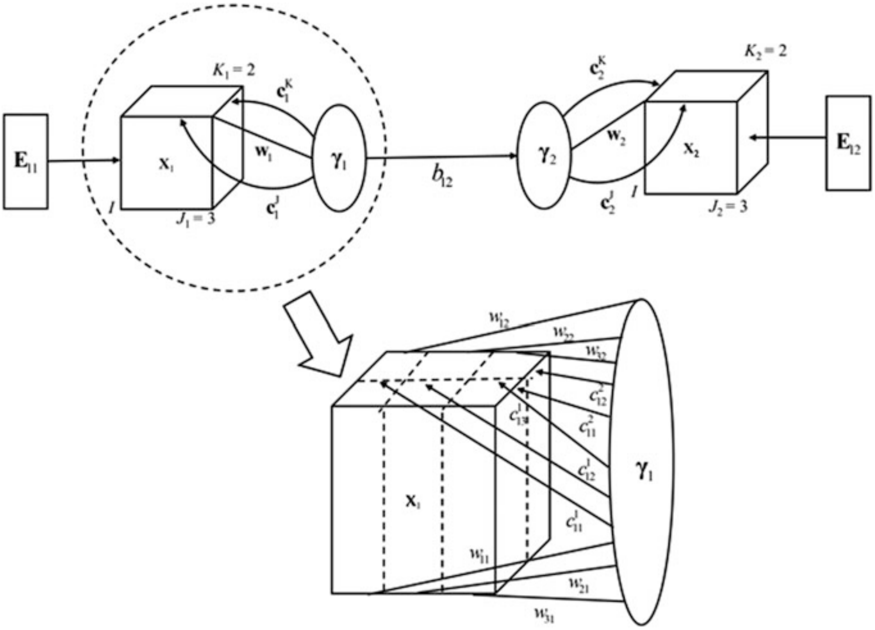


Fig. 1 A hypothetical three-way GSCA model with two latent variables

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are an I by 6 matrix unfolded from $\underline{\mathbf{X}}_1$ and $\underline{\mathbf{X}}_2$ by arranging an I by *3* frontal matrix of $\underline{\mathbf{X}}_1$ and $\underline{\mathbf{X}}_2$ next to one another, respectively. In the measurement model, $\mathbf{X}_1$ and $\mathbf{X}_2$ are assumed to load on $\boldsymbol{\gamma}_1$ on $\boldsymbol{\gamma}_2$, and hence their loading vectors $\mathbf{c}_1^{\mathrm{J}}$, $\mathbf{c}_1^{\mathrm{K}}$, $\mathbf{c}_2^{\mathrm{J}}$, and $\mathbf{c}_2^{\mathrm{K}}$ in (2) would be written as

$$\mathbf{c}_1^{\mathrm{J}} = \begin{bmatrix} c_{11}^{\mathrm{J}} \\ c_{12}^{\mathrm{J}} \\ c_{13}^{\mathrm{J}} \end{bmatrix}, \mathbf{c}_1^{\mathrm{K}} = \begin{bmatrix} c_{11}^{\mathrm{K}} \\ c_{12}^{\mathrm{K}} \end{bmatrix}, \mathbf{c}_2^{\mathrm{J}} = \begin{bmatrix} c_{21}^{\mathrm{J}} \\ c_{22}^{\mathrm{J}} \\ c_{23}^{\mathrm{J}} \end{bmatrix}, \mathbf{c}_2^{\mathrm{K}} = \begin{bmatrix} c_{21}^{\mathrm{K}} \\ c_{22}^{\mathrm{K}} \end{bmatrix}. \tag{7}$$

The structural model in Fig. 1 hypothesizes the effect of $\boldsymbol{\gamma}_1$ on $\boldsymbol{\gamma}_2$ and $\mathbf{B}$ in (5) can be expressed as

$$\mathbf{B} = \begin{bmatrix} 0 & b_{12} \\ 0 & 0 \end{bmatrix}. \tag{8}$$

To estimate the parameters of three-way GSCA (weights, loadings, and path coefficients), we aim to minimize the following least squares criterion:

$$\phi = \sum_{p=1}^{P} \mathrm{SS}\left(\mathbf{X}_p - \boldsymbol{\gamma}_p\left(\mathbf{c}_p^{\mathrm{K}} \otimes \mathbf{c}_p^{\mathrm{J}}\right)'\right) + \mathrm{SS}(\boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{B}), \tag{9}$$

subject to $\boldsymbol{\gamma}_p'\boldsymbol{\gamma}_p = 1$, $\mathbf{c}_p^{\mathrm{J}'}\mathbf{c}_p^{\mathrm{J}} = 1$, and $\mathbf{c}_p^{\mathrm{K}'}\mathbf{c}_p^{\mathrm{K}} = 1$, where SS($\mathbf{M}$) = tr($\mathbf{M}'\mathbf{M}$).

An Alternating Least Squares (ALS) algorithm (de Leeuw et al. 1976) is developed to minimize (9). It alternates three steps until convergence: each set of the unknown parameters, $\mathbf{w}_p$, $\mathbf{c}_p^{\mathrm{J}}$ and $\mathbf{c}_p^{\mathrm{K}}$, and $\mathbf{B}$, is updated alternately while the other sets are fixed. A detailed description of the ALS algorithm is provided in the Appendix.

As in GSCA, we can measure an overall measure of fit, called FIT (Hwang and Takane 2004). This index shows how much variance of all observed and latent variables is accounted for by the specified model. It is calculated as

$$\mathrm{FIT} = 1 - \frac{\sum_{p=1}^{P} \mathrm{SS}\left(\mathbf{X}_p - \boldsymbol{\gamma}_p\left(\mathbf{c}_p^{\mathrm{K}} \otimes \mathbf{c}_p^{\mathrm{J}}\right)'\right) + \mathrm{SS}(\boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{B})}{\sum_{p=1}^{P} \mathrm{SS}\left(\mathbf{X}_p\right) + \mathrm{SS}(\boldsymbol{\Gamma})}. \tag{10}$$

The FIT index would range from 0 to 1, and the larger FIT index, the more variation of endogenous variables is explained by the model.

In three-way GSCA, the bootstrap method (Efron 1982) is used to estimate the standard errors or confidence intervals of parameter estimates, which can be used for testing their statistical significance.

3 A Simulation Study

We performed a Monte Carlo simulation to investigate the parameter recovery capability of three-way GSCA. For the simulation study, we considered two three-way data sets ($P = 2$), each of which consisted of four entities in the second and third modes ($J_p = K_p = 4$). We chose parameter values as follows: all loadings in $\mathbf{c}_p^{\mathrm{J}}$ and $\mathbf{c}_p^{\mathrm{K}}$ were equal to 0.5; all weights in $\mathbf{w}_p$ were 0.2; and the path coefficient in $\mathbf{B}$ was 0.3. The prescribed loadings, weights, and path coefficient were combined into $\mathbf{V} = [\mathbf{I} \ \ \mathbf{W}]$ and $\mathbf{A} = [\mathbf{C} \ \ \mathbf{B}]$, where $\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_2 \end{bmatrix}$ and $\mathbf{C} = \begin{bmatrix} \mathbf{h}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{h}_2' \end{bmatrix}$ whose diagonal element is calculated as $\mathbf{h}_p' = \left(\mathbf{c}_p^{\mathrm{K}} \otimes \mathbf{c}_p^{\mathrm{J}}\right)'$. We aligned the two three-way data sets as $\mathbf{X}^* = [\mathbf{X}_1 \ \ \mathbf{X}_2]$ and re-expressed (4) in matrix notation as follows: $\mathbf{X}^*(\mathbf{V} - \mathbf{WA}) = \mathbf{E}$, where $\mathbf{E} = [\mathbf{E}_1 \ \ \mathbf{E}_2]$. In this study, each column of $\mathbf{E}$ was assumed to follow the standard normal distribution. Then, data were generated from $\mathbf{X}^* = \mathbf{EQ}'(\mathbf{QQ}')^{-1}$, where $\mathbf{Q} = \mathbf{V} - \mathbf{WA}$.

We considered five levels of sample size (i.e., the number of entities in the first mode): $I = 50$, 100, 200, 500, and 1000. At each sample size, 500 replications were obtained, thus yielding a total of 2,500 data sets. Three-way GSCA was applied to each of the generated data sets to estimate the four sets of parameters (i.e., $\mathbf{w}_p$, $\mathbf{c}_p^{\mathrm{J}}$ and $\mathbf{c}_p^{\mathrm{K}}$, and $\mathbf{B}$). To evaluate parameter recovery of three-way GSCA, we computed the average relative biases, standard deviations, and mean squared errors (MSE) of the estimates.

3.1 Results

Table 1 provides the average relative biases, standard deviations, and MSE of the estimates across the five different sample sizes. In this study, we regarded absolute values of relative bias greater than 10% as indicative of an unacceptable degree of bias (Bollen et al. 2007; Lei 2009).

As shown in Table 1, on average, when $I = 50$, all sets of estimates showed large relative biases (greater than 10%), indicating that the estimates of $\mathbf{w}_p$, $\mathbf{c}_p^{\mathrm{J}}$ and $\mathbf{c}_p^{\mathrm{K}}$ were positively biased, whereas those of $\mathbf{B}$ were negatively biased. On the other hand, the relative biases of the estimates became smaller with sample sizes, and when $I \geq 200$, they were considerably smaller than 10% in absolute value. The standard deviations of the estimates became smaller when the sample size increased. In regards to the MSE of the estimates, on average, although the MSE values tended to be somewhat large at $I = 50$, they decreased and became quite close to zero as the sample size increased. Overall, these simulation results suggested that three-way GSCA seems to recover population parameters reasonably well unless the sample size was too small ($I \leq 50$).

Table 1 The average values of relative biases (rbias), standard deviations (SD), and mean square errors (MSE) of parameter estimates across different sample sizes

Parameter	Sample size	rbias	SD	MSE
W	I = 50	40.92	0.15	0.082
	I = 100	23.38	0.12	0.017
	I = 200	8.31	0.09	0.007
	I = 500	0.79	0.05	0.002
	I = 1000	−1.24	0.03	0.001
$\mathbf{c}^{\mathrm{J}}$	I = 50	22.47	0.31	0.112
	I = 100	11.80	0.23	0.059
	I = 200	4.32	0.14	0.022
	I = 500	1.36	0.08	0.007
	I = 1000	0.64	0.06	0.003
$\mathbf{c}^{\mathrm{K}}$	I = 50	22.69	0.32	0.113
	I = 100	12.94	0.24	0.065
	I = 200	5.20	0.16	0.026
	I = 500	1.61	0.09	0.008
	I = 1000	0.75	0.06	0.004
B	I = 50	−18.13	0.30	0.093
	I = 100	3.47	0.16	0.025
	I = 200	3.62	0.09	0.008
	I = 500	6.26	0.04	0.002
	I = 1000	8.49	0.03	0.002

4 An Empirical Application

The present example is based on a sub-sample of children born in 1982 from the National Longitudinal Survey of Youth 1979-Children (NLSY79-C), a longitudinal study following up children of female participants of the NLSY79 every 2 years starting in 1986 (Center for Human Resource Research 2004). It contains three sets of three-way data, which are an array of subjects by variables by time points. More specifically, as described in Table 2, each latent variable in the model was assumed to be linked to a set of observed variables.

The first latent variable *problem behavior* was assumed to be associated with six variables (ANTI, ANX, DEP, HEAD, HYPR, and PEER), measured by Behavior Problems Index (Peterson and Zill 1986) across five different time points between 6–14 years of age with a 2-year interval. The second latent variable was *home environment* with two age-standardized variables (COGNZ and EMOTZ), measured using the Home Observation for Measurement of the Environment (Bradley and Caldwell 1984) across the five time points. The third latent variable *cognitive performance* with three variables (MATHZ, RECOGZ, and COMPZ) was measured by the Peabody Individual Achievement Test across the five time points. Higher values of *problem behavior* represent greater extents of misbehaviors, whereas those of *home environment* and *cognitive performance* indicate more stable

Table 2 A summary of latent and observed variables for the national longitudinal survey of youth 1979-Children (NLSY79-C) data

Latent variables	Observed variables
Problem behavior	ANTI: antisocial
	ANX: anxious/depressed
	DEP: dependent
	HEAD: headstrong
	HYPR: hyperactive
	PEER: peer conflict/withdrawn
Home environment	COGNZ: cognitive stimulation
	EMOTZ: emotional support
Cognitive performance	MATHZ: math
	RECOGZ: reading recognition
	COMPZ: reading comprehension (total)

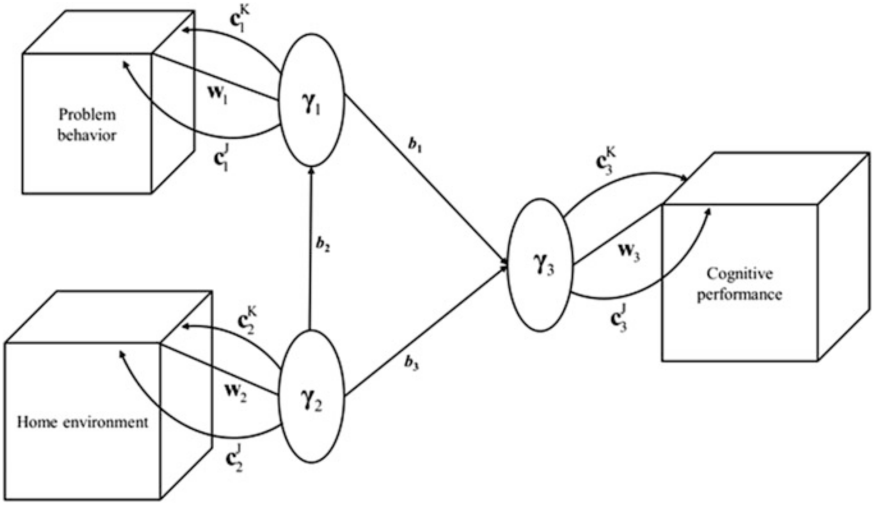


Fig. 2 A three-way GSCA model for the National Longitudinal Survey of Youth 1979-Children (NLSY79-C) data

emotional states and higher levels of competence. Figure 2 displays the hypothesized relationships among the three latent variables in the structural model. It was assumed that *problem behavior* influenced *cognitive performance*, and *home environment* was also assumed to affect both *problem behavior* and *cognitive performance.*

Tables 3, 4 and 5 present the estimates of weights, loadings, and path coefficients obtained from three-way GSCA. In three-way GSCA, a weight is estimated for the j_pth variable at the k_pth time point, which shows the contribution of the variable to defining its latent variable at a particular time point.

Table 3 Weight estimates obtained from three-way GSCA for the national longitudinal survey of youth 1979-Children (NLSY79-C) data

Latent	Variables	Time points	Estimate	Latent	Variables	Time points	Estimate
Problem behavior	ANTI	T1	0.07	Home environment	COGNZ	T1	0.03
		T2	0.06			T2	0.04
		T3	0.03			T3	0.01
		T4	0.06			T4	0.10
		T5	0.07			T5	0.10
	ANX	T1	0.04		EMOTZ	T1	0.10
		T2	0.09			T2	0.06
		T3	0.07			T3	0.16
		T4	0.02			T4	0.14
		T5	0.07			T5	0.07
	DEP	T1	0.06	Cognitive performance	MATHZ	T1	0.11
		T2	0.05			T2	0.10
		T3	0.08			T3	0.12
		T4	0.05			T4	0.11
		T5	0.05			T5	0.08
	HEAD	T1	0.05		RECOGZ	T1	0.13
		T2	0.09			T2	0.20
		T3	0.02			T3	0.26
		T4	0.03			T4	0.16
		T5	0.05			T5	0.21
	HYPR	T1	0.06		COMPZ	T1	0.13
		T2	0.07			T2	0.11
		T3	0.06			T3	0.06
		T4	0.05			T4	0.14
		T5	0.07			T5	0.09
	PEER	T1	0.05				
		T2	0.04				
		T3	0.05				
		T4	0.07				
		T5	0.05				

Table 4 presents the loading estimates and their 95% confidence intervals obtained from three-way GSCA. The loading estimates signify how each latent variable is associated with entities in the second and the third mode. As shown in the table, *problem behavior* was positively and statistically significantly related to their corresponding entities in the second mode indicating that higher values of problem behavior represented greater extents of misbehaviors (e.g., more antisocial, anxious, and/or dependent). It had the highest association with the variables HYPR and HEAD, followed by ANTI, ANX, PEER, and DEP. This latent variable was

Table 4 Loading estimates obtained from three-way GSCA for the national longitudinal survey of youth 1979-Children (NLSY79-C) data

Latent	Entities in the second mode			Entities in the third mode		
	Variables	Estimate	95% CI	Time points	Estimate	95% CI
Problem behavior	ANTI (c^{J}_{11})	0.45	0.39–0.50	T1 (c^{K}_{11})	0.42	0.320.52
	ANX (c^{J}_{12})	0.35	0.28–0.42	T2 (c^{K}_{12})	0.45	0.40–0.51
	DEP (c^{J}_{13})	0.30	0.23–0.36	T3 (c^{K}_{13})	0.48	0.39–0.56
	HEAD (c^{J}_{14})	0.49	0.44–0.53	T4 (c^{K}_{14})	0.44	0.350.52
	HYPR (c^{J}_{15})	0.50	0.45–0.55	T5 (c^{K}_{15})	0.44	0.38–0.51
	PEER (c^{J}_{16})	0.31	0.19–0.39			
Home environment	COGNZ (c^{J}_{21})	0.51	0.46–0.57	T1 (c^{K}_{21})	0.15	0.03–0.25
	EMOTZ (c^{J}_{21})	0.65	0.61–0.69	T2 (c^{K}_{22})	0.49	0.40–0.56
				T3 (c^{K}_{23})	0.51	0.46–0.56
				T4 (c^{K}_{24})	0.49	0.43–0.56
				T5 (c^{K}_{25})	0.49	0.43–0.57
Cognitive performance	MATHZ (c^{J}_{31})	0.56	0.51–0.59	T1 (c^{K}_{31})	0.47	0.37–0.54
	RECOGZ (c^{J}_{32})	0.79	0.71–0.88	T2 (c^{K}_{32})	0.55	0.48–0.64
	COMPZ (c^{J}_{33})	0.62	0.47–0.71	T3 (c^{K}_{33})	0.51	0.41–0.60
				T4 (c^{K}_{34})	0.32	0.21–0.41
				T5 (c^{K}_{35})	0.35	0.26–0.42

also positively and statistically significantly related to the five time points in the third mode. All the loading estimates for the time points were large, which was consistent with that those who exhibited problematic behaviours tended to continue to show these behaviors over time (Biederman et al. 2001). *Home environment* was positively and statistically significantly related to two observed variables in the second mode (COGNZ and EMOTZ), suggesting that its higher values indicated more cognitive and emotional supports at home. It was also positively and statistically significantly associated with all the time points in the third mode, although the first time point (T1) was less strongly associated with the latent variable than the other time points. Similarly, *cognitive performance* was positively and statistically

Table 5 Path coefficients' estimates obtained from three-way GSCA for the national longitudinal survey of youth 1979-Children data

Path coefficient		Estimate	95% CI
Problem behavior → Cognitive performance	(b_1)	−0.20	−0.46 to −0.05
Home environment → Problem behavior	(b_2)	−0.16	−0.42–0.14
Home environment → Cognitive performance	(b_3)	0.50	0.27–0.70

significantly related to all the variables in the second mode, indicating that its higher values represented higher levels of competence in mathematics and reading comprehension. It was most highly associated with RECOGZ, followed by COMPZ and MATHZ. It was also positively and statistically significantly related to all the time points in the third mode, although it was more highly correlated with earlier time points (e.g., T1–T3).

Table 5 displays the estimated path coefficients and their 95% confidence intervals. *Problem behavior* had a negative and statistically significant effect on *cognitive performance*, suggesting that children with a higher level of *problem behavior* were more likely to have a disrupted performance on cognitive tasks. *Home environment* had a negative yet statistically non-significant impact on *problem behavior*, whereas it had a positive and statistically significant effect on *cognitive performance*. This indicates that children in more stimulating and supportive environments were more likely to show better cognitive functioning, which is consistent with previous studies (Totsika and Sylva 2004).

5 Concluding Remarks

We generalized GSCA to the analysis of three-way data. Three-way GSCA enables to describe the directional relationships among latent variables as well as the relationships between entities in the second and third modes and the latent variables. A simulation study was conducted to evaluate the parameter recovery capacity of three-way GSCA. Three-way GSCA was found to recover parameters in a given model sufficiently well unless the sample size was too small. The usefulness of the proposed approach was also demonstrated through the analysis of real data. Besides investigating the interrelations among observed and latent variables, three-way GSCA enabled to examine which entities in the third mode were highly related to latent variables.

We may extend three-way GSCA to further improve its applicability. For example, we may extend three-way GSCA to accommodate so-called functional data (Ramsay and Silverman 2005). When a mode's responses can be sequenced along a continuum, such as time, frequency, or spatial location, and are intensively recorded at more than a handful points, it may be more natural to view them as a single connected entity or a function varying over the continuum. We can generalize three-way GSCA to permit observed responses in a mode to be functional rather than multivariate.

At present, three-way GSCA estimates all parameters by aggregating data across entities in the first mode (e.g., subjects) under the assumption that such entities are drawn from a homogenous population. Nonetheless, such an assumption may be easily violated in practice, and rather it is more plausible to assume that there exist heterogeneous subgroups or clusters of the population (e.g., Mun et al. 2008). To address this issue, we may combine three-way GSCA with a clustering method (e.g., non-overlapping *k*-means clustering (Hartigan and Wong 1979) or fuzzy *c*-means clustering (Bezdek 1974) in order to allow estimating cluster memberships of entities in the first mode as well as cluster-specific parameters.

Appendix

The ALS algorithm repeats the following three steps until convergence.

<u>*Step 1*</u>. Update weights ($\mathbf{w}_p$'s) for fixed $\mathbf{c}_p^{\mathrm{J}}$, $\mathbf{c}_p^{\mathrm{K}}$, and $\mathbf{B}$. This is equivalent to minimizing

$$\begin{aligned}\phi &= \sum_{p=1}^{P} \mathrm{SS}\left(\mathbf{X}_p - \boldsymbol{\gamma}_p(\mathbf{c}_p^{\mathrm{K}} \otimes \mathbf{c}_p^{\mathrm{J}})'\right) + \mathrm{SS}\left(\boldsymbol{\gamma}_p\mathbf{e}_p' + \boldsymbol{\Gamma}^{(-p)} - \boldsymbol{\gamma}_p\mathbf{b}_p' - \boldsymbol{\Gamma}^{(-p)}\mathbf{B}^{(-p)}\right) \\ &= \sum_{p=1}^{P} \mathrm{SS}\left(\mathbf{X}_p - \boldsymbol{\gamma}_p(\mathbf{c}_p^{\mathrm{K}} \otimes \mathbf{c}_p^{\mathrm{J}})'\right) + \mathrm{SS}(\boldsymbol{\gamma}_p\mathbf{t}_p - \boldsymbol{\Delta}_p) \\ &= \sum_{p=1}^{P} \mathrm{SS}\left(\mathbf{X}_p - \boldsymbol{\gamma}_p\mathbf{q}_p'\right) + \mathrm{SS}(\boldsymbol{\gamma}_p\mathbf{t}_p - \boldsymbol{\Delta}_p),\end{aligned} \tag{A.1}$$

subject to $\boldsymbol{\gamma}_p'\boldsymbol{\gamma}_p = 1$, where $\mathbf{q}_p = (\mathbf{c}_p^{\mathrm{K}} \otimes \mathbf{c}_p^{\mathrm{J}})$, $\mathbf{t}_p = \mathbf{e}_p' - \mathbf{b}_p'$, and $\boldsymbol{\Delta}_p = \boldsymbol{\Gamma}^{(-p)}\mathbf{B}^{(-p)} - \boldsymbol{\Gamma}^{(-p)}$. In (A.1), $\boldsymbol{\Gamma}^{(-p)}$ and $\mathbf{B}^{(-p)}$ indicate $\boldsymbol{\Gamma}$ and $\mathbf{B}$, whose columns are all zero vectors except the *p*th column, respectively, and $\mathbf{e}_p'$ indicates a 1 by *P* vector, whose elements are all zero except the *p*th element being unity. Based on (1), (A.1) can be re-expressed as

$$\begin{aligned}\phi &= \sum_{p=1}^{P} \mathrm{SS}\left(\mathbf{X}_p - \mathbf{X}_p\mathbf{w}_p\mathbf{q}_p'\right) + \mathrm{SS}(\mathbf{X}_p\mathbf{w}_p\mathbf{t}_p - \boldsymbol{\Delta}_p) \\ &= \sum_{p=1}^{P} \left(\mathrm{tr}(\mathbf{X}'{}_p\mathbf{X}_p) - 2\mathbf{w}_p'\mathbf{X}_p'\mathbf{X}_p\mathbf{q}_p + \mathbf{w}_p'\mathbf{X}_p'\mathbf{X}_p\mathbf{w}_p\mathbf{q}_p'\mathbf{q}_p\right) \\ &\qquad + \mathbf{w}_p'\mathbf{X}_p'\mathbf{X}_p\mathbf{w}_p\mathbf{t}_p\mathbf{t}_p' - 2\mathbf{w}_p'\mathbf{X}_p'\Delta_p\mathbf{t}_p' + \mathrm{tr}\left(\boldsymbol{\Delta}_p'\boldsymbol{\Delta}_p\right).\end{aligned} \tag{A.2}$$

Solving $\frac{\partial\phi}{\partial\mathbf{w}_p} = \mathbf{0}$, $\mathbf{w}_p$ is updated by

$$\hat{\mathbf{w}}_p = \left(\mathbf{q}_p\mathbf{q}_p'\mathbf{X}_p'\mathbf{X}_p + \mathbf{t}_p\mathbf{t}_p'\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\left(\mathbf{X}_p'\mathbf{X}_p\mathbf{q}_p + \mathbf{X}_p'\boldsymbol{\Delta}_p\mathbf{t}_p'\right). \tag{A.3}$$

Subsequently, $\boldsymbol{\gamma}_p$ is updated by $\gamma_p = \mathbf{X}_p\widehat{\mathbf{w}}_p$ and normalized to satisfy the constraint $\boldsymbol{\gamma}_p'\boldsymbol{\gamma}_p = 1$.

<u>*Step 2*</u>. Update $\mathbf{c}_p^{\mathrm{J}}$ and $\mathbf{c}_p^{\mathrm{K}}$ for fixed $\mathbf{w}_p$ and $\mathbf{B}$. This is equivalent to applying parallel factor analysis (PARAFAC) (Harshman 1970), subject to $\mathbf{c}_p^{\mathrm{J}'}\mathbf{c}_p^{\mathrm{J}} = 1$, and $\mathbf{c}_p^{\mathrm{K}'}\mathbf{c}_p^{\mathrm{K}} = 1$.

We can simply use the ALS algorithm for PARAFAC to update $\mathbf{c}_p^{\mathrm{J}}$ and $\mathbf{c}_p^{\mathrm{K}}$ (Acar and Yener 2009; Harshman 1970; Olivieri et al. 2015).

Step 3. Update $\mathbf{B}$ for fixed $\mathbf{w}_p$, $\mathbf{c}_p^{\mathrm{J}}$ and $\mathbf{c}_p^{\mathrm{K}}$. This is equivalent to minimizing

$$\begin{aligned} \phi_B &= \mathrm{SS}(\boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{B}) \\ &= \mathrm{SS}\left(\mathrm{vec}(\boldsymbol{\Gamma}) - \left(\mathbf{I}_p \otimes \boldsymbol{\Gamma}\right)\mathrm{vec}(\mathbf{B})\right) \\ &= \mathrm{SS}(\mathrm{vec}(\boldsymbol{\Gamma}) - \boldsymbol{\Psi}\mathbf{u}) \end{aligned} \tag{A.4}$$

where vec(**S**) is a super vector formed by stacking all columns of **S** in order, **u** denotes free parameters to be estimated in vec(**B**), and $\boldsymbol{\Psi}$ is a matrix consisting of the columns of $\mathbf{I}_p \otimes \boldsymbol{\Gamma}$ corresponding to the free parameters of vec(**B**) The estimate of **u** is obtained by

$$\hat{\mathbf{u}} = \left(\boldsymbol{\Psi}'\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}'\,\mathrm{vec}(\boldsymbol{\Gamma}). \tag{A.5}$$

Then, $\widehat{\mathbf{B}}$ is reconstructed from $\widehat{\mathbf{u}}$.

References

Acar, E., & Yener, B. (2009). Unsupervised multiway data analysis: A literature survey. *IEEE Transactions on knowledge and Data Engineering, 21*(1), 6–20.

Andersen, C. M., & Bro, R. (2003). Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *Journal of Chemometrics, 17*(4), 200–215.

Bezdek, J. C. (1974). Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology, 1*(1), 57–71.

Biederman, J., Monuteaux, M. C., Greene, R. W., Braaten, E., Doyle, A. E., & Faraone, S. V. (2001). Long-term stability of the child behavior check list in a clinical sample of youth with attention deficit hyperactivity disorder. *Journal of Clinical Child Psychology, 30*(4), 492–502.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research, 36*(1), 48–86.

Bradley, R. H., & Caldwell, B. M. (1984). The HOME inventory and family demographics. *Developmental Psychology, 20,* 315–320.

Bro, R. (1996). Multiway calidration. multilinear PLS. *Journal of Chemometrics, 10,* 47–61.

Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems, 38*(2), 149–171.

Bro, R., & Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics, 17*(1), 16–33.

Center for Human Resource Research. (2004). *NLSY79 Child and Young Adult Data Users Guide*. Columbus. OH: Ohio State University.

Christensen, J., Becker, E. M., & Frederiksen, C. S. (2005). Fluorescence spectroscopy and PARAFAC in the analysis of yogurt. *Chemometrics and Intelligent Laboratory Systems, 75*(2), 201–208.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research, 29*(3), 162–173.

de Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika, 41,* 471–514.

De Roover, K., Ceulemans, E., Timmerman, M. E., Vansteelandt, K., Stouten, J., & Onghena, P. (2012). Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock da ta. *Psychological Methods, 17*(1), 100.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: SIAM.

Ferrer, E., & McArdle, J. J. (2010). Longitudinal modeling of develo mental changes in psychological research. *Current Directions in Psycho logical Science, 19*(3), 149–154.

Germond, L., Dojat, M., Taylor, C., & Garbay, C. (2000). A cooperative framework for segmentation of MRI brain scans. *Artificial Intelligence in Medicine, 20*(1), 77–93.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics, 16*, 1–84.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 100–108.

Hwang, H., Desarbo, W. S., & Takane, Y. (2007a). Fuzzy clusterwise generalized structured component analysis. *Psychometrika, 72*(2), 181–198.

Hwang, H., Ho, M. R., & Lee, J. (2010). Generalized structured component analysis with latent interactions. *Psychometrika, 75*(2), 228–242.

Hwang, H., & Takane, Y. (2004). Generalized structured component analysis. *Psychometrika, 69* (1), 81–99.

Hwang, H., & Takane, Y. (2014). *Generalized structured component analysis: A component-based approach to structural equation modeling*. Boca Raton, FL: Chapman & Hall/CRC Press.

Hwang, H., Takane, Y., & Malhotra, N. (2007b). Multilevel generalized structured component analysis. *Behaviormetrika, 34*(2), 95–109.

Kroonenberg, P. M. (1987). Multivariate and longitudinal data on growing children. Solutions using a three-mode principal component analysis and some comparison results with other approaches. In F. M. J. M. P. J. Janssen (Ed.), *Data analysis. The ins and outs of solving real problems.* (*pp. 89–112*). New York: Plenum.

Kroonenberg, P. M. (2008). *Applied multiway data analysis* (Vol. 702). Wiley.

Kuze, T., Goto, M., Ninomiya, K., Asano, K., Miyazawa, S., Munekata, H., et al. (1985). A longitudinal study on development of adolescents' social attitudes. *Japanese Psychological Research, 27*(4), 195–205.

Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality & Quantity, 43*(3), 495–507.

Mun, E. Y., von Eye, A., Bates, M. E., & Vaschillo, E. G. (2008). Finding groups using model-based cluster analysis: Heterogeneous emotional self-regulatory processes and heavy alcohol use risk. *Developmental Psychology, 44*(2), 481.

Olivieri, A. C., Escandar, G. M., Goicoechea, H. C., & de la Peña, A. M. (2015). *Fundamentals and analytical applications of multi-way calibration* (Vol. 29). Elsevier.

Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology, 54*(1), 49–78.

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relation ships, and behavioral problems in children. *Journal of Marriage and the Family, 48*(2), 295–307.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer.

Tenenhaus, A., Le Brusquet, L., & Lechuga, G. (2015). Multiway Regularized Generalized Canonical Correlation Analysis. In *47èmes Journée de Statistique de la SFdS (JdS 2015).*

Thirion, B., & Faugeras, O. (2003). Dynamical components analysis of fMRI data through kernel PCA. *NeuroImage, 20*(1), 34–49.

Totsika, V., & Sylva, K. (2004). The home observation for measurement of the environment revisited. *Child and Adolescent Mental Health, 9*(1), 25–35.

Combining Factors from Different Factor Analyses Based on Factor Congruence

Anikó Lovik, Vahid Nassiri, Geert Verbeke and Geert Molenberghs

Abstract While factor analysis is one of the most often used techniques in psychometrics, comparing or combining solutions from different factor analyses can be cumbersome even though it is necessary in several situations. For example, when applying multiple imputation (to account for incompleteness) or multiple outputation (which can be used to deal with clustering in multilevel data) often tens or hundreds of results have to be combined into one final solution. While different solutions have been in use, we propose a simple and easy to implement solution to match factors from different analyses based on factor congruence. To demonstrate this method, the Big Five Inventory data collected under the auspices of the Divorce in Flanders study was analysed combining multiple outputation and factor analysis. This multilevel sample consists of 7533 individuals coming from 4460 families with about 10% of missing values.

Keywords Big Five Inventory · Divorce in Flanders · Factor analysis · Factor congruence · Factor matching · Modified Tucker's congruence coefficient (mTCC)

A. Lovik (✉) · V. Nassiri · G. Verbeke · G. Molenberghs
I-BioStat, KU Leuven—University of Leuven, Leuven, Belgium
e-mail: Aniko.Lovik@kuleuven.be
URL:http://ibiostat.be

V. Nassiri
e-mail: Vahid.Nassiri@kuleuven.be

G. Verbeke
e-mail: Geert.Verbeke@kuleuven.be

G. Molenberghs
e-mail: Geert.Molenberghs@uhasselt.be

G. Verbeke · G. Molenberghs
I-BioStat, Hasselt University, Diepenbeek, Belgium

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_18

1 Introduction

While factor analysis is one of the most often used techniques in psychometrics, comparing and/or combining solutions from different factor analyses can be cumbersome even though combining factors is necessary in several situations. Such situations include using factor analysis with multiple imputation (to account for incompleteness; Rubin 1976; Little and Rubin 2002; Carpenter and Kenward 2012) or with multiple outputation (which is a simple solution to deal with multilevel data when one wants to use methods that require independent observations; Follmann et al. 2003). In both situations, often tens or hundreds of results have to be combined into one final solution. Also, assessing the factorial invariance for factors from different studies or samples, for example in the case of a cross-cultural study, may be of interest.

Although several solutions have been proposed, such as a confirmatory maximum likelihood procedure (e.g., Jöreskog 1966) or methods based on procrustes rotations (e.g., Korth and Tucker 1975), they are often complicated and difficult to implement (Lorenzo-Seva and ten Berge 2006) or limited in use. For the latter, a good example is Cattell's index of proportionality, which assumes "that the variance of each factor in one experiment shall be different . . . from that of the corresponding factor in the second" (Cattell 1951 cited by Pinneau and Newhouse 1964, p. 276). Another, often occurring, disadvantage is that many of these methods change the factor loadings.

A possibility that does not affect the factor loadings would be to compare each pair of factors from the different analyses and to select the most similar pairs, for example by minimising the difference between the factor loadings. While this is certainly possible, we propose finding the matching pairs based on the modified Tucker's congruence coefficient (mTCC).

The original Tucker's congruence coefficient (TCC) has been around for nearly 70 years and has been in use ever since to assess factor similarity across samples for the same variables (Tucker 1951, Lorenzo-Seva and ten Berge 2006). It is merely the cosine of the angle of two uncentered vectors. The interpretation is quite arbitrary as it is difficult to establish reference values. Lorenzo-Seva and ten Berge (2006) found in an empirical study that 0.95 may be a suitable cut-off value. However, the TCC is sensitive to changes in signs over different analyses. This may result in overestimating/underestimating congruence when the signs of the variable pairs are predominantly the same/different (Pinneau and Newhouse 1964; Barrett 1986; Lovik et al. 2017). Another practical issue is related to negatively framed items (statements in a questionnaire which have an opposing meaning, a negative association with the factor they belong to) which results in factor loadings of the same magnitude with reversed sign, since the factor loadings are calculated based on the correlation/covariance matrix of all items. For these reasons, a modified congruence coefficient was proposed, which may be more useful in combining factor analyses (Lovik et al. 2017). This coefficient uses the absolute values of the products in the numerator of the TCC:

$$\psi(x,y) = \frac{\sum |x_i y_i|}{\sqrt{\sum x_i^2 \sum y_i^2}}. \tag{1}$$

The mTCC does not have a nice geometric interpretation but has several advantages. To begin with, all of the advantages of the TCC are preserved: the new coefficient is insensitive to scalar multiplication of x and y, and to changes in the sign of any pair (x, y) but sensitive to additive constants. Furthermore, it is also still a continuous function of x_i and y_i. Obviously, the mTCC is always at least as large as the TCC and it varies between 0 and 1. It should be noted that there is no direct relationship between the TCC and the mTCC. The reason it may be more useful for combining factors is that very low values (between -1 and -0.90) normally arise when two factors are very similar (in interpretation) but all signs are reversed for one factor compared to the other. In such cases, the TCC would erroneously reject the possibility that the two are equal, while the modified coefficient results in a value above 0.90.

The method to combine M factor analyses with k factors each based on mTCC (or TCC) is very simple:

Notation Suppose we have M sets of factor loading matrices each with k factors: $L_1, L_2, \ldots, L_M$. We denote a re-ordered set of factor loadings by $\widetilde{L}_i$ ($i = 1, \ldots, M$) and $\widehat{L}_r$ represents the combination of r sets of factor loadings. Furthermore, a congruence matrix is a symmetric $k \times k$ matrix containing the congruence coefficients between all possible pairs of factors of two sets of factor loadings (from two separate factor analyses).

Algorithm $\widetilde{L}_r$, the re-ordered L_r based on L_s, is computed as follows:

1. The $k \times k$ congruence matrix for L_r and L_s is constructed.
2. The location of maximum congruence coefficient in each column of the congruence matrix is determined.
3. $\widetilde{L}_r$ is constructed by re-ordering the columns of L_r based on the maximum locations obtained from Step 2.

Now the combination algorithm is as follows:

1. **Begin**. $\widehat{L}_1 = L_1$.
2. **Iteration**. $\widehat{L}_r = \frac{(r1)\widehat{L}_{r1}+\widetilde{L}_r}{r}$, for $r = 2, \ldots, M$.

We demonstrate the method on a dataset that contains responses to a Big Five personality inventory.

2 Motivating Dataset: The Divorce in Flanders Study

The dataset we use for analysis is a subsample from the *Divorce in Flanders* (DiF) project, which contains a sample of marriages registered between 1971 and 2008 with oversampling of divorces (1/3 intact and 2/3 dissolved marriages at the sampling date) drawn from the Belgian National Register (see details in Mortelmans et al.

2011). Family members across three generations were surveyed during the original data collection, more than 10,000 people. The data were collected in 2008 and the validated Dutch language version of the BFI was administered among a battery of tests with the aim of studying the phenomenon of divorce in families. In this paper we use data from 4460 families, 7533 people in total (3362 mothers, 2920 fathers and 1251 children). We excluded new partners of the ex-spouses and parents of the selected sample. One of the main advantages of the data collection is the ability to assess, among others, the patterns of matching personality traits between family members, predicting personality traits by studying the intergenerational transmission of personality, associating personality traits with fertility and personality traits with divorce.

As part of this study the personality of each participant was assessed with the validated Dutch version (Denissen et al. 2008) of the Big Five Inventory (BFI, John and Srivastava 1999), a personality test which is a commonly used tool to assess personality measuring the five factors of personality (Neuroticism, Extraversion, Openness to Experience, Conscientiousness and Agreeableness; e.g. John and Srivastava 1999; Digman 1990). When clustering is not taken into account, the five factors tend to emerge clearly from the data (Lovik et al. 2017).

Participants were asked to rate their agreement with each item regarding their perceptions of themselves using a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The 44-item questionnaire contains 14 reversed items. The items were reversed before any analysis has taken place.

Since the DiF was a multi-actor where several family members from the same family were invited to participate, the observations are not independent. To account for clustering, we decided to use multiple outputation (Follmann et al. 2003). Multiple outputation is a within-cluster sampling method that “throws out excess data” by sampling exactly one observation from each cluster to create multiple subsets without clustering, which allows using statistical methods where observations are assumed to be independent. 1000 random subsets were generated from the original dataset using simple random sampling, thus one individual from each family was selected, resulting in 1000 samples of size 4460. On each of these 1000 datasets factor analysis with principal component extraction was performed and the results were rotated using direct oblimin rotation. The analyses were combined with the method described previously based on TCCs and mTCCS. We wanted to examine whether the order of the factor analyses had an effect on the final result. To this end, we randomly re-ordered the 1000 datasets and repeated the analysis. This process was repeated 10 times, the factor structure did not change.

3 Results

Descriptive statistics for TCCs and mTCCs are given in Table 1 for congruent and incongruent factors separately.

Figures 1 and 2 allow to compare the results based on either TCCs or mTCCs. As it can be seen in Fig. 1, although the five factor structure is quite clear in most of

Table 1 Descriptives of TCCs and mTCCs based on factor congruence

TCC	mTCC	Valid N	TCC				mTCC			
			Mean	Std	Min	Max	Mean	Std	Min	Max
Incongruent	Incongruent	19,975	−0.005	0.116	−0.334	0.533	0.431	0.068	0.282	0.702
Incongruent	Congruent	5	−0.738	0.029	−0.767	−0.701	0.772	0.020	0.748	0.794
Congruent	Incongruent	5	0.446	0.033	0.397	0.480	0.701	0.041	0.628	0.723
Congruent	Congruent	4,990	0.993	0.011	0.830	1.000	0.994	0.009	0,885	1.000
All groups		24,975	0.194	0.413	−0.767	1.000	0.544	0.233	0.282	1.000

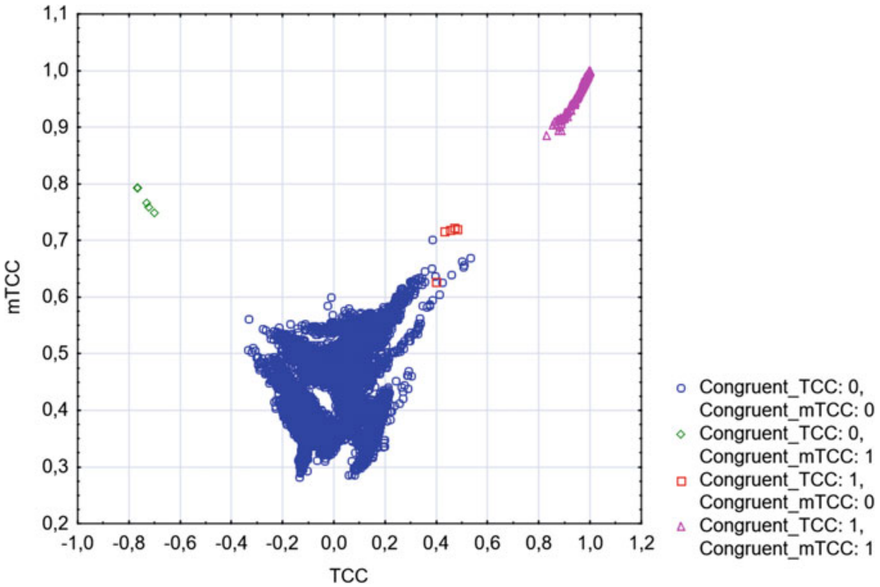


Fig. 1 Scatterplot of TCCs and mTCCs in the DiF study

the datasets, the congruence coefficients show quite a bit of variability. The TCCs range from −0.7671 to 0.9996, while the mTCCs are between 0.2822 and 0.9996. The equality of the maximum is not surprising, the higher the TCC the smaller the difference with the associated mTCC. In case all factor loadings are positive, the two coefficients are, of course, equal. This happens in 2% of the cases in our example.

There are 443/999 pairs which have a different initial factor ordering than the reference. This shows that finding the right factor order before combining factor analyses is extremely important.

Figure 1 shows that out of the 999 × 5 factor matches, five analyses will be ordered differently based on TCCs compared to mTTCs. The reason is that the correct matches have high negative loadings and TCCs select the highest positive match resulting in assigning two factors to one. Nevertheless, in this example matching based on TCCs works in 99.5%, based on mTCCs in 100% of the cases. For this reason, Table 2 presents the final factor loadings after combining the 1000 datasets based on mTCCs. Should we choose to ignore the mismatches and still combine factors based on TCCs, the difference between the result without mismatches (based on mTCCs) and the result based on TCCs is smaller than 10^{-8} for each factor loading. Of course, one has to take into account that the sample was rather big, especially for an exploratory factor analysis. In fact, when combining multiple outputation with split-sampling, resulting in factor analyses based on samples of size 1000 instead of 4460, the number of mismatches is 97/999 for factor matching based on TCCs and 52/999 in case of mTCCs, with an overlap of 45 cases. The difference between the final analyses is still below 10^{-3} for all factor loadings. The mismatches with the analyses based on mTCCs are due to an increased number of high cross-loadings for at least two factors. In fact, the factor structure is quite unclear for these analyses.

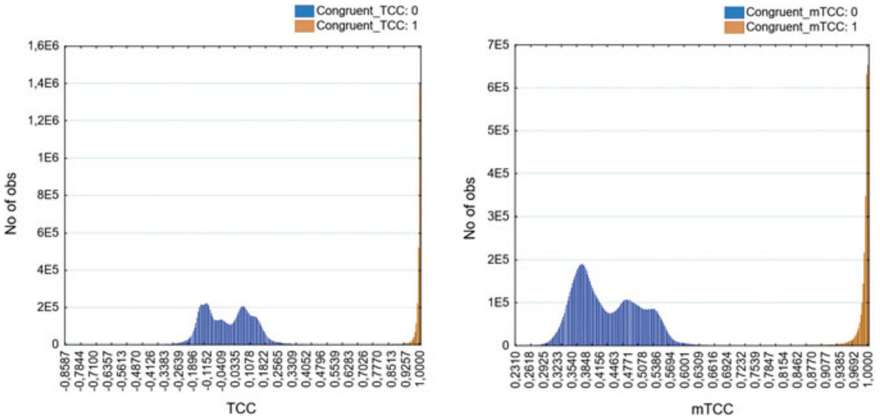


Fig. 2 Histogram of TCCs (left) and mTCCs (right) based on congruence

Table 2 Final factor loadings (combining 1000 analyses based on mTCCs)

No.	Questionnaire item	N	E	O	C	A
19.	…worries a lot	**0.71**	−0.06	0.04	0.10	−0.03
14.	…can be tense	**0.70**	−0.05	0.07	0.17	−0.15
9r.	…is relaxed, handles stress well	**0.63**	−0.08	−0.17	−0.15	0.03
39.	…gets nervous easily	**0.75**	−0.01	0.00	0.02	−0.10
24r.	…is emotionally stable, not easily upset	**0.49**	−0.05	−0.16	−0.20	0.02
34r.	…remains calm in tense situations	**0.56**	0.08	−0.19	−0.22	−0.05
4.	…is depressed, blue	**0.45**	−0.29	0.06	−0.02	−0.12
29.	…can be moody	**0.37**	−0.02	0.06	0.09	**−0.48**
1.	…is talkative	0.10	**0.68**	0.02	0.07	−0.06
21r.	…tends to be quiet	**−0.06**	**0.77**	−0.15	−0.09	−0.04
16.	…generates a lot of enthusiasm	−0.02	**0.50**	0.28	0.28	0.05
36.	…is outgoing, sociable	0.10	**0.53**	0.14	0.11	**0.31**
6r.	…is reserved	−0.20	**0.67**	−0.19	−0.05	0.04
31r.	…is sometimes shy, inhibited	**−0.30**	**0.58**	−0.25	−0.03	−0.02
11.	…is full of energy	−0.25	**0.35**	0.13	**0.34**	−0.04
26.	…has an assertive personality	−0.26	**0.32**	0.06	**0.35**	−0.25
40.	…likes to reflect, play with ideas	−0.04	0.00	**0.46**	**0.38**	−0.07
25.	…is inventive	−0.19	0.13	**0.45**	**0.33**	−0.11
30.	…values artistic, aesthetic experiences	0.01	−0.17	**0.69**	0.08	0.12
5.	…is original, comes up with new ideas	−0.09	0.17	**0.43**	0.25	−0.11

(continued)

Table 2 (continued)

No.	Questionnaire Item	N	E	O	C	A
15.	…is ingenious, a deep thinker	0.09	−0.05	**0.36**	**0.48**	−0.15
20.	…has an active imagination	0.03	0.19	**0.54**	−0.03	−0.11
10.	…is curious about many different things	−0.12	0.22	**0.40**	0.26	−0.10
44.	…is sophisticated in art, music, or literature	−0.04	−0.12	**0.61**	−0.05	0.11
41r.	…has few artistic interests	−0.12	−0.15	**0.52**	−0.04	0.14
35r.	…prefers work that is routine	−0.25	0.01	0.19	−0.06	−0.14
3.	…does a thorough job	0.03	0.02	−0.03	**0.64**	−0.09
28.	…perseveres until the task is finished	−0.01	0.00	−0.04	**0.69**	0.03
18r.	…tends to be disorganized	−0.03	−0.08	**−0.45**	**0.48**	0.20
23r.	…tends to be lazy	−0.04	0.00	**−0.35**	**0.49**	0.21
13.	…is a reliable worker	0.06	0.07	0.04	**0.58**	0.04
33.	…does things efficiently	−0.03	0.00	0.01	**0.69**	0.07
38.	…makes plans and follows through with them	−0.08	0.16	0.09	**0.60**	−0.06
43r.	…is easily distracted	**−0.33**	−0.10	−0.28	**0.39**	0.17
8r.	…can be somewhat careless	0.03	−0.13	**−0.41**	**0.38**	0.24
32.	…is considerate and kind to almost everyone	0.18	0.19	0.22	0.20	**0.47**
17.	…has a forgiving nature	0.10	0.16	0.25	0.07	**0.41**
7.	…is helpful and unselfish with others	0.14	0.06	0.17	0.21	0.24
12r.	…starts quarrels with others	−0.25	−0.13	−0.01	0.03	**0.54**
37r.	…is sometimes rude to others	−0.13	−0.14	−0.03	0.04	**0.68**
27r.	…can be cold and aloof	0.02	0.27	−0.07	−0.06	**0.58**
22.	…is generally trusting	0.03	0.19	0.28	−0.10	**0.33**
2r.	…tends to find fault with others	−0.18	−0.19	−0.09	−0.02	**0.55**
42.	…likes to cooperate with others	0.04	**0.32**	0.12	0.18	0.23

N Neuroticism, *E* Extraversion, *O* Openness to Experience, *C* Conscientiousness, *A* Agreeableness

Figure 2 shows that both TCCs (left) and mTCCs (right) separate congruent and incongruent factors well.

The above mentioned results belong to the first analysis. As mentioned previously, the order of the analyses may influence the end result and for this reason, the entire analysis was repeated several times. We found no substantial differences.

4 Conclusion

In this paper, we used a modified Tucker's congruence coefficient to combine factor analyses by matching factors based on the maximum of all calculated mTCCs. In our example, both TCCs and mTCC work well in separating congruent and incongruent factors. However, it should be noted that this method may not work perfectly if the factor structure is not clear for more than a few of the factor analyses that need to be combined. Depending on the analysis, throwing out the "unclear" factor analyses may cause bias. Therefore, if mismatches happen one needs to assess the effect of keeping/deleting the analyses that cause the mismatch and a sensitivity analysis might prove useful.

References

Barrett, P. (1986). Factor comparison: An examination of three methods. *Personality and Individual Differences*, *7*(3), 327–340.

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. Wiley.

Denissen, J. J. A., Geenen, R., van Aken, M. A. G., Gosling, S. D., & Potter, J. (2008). Development and validation of a Dutch translation of the Big Five Inventory (BFI). *Journal of Personality Assessment*, *90*(2), 152–157.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417–440.

Follmann, D., Proschan, M., & Leifer, E. (2003). Multiple outputation: Inference for complex clustered data by averaging analyses from independent data. *Biometrics*, *59*(2), 420–429.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). New York: Guilford.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika*, *31*(2), 165–178.

Korth, B., & Tucker, L. R. (1975). The distribution of chance congruence coefficients from simulated data. *Psychometrika*, *40*(3), 361–372.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*(2), 57–64.

Lovik, A., Nassiri, V., Verbeke, G., & Molenberghs, G. (2017). *A modified Tucker's congruence coefficient*. Submitted.

Lovik, A., Nassiri, V., Verbeke, G., Molenberghs, G., & Sodermans, A. K. (2017). Psychometric properties and comparison of different techniques for factor analysis on the Big Five Inventory from a Flemish sample. *Personality and Individual Differences*, *117*, 122–129.

Mortelmans, D., Pasteels, I., Bracke, P., Matthijs, K., Bavel, V. J., & Van Peer, C. (2011). *Scheiding in Vlaanderen*. Leuven: Acco.

Pinneau, S. R., & Newhouse, A. (1964). Measures of invariance and comparability in factor analysis for fixed variables. *Psychometrika*, *29*(3), 271–281.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Tucker, L. R (1951). A method for synthesis of factor analysis studies (No. PRS-984). Princeton: Educational Testing Service.

On the Bias in Eigenvalues of Sample Covariance Matrix

Kentaro Hayashi, Ke-Hai Yuan and Lu Liang

Abstract Principal component analysis (PCA) is a multivariate statistical technique frequently employed in research in behavioral and social sciences, and the results of PCA are often used to approximate those of exploratory factor analysis (EFA) because the former is easier to implement. In practice, the needed number of components or factors is often determined by the size of the first few eigenvalues of the sample covariance/correlation matrix. Lawley (1956) showed that if eigenvalues of population covariance matrix are distinct, then each sample eigenvalue contains a bias of order 1/*N*, which is typically ignored in practice. This article further shows that, under some regulatory conditions, the order of the bias term is *p*/*N*. Thus, when *p* is large, the bias term is no longer negligible even when *N* is large.

Keywords Factor analysis • Principal component analysis • High dimension
Large *p* small *N*

K. Hayashi (✉)
Department of Psychology, University of Hawaii at Manoa,
2530 Dole Street, Sakamaki C400, Honolulu, HI 96822, USA
e-mail: hayashik@hawaii.edu

K.-H. Yuan
Department of Psychology, University of Notre Dame,
123A Haggar Hall, Notre Dame, IN 46556, USA
e-mail: kyuan@nd.edu

L. Liang
Department of Psychology, Florida International University,
11200 S.W. 8th Street, Miami, FL 33199, USA
e-mail: luliang@fiu.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_19

1 Introduction

Principal component analysis (PCA; Hotelling 1933) is a multivariate statistical technique for data reduction frequently employed in research in behavioral and social sciences. PCA has been a default dimension reduction technique in statistical software SPSS (IBM Corp. 2016), which is most widely used by researchers in social sciences. Anderson (1963) derived the asymptotic distribution of the eigenvalues and standardized eigenvectors of a sample covariance matrix when the observations follow a multivariate normal distribution whose covariance matrix can have eigenvalues with more than one multiplicity. Lawley (1956; see also Muirhead 1982) gave the formulas for the asymptotic expansion for both the mean and the variance of eigenvalues of the sample covariance matrix up to the order of $1/N$ when their population counterparts are distinct. The current work is an extension of Lawley's work. We show that the bias term in sample eigenvalues is of order p/N when the number of variables p is not negligible.

From a practical point of view, PCA is often used to approximate the results of exploratory factor analysis (EFA; see, e.g., Hwang and Takane 2004). It has been well known that PCA and EFA often yield approximately comparable loading matrices (cf., Velicer and Jackson 1990). Conditions under which the two matrices are close to each other have been studied extensively (Bentler and Kano 1990; Guttman, 1956; Krijnen, 2006; Schneeweiss and Mathes 1995). Because the computation for the estimates of PCA loadings is much simpler than that for the estimates of FA loadings in that the former is just an eigenvalue-eigenvector decomposition of the sample covariance matrix, it is attractive if PCA can be used as an approximation for FA especially when p is large.

2 Principal Component Analysis

Let $\mathbf{\Lambda}^+$ be the $p \times p$ matrix whose columns are the standardized eigenvectors corresponding to the eigenvalues of $\mathbf{\Sigma}$ in descending order; $\mathbf{\Omega}^+ = diag(\omega_1, \omega_2, \ldots, \omega_p)$ be the $p \times p$ diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{\Sigma}$, that is $\mathbf{\Sigma\Lambda}^+ = \mathbf{\Lambda}^+\mathbf{\Omega}^+$; $\mathbf{\Lambda}$ be the $p \times m$ matrix corresponding to the first m largest eigenvalues of $\mathbf{\Sigma}$, $\mathbf{\Omega}$ be the $m \times m$ diagonal matrix whose diagonal elements are the m largest eigenvalues of $\mathbf{\Sigma}$; and $\mathbf{\Omega}^{1/2}$ be the $m \times m$ diagonal matrix whose diagonal elements are the square root of those in $\mathbf{\Omega}$. Then the PCs with m elements are obtained as (c.f., Anderson 2003):

$$\boldsymbol{f} = \mathbf{\Lambda}' y, \tag{1}$$

where $\boldsymbol{y}$ is the vector of p manifest variables. Clearly, the PCs are uncorrelated with covariance matrix:

$$Cov(\boldsymbol{f}) = \boldsymbol{\Lambda}'\boldsymbol{\Sigma}\boldsymbol{\Lambda} = \boldsymbol{\Omega}. \tag{2}$$

When m is properly chosen, we have

$$\boldsymbol{\Sigma} \approx \boldsymbol{\Lambda}\boldsymbol{\Omega}\boldsymbol{\Lambda}' = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime}, \tag{3}$$

where

$$\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{\Omega}^{1/2} \tag{4}$$

is the $p \times m$ matrix of PCA loadings. If we define $\boldsymbol{f}^* = \boldsymbol{\Omega}^{-1/2}\boldsymbol{f}$, then $Cov(\boldsymbol{f}^*) = I_m$ and we can express PCA similar to EFA, that is,

$$\boldsymbol{y} = \boldsymbol{\Lambda}^*\boldsymbol{f}^* + \boldsymbol{\varepsilon}^*. \tag{5}$$

where $\boldsymbol{\varepsilon}^* = \boldsymbol{\Lambda}^-\boldsymbol{f}^-$, with $\boldsymbol{\Lambda}^-$ being the $p \times (p - m)$ matrix whose columns are the standardized eigenvectors corresponding to the $p - m$ smallest eigenvalues of $\boldsymbol{\Sigma}$, and $\boldsymbol{f}^- = \boldsymbol{\Lambda}^{-\prime}\boldsymbol{y}$. Obviously, $Cov(\boldsymbol{f}^*, \boldsymbol{\varepsilon}^*) = 0$.

3 Bias in Eigenvalues of Sample Covariance Matrix

Suppose that y has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, that is, $y \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a fixed value of p, Lawley (1956) showed that if the eigenvalues of $\boldsymbol{\Sigma}$ are distinct, that is, if $\omega_1 > \omega_2 > \ldots > \omega_p (> 0)$, then the mean of the i-th largest eigenvalue l_i of the sample covariance matrix $\boldsymbol{S}$ can be expanded as

$$E(l_i) = \omega_i + \frac{\omega_i}{n} \sum_{\substack{j=1 \\ j \neq i}}^{p} \frac{\omega_j}{\omega_i - \omega_j} + O(n^{-2}) = \omega_i + \frac{1}{n} f_i(\boldsymbol{\Omega}) + O(n^{-2}), \tag{6}$$

where $n = N - 1$ and

$$f_i(\boldsymbol{\Omega}) = \sum_{\substack{j=1 \\ j \neq i}}^{p} \frac{\omega_i\omega_j}{\omega_i - \omega_j} = \sum_{\substack{j=1 \\ j \neq i}}^{p} \omega_j\left(1 - \frac{\omega_j}{\omega_i}\right)^{-1} = \sum_{\substack{j=1 \\ j \neq i}}^{m} \omega_j\left(1 - \frac{\omega_j}{\omega_i}\right)^{-1} + \sum_{\substack{j=m+1 \\ j \neq i}}^{p} \omega_j\left(1 - \frac{\omega_j}{\omega_i}\right)^{-1} = f_{1i}(\boldsymbol{\Omega}) + f_{2i}(\boldsymbol{\Omega}) \tag{7}$$

See also Sect. 9.3 of Muirhead (1982, p. 388). The term $(1/n)f_i(\boldsymbol{\Omega})$ on the right-hand side of Eq. (6) dominates the bias. If $f_i(\boldsymbol{\Omega})$ in Eq. (7) is of order 1, the bias term is of order 1/*N*. This is the case with p fixed, because then p does not affect the order. More generally, when the number of variables p is negligible relative to the sample size N, the effect of $1/N$ dominates as N increases, and the sample eigenvalues are still asymptotically unbiased. However, when p is large or p/N is not negligible, we encounter a different situation. For example, a substantial bias is noted in by Arruda and Bentler (2017) in the context of varying values of N while holding p constant, in which the largest bias was found when N is the smallest.

Now, we show that the bias term is of order $O(p/N)$ when p is not negligible. To see this, in addition to the above assumption of distinct eigenvalues, we further assume the following:

(A1) For the m largest eigenvalues, $\omega_i = O(p) \to \infty$ as $p \to \infty$.
(A2) The m largest eigenvalues are well separated with each other. That is, for every different pair ω_j/ω_i does not converge to 1 as $p \to \infty$.
(A3) For the rest $p - m$ eigenvalues, $\omega_i = O(1)$ as $p \to \infty$.
(A4) The ratio $m/p \to 0$ as $p \to \infty$.

Then, regarding the order of the function $f_i(\mathbf{\Omega})$ with respect to the number of variables p, we proceed as follows:

Case 1: Suppose the subscript i is in $\{1, 2, \ldots, m\}$, that is, $i \leq m$.
(C1.1) If $j \leq m$ with $j \neq i$, both ω_i and ω_j are $O(p)$ by (A1), so that ω_j/ω_i is $O(1)$. Thus, $(1-\omega_j/\omega_i)^{-1}$ is still $O(1)$ by (A2), and finally, $\omega_j(1-\omega_j/\omega_i)^{-1} = O(p)\cdot O(1) = O(p)$. This leads to: $f_{1i}(\mathbf{\Omega}) = \sum_{\substack{j=1 \\ j\neq i}}^{m} \omega_j(1-\omega_j/\omega_i)^{-1} = (m-1)\cdot O(p) = O(p)$ by (A4).

(C1.2) Next, if $j \geq m+1$, $\omega_i = O(p)$ and $\omega_j = O(1)$ by (A1) and (A3), so $\omega_j/\omega_i = O(p^{-1}) \to 0$. Thus, $(1-\omega_j/\omega_i)^{-1} \to 1$ and $\omega_j(1-\omega_j/\omega_i)^{-1} = O(1)$ by (A3). By collecting $p - m$ such terms, along with (A4), $f_{2i}(\mathbf{\Omega}) = \sum_{j=m+1}^{p} \omega_j(1-\omega_j/\omega_i)^{-1} = (p-m)\cdot O(1) = O(p)$.

Combining (C1.1) and (C1.2) yields $f_i(\mathbf{\Omega}) = f_{1i}(\mathbf{\Omega}) + f_{2i}(\mathbf{\Omega}) = O(p) + O(p) = O(p)$. In summary, if $i \leq m$, in Eq. (6), the first term ω_i is $O(p)$, and the second term $(1/n)f_i(\mathbf{\Omega})$ is $O(p/N)$.

Case 2: Suppose the subscript i is in $\{m + 1, \ldots, p\}$, that is, $i \geq m+1$.
(C2.1) If $j \leq m$, then $\omega_i = O(1)$ and $\omega_j = O(p)$ according to assumptions (A1) and (A3), so that $\omega_i\omega_j/(\omega_i - \omega_j) = O(1)\cdot O(p)/\{O(1) - O(p)\} = O(1)$. Thus, $f_{1i}(\mathbf{\Omega}) = \sum_{j=1}^{m} \omega_j(1-\omega_j/\omega_i)^{-1} = m\cdot O(1) = O(1)$.
(Note: Here, $f_{1i}(\mathbf{\Omega})$ is always negative because $\omega_i - \omega_j < 0$.)
(C2.2) If $j \geq m+1$, then $\omega_i = O(1)$ and $\omega_j = O(1)$ according to (A3), so that ω_j/ω_i is $O(1)$. Thus, $(1-\omega_j/\omega_i)^{-1}$ is still $O(1)$ by (A2), and also, $\omega_j(1-\omega_j/\omega_i)^{-1} = O(1)\cdot O(1) = O(1)$. By collecting $p - m - 1$ such terms, along

with (A4), $f_{2i}(\mathbf{\Omega}) = \sum_{\substack{j=m+1 \\ j \neq i}}^{p} \omega_j (1-\omega_j/\omega_i)^{-1} = (p-m-1)\cdot O(1) = O(p)$.

Therefore, combining (C2.1) and (C2.2), $f_i(\mathbf{\Omega}) = f_{1i}(\mathbf{\Omega}) + f_{2i}(\mathbf{\Omega}) = O(1) + O(p) = O(p)$. In summary, if $i \geq m+1$, in Eq. (6), the first term $\omega_{\mathbf{i}}$ is $O(1)$, and the second term $(1/n) f_i(\mathbf{\Omega})$ is $O(p/N)$.

Thus, by combining Cases 1 and 2 the order of the bias terms as a whole is $O(p/N)$. Therefore, we can write Eq. (6) as:

$$E(l_i) = \omega_i + O(p/N), \tag{8}$$

where

$$\begin{aligned} \omega_i &= O(p) \quad i = 1, \ldots, m, \text{ and} \\ \omega_i &= O(1) \quad i = m+1, \ldots, p, \end{aligned}$$

by (A1) and (A3).

In summary, our analytical results imply that (i) when p is ignorable relative to N, the order of the bias of the eigenvalues of the sample covariance matrix is $1/N$; (ii) when p is not ignorable relative to N, the order of the bias of the sample eigenvalues is p/N; (iii) The sample eigenvalue is asymptotic unbiased if p/N goes to zero. It is obvious that the result in case (ii) is more general than that in case (i), and the sample eigenvalues are always asymptotically unbiased when $N \to \infty$ while holding p constant.

Here, it is important to note that problems might arise if $O(p/N)$ exceeds $o(1)$, in which case the assumption of $p/N \to 0$ in (A3) does not hold. This indicates that the estimated eigenvalues are consistent if and only if $p/N \to 0$ (See also, e.g., the consistency result in Theorem 1 of Johnstone and Lu 2009 on this point).

4 Simulation

4.1 Method

Our result indicates that the bias term in the eigenvalues of the sample covariance matrix should increase as the number of variables p increases, and that the bias should decrease as the sample size N increases. To verify the result, we conducted a small simulation. The number of variables p is chosen as either 20 or 100. For each p, the sample sizes N varies according to $2p$, $4p$, $6p$, $8p$, and $10p$. At p = 20, the population eigenvalues are chosen proportional to 10.0, 8.0, 6.0, 2.0, 1.9, 1.8, 1.7, 1.6, 1.5, 1.4, 1.3, 1.2, 1.1, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, and then these 20 numbers are scaled so that the sum of the eigenvalues to be exactly 20 (See Fig. 1 for the scree plot) to satisfy the need for a population correlation matrix. For p = 100,

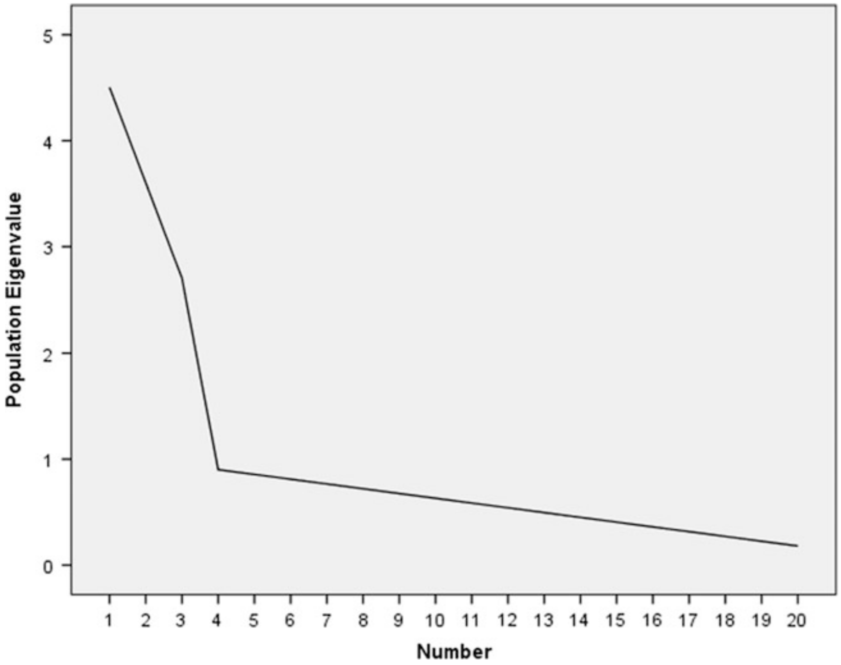


Fig. 1 Population eigenvalues of the condition with the number of variables $p = 20$

we generated population eigenvalues using the series $(p)(2k)^{-1}/\sum_{i=1}^{p}(2k)^{-1}$, $k = 1, \ldots, p = 100$ (See Fig. 6 for the scree plot).

At each p, we generated an orthogonal matrix with the pre-specified eigenvalues using the algorithm by Stewart (1980). Next, we applied the Givens rotations to the orthogonal matrix (Bendel and Mickey 1978). Finally, we converted the rotated matrix into a correlation matrix (Davies and Higham 2000). The entire procedure is implemented in the SAS Procedure IML (SAS Institute). Now, a random sample of size N were generated from the multivariate normal distribution with mean vector $\mathbf{0}$ and the correlation matrix created by the procedure described above. For each data set with p variables and N observations, we computed a sample correlation matrix, and obtained the p sample eigenvalues. The number of replications is 1,000 for each crossed condition of p and N. The 1,000 replications were averaged in obtaining the mean sample eigenvalues, which were compared against the corresponding population eigenvalues, and an empirical bias is thus obtained for each of the p sample eigenvalues.

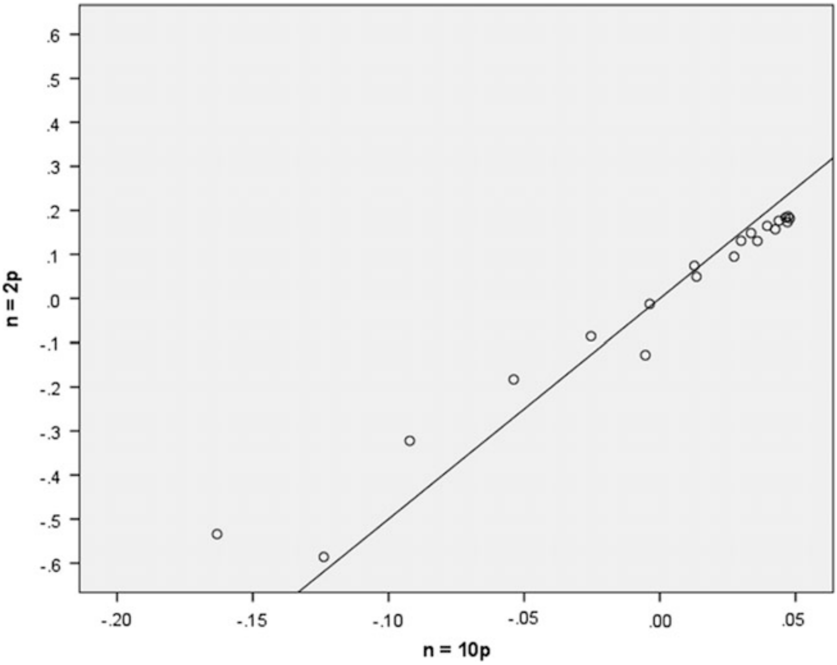


Fig. 2 Scatterplot of average biases between the case with $N = 2p = 40$ and the case with $N = 10p = 200$ for the number of variables $p = 20$. The inserted line is $y = 5x$

4.2 Results

Figures 2 through 5 describe the results with $p = 20$. The plot in Fig. 2 contrasts the empirical bias at $N = 40$ ($p/N = 0.5$) on the vertical axis against that at $N = 200$ ($p/N = 0.1$) on the horizontal axis. The slope of the solid line is equal to the ratio of two *sample sizes*, 200/40 = 5 (i.e., $y = 5x$) corresponding to the theoretical result obtained in the previous section. In Figs. 3, 4 and 5, the values on the horizontal axis remain the same whereas the values on the vertical axis are changed to those corresponding to $N = 80$, 120, and 160, respectively, and so are the slopes of the solid lines. In parallel, Figs. 7 through 10 contain the plots of the empirical bias for $p = 100$, where the horizontal axis is for the condition of $N = 1{,}000$ while the conditions for vertical axis vary from $N = 200$ to 800. Corresponding to the ratio of the sample sizes, the slopes of the solid line in the four figures are 5/1 = 5, 5/2 = 2.5, 5/3, and 5/4 = 1.25, respectively.

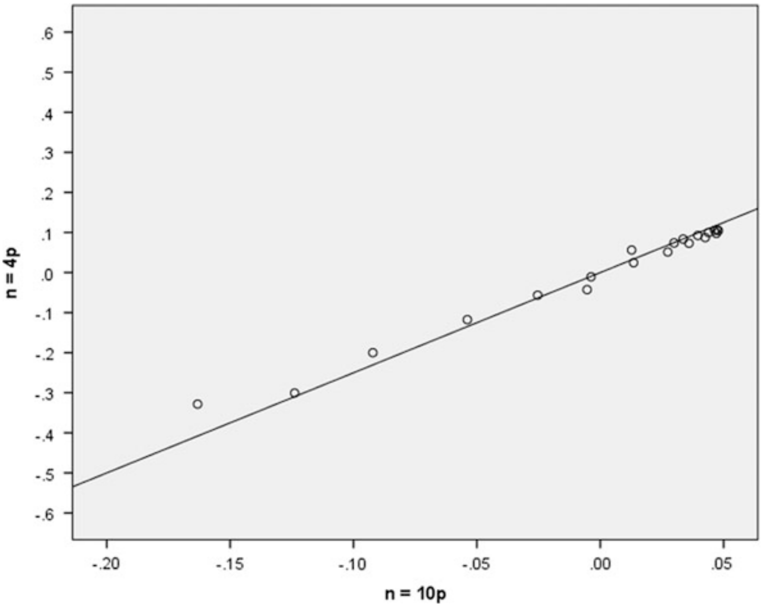


Fig. 3 Scatterplot of average biases between the case with $N = 4p = 80$ and the case with $N = 10p = 200$ for the number of variables $p = 20$. The inserted line is $y = 2.5x$

Note that, with the same value of p, the slope of the solid line in each figure is also identical to the ratio of the values of p/N. If the points are on the line, we have a support to our finding that the bias term is of order p/N.

For the scenario with $p = 20$, there are some fluctuations from the expected results in Fig. 2, due to the small sample size ($N = 40$) on the vertical axis. In Fig. 3, where the vertical axis is with a sample size of $N = 80$, the points are very close to the expected line. As the sample size increases, the points in both Figs. 4 and 5 are essentially on the solid line (Fig. 6).

The results at $p = 100$ are similar to those at $p = 20$. In Figs. 7 and 8 when the sample sizes (of $N = 200$ and 400, respectively) on the vertical axis are relatively small, the points somewhat deviate downward from the theoretical lines at the large end of the plots. However, in Figs. 9 and 10, where the sample sizes for the vertical axes are $N = 600$ and 800, respectively, the points are mostly on the theoretical solid lines except a few at the low end.

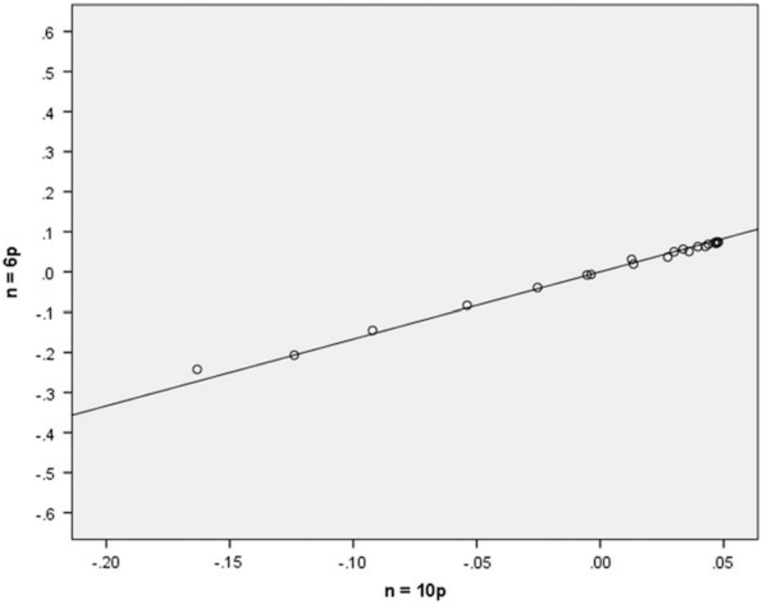


Fig. 4 Scatterplot of average biases between the case with $N = 6p = 120$ and the case with $N = 10p = 200$ for the number of variables $p = 20$. The inserted line is $y = (5/3)x$

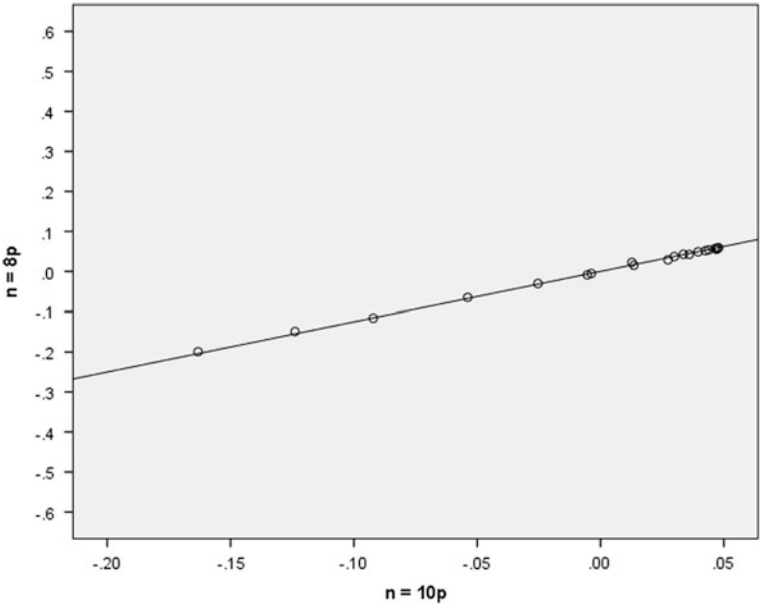


Fig. 5 Scatterplot of average biases between the case with $N = 8p = 160$ and the case with $N = 10p = 200$ for the number of variables $p = 20$. The inserted line is $y = 1.25x$

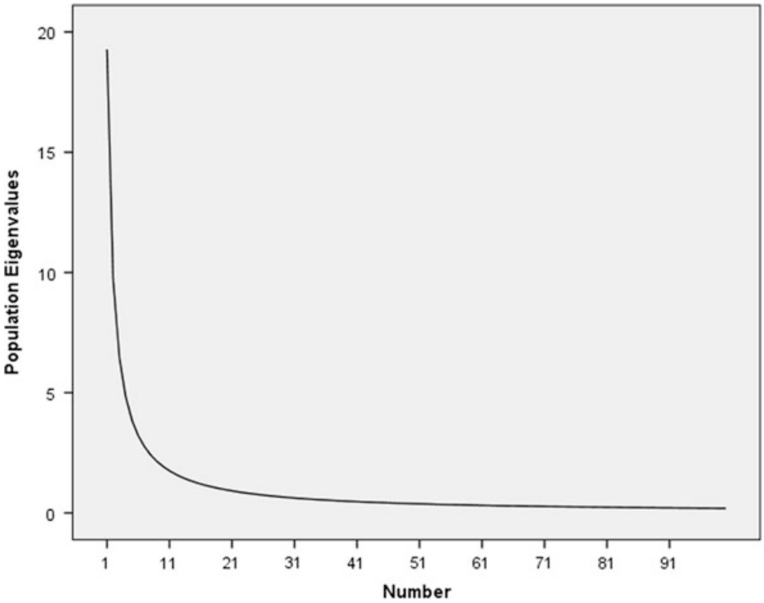


Fig. 6 Population eigenvalues of the condition with the number of variables $p = 100$

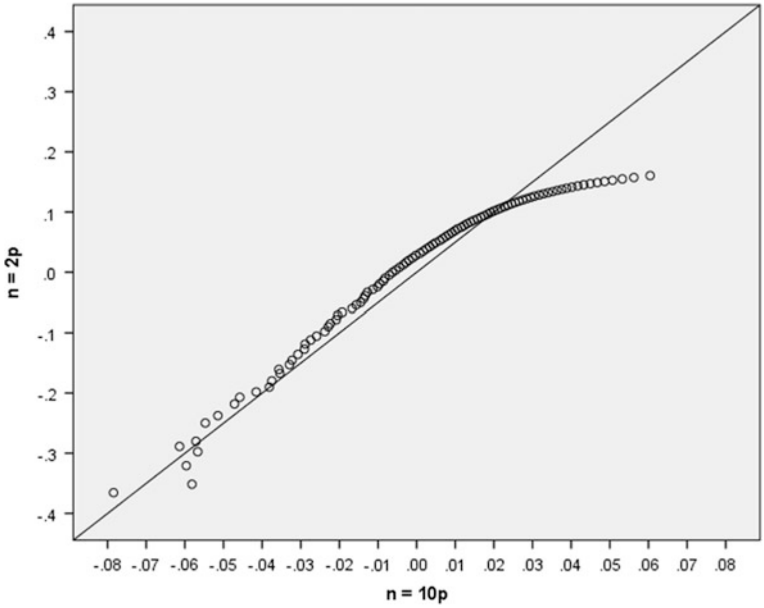


Fig. 7 Scatterplot of average biases between the case with $N = 2p = 200$ and the case with $N = 10p = 1{,}000$ for the number of variables $p = 100$. The inserted line is $y = 5x$

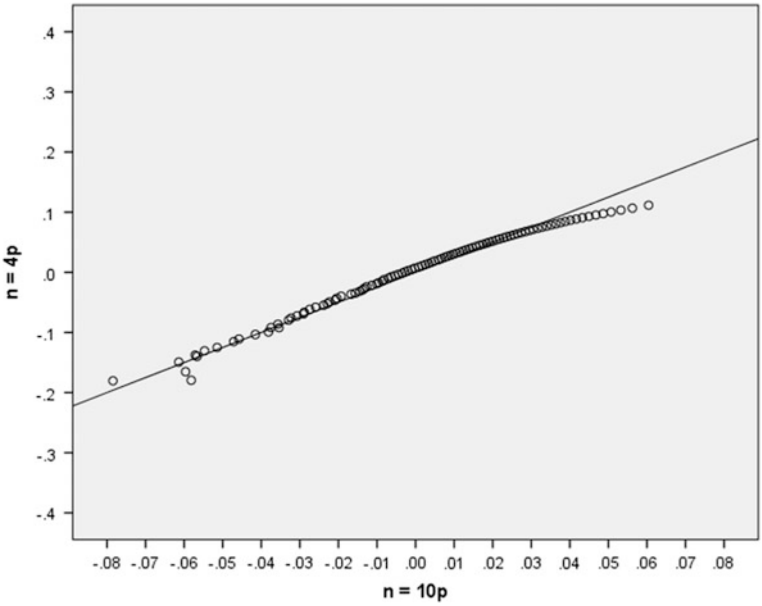


Fig. 8 Scatterplot of average biases between the case with $N = 4p = 400$ and the case with $N = 10p = 1{,}000$ for the number of variables $p = 100$. The inserted line is $y = 2.5x$

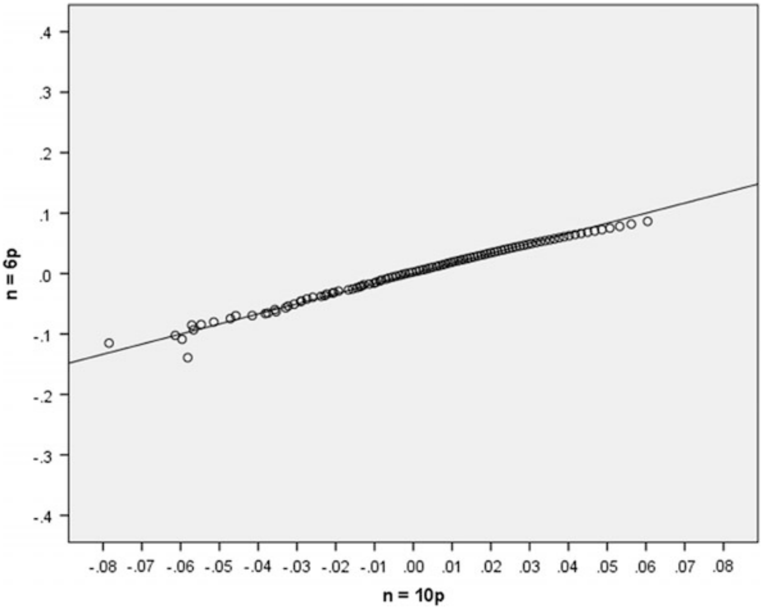


Fig. 9 Scatterplot of average biases between the case with $N = 6p = 600$ and the case with $N = 10p = 1{,}000$ for the number of variables $p = 100$. The inserted line is $y = (5/3)x$

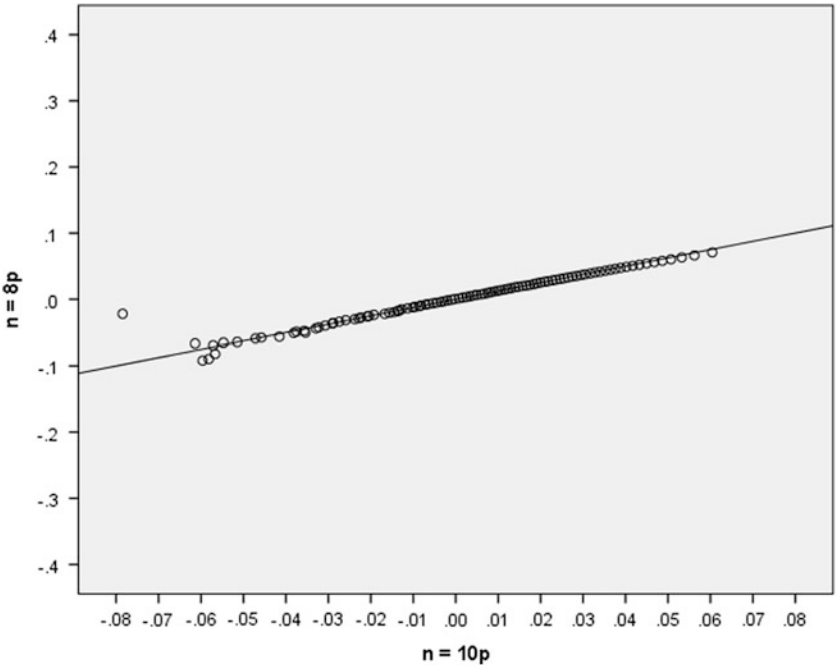


Fig. 10 Scatterplot of average biases between the case with $N = 8p = 800$ and the case with $N = 10p = 1{,}000$ for the number of variables $p = 100$. The inserted line is $y = 1.25x$

5 Concluding Remarks

We showed that the order of the bias of the eigenvalues of the sample covariance matrix is $O(p/N)$ and is not negligible when p is large. We confirmed our finding by a simulation study. We suspect that higher order terms such as $p^{1/2}N^{-3/4}$ and $p^{3/4}N^{-1/2}$ might explain the deviations seen in Figs. 7 and 8 (Yanagihara, personal communication). The bias term of order $O(p/N)$ may justify the use of p/N as the ridge tuning constant in the ridge methods for structural equation models (Yuan and Chan 2008, 2016).

Acknowledgements The authors are thankful to Dr. Dylan Molenaar for his very helpful comments. Ke-Hai Yuan's work was supported by the National Science Foundation under Grant No. SES-1461355.

References

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics, 34,* 122–148.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley.

Arruda, E. H., & Bentler, P. M. (2017). A regularized GLS for structural equation modeling. *Structural Equation Modeling, 24,* 657–665.

Bendel, R. B. & Mickey, M. R. (1978). Population correlation matrices for sampling experiments. *Communications in Statistics—Simulation and Computation*, *7*, 163–182.

Bentler, P. M., & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research, 25,* 67–74.

Davies, P. I., & Higham, N. J. (2000). Numerically stable generation of correlation matrices and their factors. *BIT, 40,* 640–651.

Guttman, L. (1956). "Best possible" estimates of communalities. *Psychometrika, 21,* 273–286.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417–441, 498–520.

Hwang, H., & Takane, Y. (2004). Generalized structured component analysis. *Psychometrika, 69,* 81–99.

IBM Corp. (2016). *IBM SPSS statistics for windows*, *Version 24.0*. Armonk, NY: IBM Corp.

Johnstone, I. M., & Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association, 104,* 682–703.

Krijnen, W. P. (2006). Convergence of estimates of unique variances in factor analysis, based on the inverse of sample covariance matrix. *Psychometrika, 71,* 193–199.

Lawley, D. N. (1956). Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika, 43,* 128–136.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. New York: Wiley.

SAS Institute. Appendix C: Generating random correlation matrices. Retrieved April 28, 2017, from https://support.sas.com/publishing/authors/extras/65378_Appendix_C_Generating_Random_Correlation_Matrices.pdf.

Schneeweiss, H., & Mathes, H. (1995). Factor analysis and principal components. *Journal of Multivariate Analysis, 55,* 105–124.

Stewart, G. W. (1980). The efficient generation of random orthogonal matrices with an application to condition estimation. *SIAM Journal on Numerical Analysis, 17,* 403–409.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25,* 1–28.

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 distribution: Ridge generalized least squares. *Structural Equation Modeling, 23,* 163–179.

Using Product Indicators in Restricted Factor Analysis Models to Detect Nonuniform Measurement Bias

Laura Kolbe and Terrence D. Jorgensen

Abstract When sample sizes are too small to support multiple-group models, an alternative method to evaluate measurement invariance is restricted factor analysis (RFA), which is statistically equivalent to the more common multiple-indicator multiple-cause (MIMIC) model. Although these methods traditionally were capable of detecting only uniform measurement bias, RFA can be extended with latent moderated structural equations (LMS) to assess nonuniform measurement bias. As LMS is implemented in limited structural equation modeling (SEM) computer programs (e.g., M*plus*), we propose the use of the product indicator (PI) method in RFA models, which is available in any SEM software. Using simulated data, we illustrate how to apply this method to test for measurement bias, and we compare the conclusions with those reached using LMS in M*plus*. Both methods obtain comparable results, indicating that the PI method is a viable alternative to LMS for researchers without access to SEM software featuring LMS.

Keywords Factor analysis · Product indicators · Measurement invariance
Nonuniform measurement bias

1 Introduction

Measurement bias entails that scales function differently across groups, irrespective of true differences in the construct that the scale was designed to measure. Let T denote the construct of interest measured by a set of observed variables X. Moreover, let V be a set of variables other than T. The formal definition of measurement bias involves a violation of measurement invariance (Mellenbergh 1989):

L. Kolbe (✉) · T. D. Jorgensen
Research Institute of Child Development and Education, University of Amsterdam,
1001 NG, P.O. Box 15776 Amsterdam, The Netherlands
e-mail: L.Kolbe@uva.nl

T. D. Jorgensen
e-mail: T.D.Jorgensen@uva.nl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_20

$$f_1(X|T=t, V=v) = f_2(X|T=t) \tag{1}$$

where f_1 is the conditional distribution of X given T and V, and f_2 the conditional distribution of X given T. If measurement invariance holds (i.e., $f_1 = f_2$), the measurement of T by X is invariant with respect to V. But if measurement invariance does not hold (i.e., $f_1 \neq f_2$), the measurement of T by X is biased with respect to V. A distinction can be made between uniform and nonuniform bias. Uniform bias implies that the extent of bias is constant for all levels of the construct T, whereas nonuniform bias implies that the extent of bias varies with T.

A common method to test for measurement bias with respect to a grouping variable is multiple-group confirmatory factor analysis (MGCFA; Vandenberg and Lance 2000), which requires sufficiently large samples for each group. An alternative for testing measurement bias is restricted factor analysis (RFA; Oort 1992, 1998). An advantage of this method over MGCFA is that the potential violator V may be categorical or continuous, observed or latent, and multiple violators can be investigated simultaneously. Moreover, RFA does not require the division of the sample into subsamples by V. The latter advantage comes at the cost of additional assumptions—namely, homogeneity of residual variances across groups.[1] If these additional assumptions hold, RFA should have more power than MGCFA to detect measurement bias.

When using RFA, the potential violator V is added to a common factor model as an exogenous variable that covaries with T. Uniform bias can be assessed by testing the significance of direct effects of V on X. To assess nonuniform bias, an extension for modeling latent interactions is required. RFA is commonly extended with latent moderated structural equations (LMS; Barendse et al. 2010). This allows for assessing nonuniform bias by testing the significance of interaction effects of $T \times V$ on X. Although this method generally has high power to detect measurement bias (Barendse et al. 2010, 2012; Woods and Grimm 2011), a disadvantage is that LMS is only implemented in the commercial structural equation modeling (SEM) software M*plus* (Muthén and Muthén 2012).[2] Moreover, most traditional SEM fit indices to test for model fit are not available when using the LMS method in M*plus*, except for Akaike's Information Criterion (AIC; Akaike 1973) and Bayesian Information Criterion (BIC; Schwartz 1978).

In this chapter, we introduce the product indicator (PI) method to model latent interactions in RFA models. The PI method has received a great deal of attention in the general context of modeling interactions among latent variables in SEM (Henseler and Chin 2010; Lin et al. 2010; Little et al. 2006; Marsh et al. 2004),

[1] In traditional RFA models, common-factor variances are also assumed to be equal across groups. However, when extending RFA to include a latent interaction factor with product indicators (described immediately following), differences in common-factor variances can be captured by the covariance between the common factor and the latent interaction factor.

[2] LMS is also available in the open-source R package `nlsem` (Umbach et al. 2017), but the implementation is very limited. It is not possible to test measurement bias using RFA models in the `nlsem` package, so we do not consider it further.

but has never been studied in light of testing measurement bias. First, we discuss the detection of measurement bias using RFA models, then we introduce the PI method, and finally we demonstrate how to test for measurement bias using RFA with PI by means of an illustrative example. We compare the results of PI to LMS on the same simulated data set.

2 Restricted Factor Analysis

2.1 Detection of Measurement Bias with RFA Models

In RFA models, the construct T can be modeled as a latent factor with multiple measures X as observed indicators. The possible violator V is added to the measurement model as an exogenous single-indicator latent variable and is allowed to covary with the common factor T. The violator V may represent a grouping variable by using a dummy-coded indicator. The observed scores X are modeled as

$$x_j = \boldsymbol{\tau} + \boldsymbol{\lambda} t_j + b g_j + c t_j g_j + \boldsymbol{\delta}\boldsymbol{\varepsilon}_j \tag{2}$$

where x_j is a vector of observed scores, t_j is the common factor T score, g_j is a dummy code for group membership V, and $\boldsymbol{\varepsilon}_j$ is a vector of the residual scores of subject j. Moreover, the vector $\boldsymbol{\tau}$ contains intercepts, $\boldsymbol{\lambda}$ is a vector of factor loadings on the common factor T, and $\boldsymbol{\delta}$ is a vector of residual factor loadings. The vectors b and c are of special interest and contain regression coefficients. A nonzero element in b or c indicates uniform or nonuniform bias, respectively.

Figure 1 illustrates an example of an RFA model to test for measurement bias using two anchor items. The violator V is modeled as a latent variable with a single indicator G representing group membership. For visual simplicity, the measurement model of $T \times V$ is excluded from Fig. 1, but those details are discussed in the following subsection. Measurement bias can be examined by comparing the fit of an unconstrained model with several constrained models. In the unconstrained model, all items are regressed on V and $T \times V$, except for the items in the anchor set. Each constrained model involves fixing the regression of the studied item onto V and $T \times V$ at zero.

The pair of constraints for each item can be tested simultaneously, where the null hypothesis of no measurement bias implies both $\boldsymbol{b}$ and $\boldsymbol{c}$ coefficients corresponding to the studied item are zero in the population. These constraints can be tested via model comparison of a constrained and unconstrained model, producing a likelihood ratio test statistic that is distributed as χ^2 random variable with 2 *df.* A significant test statistic indicates that the studied item is biased with respect to V, and 1-*df* follow-up tests of the individual $\boldsymbol{b}$ and $\boldsymbol{c}$ coefficients can reveal whether that indicator's bias is uniform or nonuniform. Our study focuses only on the 2-*df* omnibus test for each indicator.

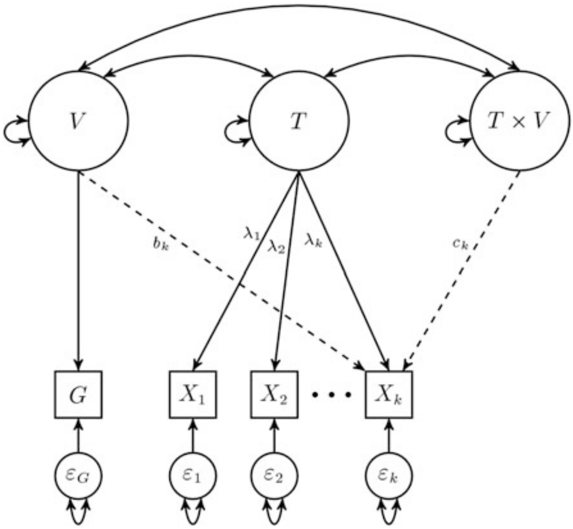


Fig. 1 An example of testing measurement bias using an RFA model. Dashed arrows represent effects that may be estimated to test for uniform and nonuniform bias. The indicators X_1 and X_2 serve as anchor items

2.2 Product Indicators

The use of PI to model interactions among latent variables was originated by Kenny and Judd (1984). The PI method involves the specification of a measurement model for the latent interaction factor. Generally, product terms are built by multiplying the indicators of the associated latent variables, which serve as indicators for the latent interaction factor. All indicators, including the product indicators, are assumed to be multivariate normally distributed if the maximum likelihood estimation procedure is used. Because products of normal variables are not themselves normally distributed, this assumption is violated. Thus, a robust maximum likelihood estimator is used to relax this assumption (see Marsh et al. 2004).

Several variants of the PI method have been proposed, among which is the double-mean-centering strategy (Lin et al. 2010) that we implement herein. The double-mean-centering strategy is superior to other strategies because it eliminates the need for a mean structure and does not involve a cumbersome estimation procedure. Although the orthogonalizing and double-mean-centering strategy perform equally well when all indicators are normally distributed, the double-mean-centering strategy performs better when the assumption that all indicators are normally distributed is violated (Lin et al. 2010).

The Double-Mean-Centering Strategy. The first step of the double-mean-centering strategy involves mean-centering the indicators of the latent variables of interest. Each of the mean-centered indicators of one latent variable are multiplied by the mean-centered indicators of the other latent variable. Then, the resulting product indicators are centered at their means and are used as indicators of the latent interaction factor. If the common factor T has I indicators and the violator variable V has J indicators, then the latent interaction factor can have up to $I \times J$ product indicators, although matching schemes have been proposed to reduce the number of product indicators (Marsh et al. 2004). In RFA, however, these matching schemes would be irrelevant when the common factor only interacts with a single-indicator violator construct (or with multiple single-indicator violators). Figure 2 shows an example of an RFA model with a latent interaction using the PI method. All possible cross-products are used in this example (i.e., each indicator of T is multiplied by the single indicator of V), and all indicators of T and V are centered at their means.[3]

3 Illustrative Example

We simulated a single data set to demonstrate how to apply the PI method in R (R Core Team 2016) to test for measurement bias, and to compare the conclusions with those reached using LMS. See Barendse et al. (2012) for M*plus* syntax to apply LMS.

3.1 Data Generation

Data were generated for two groups, each with a group size of $n = 100$. We considered a scale of $k = 10$ items, 40% of which were biased: two uniformly biased items and two nonuniformly biased items. This way, we are able to investigate the performance of LMS and PI using a hypothetical scale with a substantial degree of measurement bias. Item scores of subject j in group g were generated using the following model:

$$x_j = \boldsymbol{\tau}_g + \boldsymbol{\lambda}_g t_j + \boldsymbol{\delta}_g \boldsymbol{\varepsilon}_j \tag{3}$$

where x_j is a vector of 10 item scores, t_j is the common factor score, and $\boldsymbol{\varepsilon}_j$ is a vector of 10 unique factor scores (residuals) for subject j. Moreover, $\boldsymbol{\tau}_g$ is a vector containing 10 intercepts, $\boldsymbol{\lambda}_g$ is a vector of 10 common factor loadings, and $\boldsymbol{\delta}_g$ is a vector of 10 residual factor loadings of group g. Following Barendse et al. (2010), differences in the common factor were simulated by drawing common factor scores from a standard normal distribution for the reference group $t^r \sim N(0, 1)$ and from a normal distribution with a lower mean for the focal group $t^f \sim N(-0.5, 1)$. Residual factor scores were drawn from a standard normal distribution $\varepsilon_j \sim N(0, 1)$.

[3] In the case of a dummy-coded indicator, the mean is the proportion of the sample in Group 1. Mean-centering does not affect the variance, so a 1-unit increase in a mean-centered dummy code still represents a comparison of Group 1 to Group 0, just as the original dummy code does.

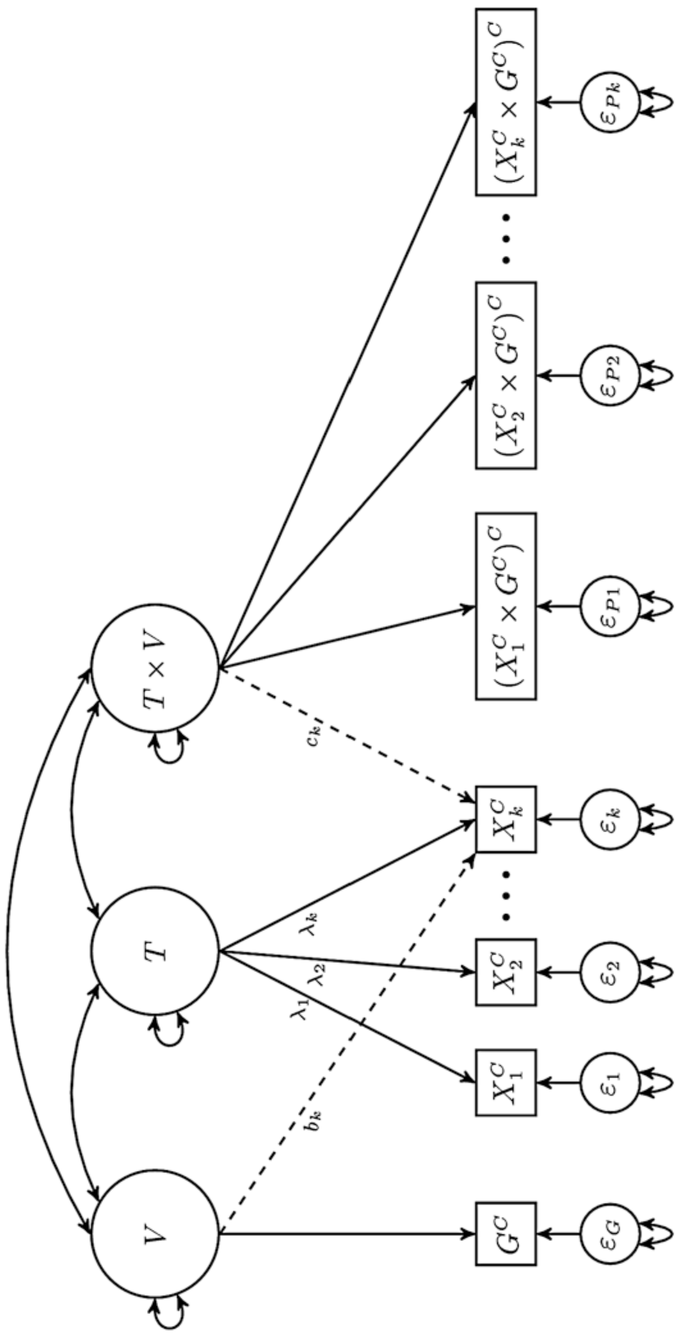


Fig. 2 An example of testing measurement bias using an RFA model with PI. Dashed arrows represent effects that may be estimated to test for uniform and nonuniform bias. The indicators X_1 and X_2 serve as anchor items

The same magnitude of uniform and nonuniform bias used by Barendse et al. (2010) was used. To introduce uniform bias, all intercepts $\boldsymbol{\tau}$ were equal to 0, except for the intercept for the second and third item in the focal group, which were chosen equal to 0.5 (small uniform bias) and 0.8 (large uniform bias), respectively. Moreover, all common factor loadings were fixed at 0.8, except for the factor loadings of the fourth and fifth item in the focal group, which were chosen equal to 0.55 (small nonuniform bias) and 0.3 (large nonuniform bias), respectively. The residual factor loadings were set equal to the square root of $1-\boldsymbol{\lambda}_g^2$. Table 1 presents R syntax to generate this data set.

Table 1 R syntax for data generation for the illustrative example

```
## set seed
RNGkind("L'Ecuyer-CMRG")
.Random.seed <- as.integer(c(407, 1945764513, -1852313839, 178524778,
-983224279,-1572978333, -68534343))
## specify group size
Nn <- 100
## draw latent-trait values
theta1 <- rnorm(Nn)
theta2 <- rnorm(Nn, -0.5, 1)
## draw scores on residual factor
residual <- matrix(NA, 2*Nn, 10)
for (j in 1:Nn) {
  for (i in 1:10) {
    residual[j, i] <- rnorm(1)
  }
}
## model parameters reference group
loading1 <- rep(0.8, 10)
delta1 <- sqrt(1 - loading1^2)
## model parameters focal group
tau2 <- c(0, -0.5, -0.8, 0, 0, 0, 0, 0, 0, 0)
loading2 <- c(0.8, 0.8, 0.8, 0.55, 0.3, 0.8, 0.8, 0.8, 0.8, 0.8)
delta2 <- sqrt(1 - loading2^2)
## simulate indicator scores reference group
x1 <- matrix(NA, Nn, 10)
for (j in 1:Nn) {
  for (i in 1:10) {
    x1[j,i] <- loading1[i] * theta1[j] + delta1[i] * residual[j, i]
  }
}
## simulate indicator scores focal group
x2 <- matrix(NA, Nn, 10)
for (j in 1:Nn) {
  for (i in 1:10) {
    x2[j,i] <- tau2[i] + loading2[i]*theta2[j] + delta2[i]*residual[j,i]
  }
}
## combine scores of both groups
dat <- as.data.frame(rbind(x1, x2))
dat$group <- rep(c(1, 2), each = Nn)
names(dat) <- paste0("x", 1:11)
```

3.2 Application

Table 2 shows R syntax for the application of PI in RFA models to detect measurement bias in the simulated data set. The RFA models with PI are fitted with the R package `lavaan` (version 0.5–23; Rosseel 2012). In our example, we apply the double-mean-centering strategy. First, the `indProd()` function in the `semTools` package (version 0.4–14; semTools Contributers 2016) with the argument `doubleMC = TRUE` is used to transform the data in order to be suitable for this strategy. This way, the indicators of the common factor *T* and violator *V* are mean-centered and indicators of the interaction factor $T \times V$ are built by multiplying the mean-centered indicator of *V* by each mean-centered indicator of *T*. The resulting product indicators are mean-centered again. After the data are prepared, one constrained model for each studied item must be specified. We use the ninth and tenth items, which are both bias-free, as anchor items, so they are not tested for measurement bias. Hence, the studied items are the first eight items, four of which are biased, which leads to eight constrained models in total. The unconstrained model is the same across items.

The first factor of the unconstrained model is the common factor *T* with 10 mean-centered observed variables X^C as indicators. The second factor is the violator *V* with a mean-centered single indicator G^C representing group membership. The residual variance of G^C is fixed at 0. The interaction factor $T \times V$ is the third factor

Table 2 R syntax for the application of PI in RFA in the illustrative example

```
## required package
library(semTools)
## prepare data
datDMC <- indProd(dat, 1:10, 11, match = FALSE, doubleMC = TRUE)
## additional parameters
paramc <- paste0("group + group.by.theta =~ x", 1:8)
## specify and fit unconstrained model
mod.un <- c('
  theta =~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10
  group =~ 1*x11
  group.by.theta =~ x1.x11 + x2.x11 + x3.x11 + x4.x11 + x5.x11 +
                    x6.x11 + x7.x11 + x8.x11 + x9.x11 + x10.x11
  x11 ~~ 0*x11', paramc)
mod.un.fit <- cfa(mod.un, data = datDMC, estimator = "MLM")
## specify and fit constrained models
out <- matrix(NA, nrow = 8, ncol = 2,
              dimnames = list(paste0("x", 1:8), c("X2", "p")))
for (i in 1:length(paramc)) {
  mod.con <- mod.un[-(i+1)] # remove b and c for the i-th studied item
  mod.con.fit <- cfa(mod.con, data = datDMC, estimator = "MLM")
  outfit <- lavTestLRT(mod.con.fit, mod.un.fit,
                       method = "satorra.bentler.2001")
  out[i,1:2] <- c(outfit[2,5], outfit[2,7])
}
## print results
out
```

of the unconstrained model with double-mean-centered product indicators. For example, the first indicator of the interaction factor is obtained by mean-centering $G^C \times X_1^C$. For all factors in the unconstrained model, the factor loading λ of the first indicator is fixed at unity for identification. Covariances between all three factors are freely estimated. Finally, factor loadings of all items on V and $T \times V$ are added, except for the anchor items. The constrained models are built by removing factor loadings of the studied item on V and $T \times V$ from the unconstrained model. The estimator to be used for the unconstrained and constrained models is set to "`MLM`", which involves maximum likelihood estimation with robust standard errors and a Satorra-Bentler scaled test statistic (Rosseel 2012).

To test each of the eight items for measurement bias, likelihood ratio test statistics are calculated using the `lavTestLRT()` function in the `lavaan` package (version 0.5–23; Rosseel, 2012). This involves comparing the fit of the unconstrained model with each constrained model. By setting the argument `method` = "`satorra.bentler.2001`", a scaled $\Delta\chi^2$ test statistic with 2 *df* is computed as described by Satorra and Bentler (2001). An item is flagged as biased with respect to violator V when the $\Delta\chi^2$ statistic is significant using a criterion of $\alpha = 0.05$.

3.3 Results of Measurement Bias Detection

Table 3 presents the results of measurement bias detection using RFA with LMS and PI. When the PI method was applied, the $\Delta\chi^2$ statistics of three out of four truly biased items were significant. The item with small nonuniform bias, Item 4, was not flagged as biased, which is consistent with previous Monte Carlo studies showing that power to detect uniform bias is greater than to detect nonuniform bias (Barendse et al. 2010, 2012). Moreover, none of the $\Delta\chi^2$ statistics of the bias-free

Table 3 Results of testing measurement bias using RFA models with PI and LMS

Item	PI		LMS	
	$\chi^2_{df=2}$	p	$\chi^2_{df=2}$	p
1	0.425	0.809	0.674	0.714
2	**19.396**	**0.000**	**17.696**	**0.000**
3	**38.755**	**0.000**	**28.000**	**0.000**
4	5.217	0.074	**6.283**	**0.043**
5	**10.105**	**0.006**	**10.656**	**0.005**
6	0.145	0.930	0.201	0.904
7	0.948	0.622	0.772	0.680
8	0.246	0.884	0.196	0.907

Note Bold cells indicate significant measurement bias. Items 9 and 10 were used as anchor items, so they were not tested for measurement bias items were significant. Thus, none of the items were incorrectly flagged as biased using PI. The LMS method obtained comparable results, but correctly flagged all truly biased items as biased with respect to violator V.

4 Discussion

In this chapter, we proposed the use of PI in RFA models as an alternative to LMS to test nonuniform measurement bias. The illustrative example showed that this method obtains results comparable to LMS. Because RFA with LMS can only be implemented in M*plus* (Muthén and Muthén 2012), knowing that PI performs at least as well as LMS provides more researchers the opportunity to test for nonuniform bias using SEM software package. An additional advantage of PI is the availability of more traditional SEM fit indices to test for model fit that are not available when using LMS in M*plus*, nor when using other available strategies for modeling interactions with latent variables (e.g., random effects models which treat item responses as cross-nested within items and subjects). However, several aspects of the use of PI in RFA models are yet unclear, for example, which items should serve as product indicators for the interaction factor (e.g., all items, only anchor items, or anchor items and studied items). In addition, RFA models assume strict invariance, that is, equal residual variances across groups. Future research could investigate how violations of strict invariance affect Type I error rates.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademia Kiado. https://doi.org/10.1007/978-1-4612-1694-0_15.

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias: A simulation study. *Advances in Statistical Analysis, 94,* 117–127. https://doi.org/10.1007/s10182-010-0126-1.

Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling, 19,* 561–579. https://doi.org/10.1080/10705511.2012.713261.

Henseler, J., & Chin, W. W. (2010). A comparison of approaches for the analysis of interaction effects between latent variables using partial least squares path modeling. *Structural Equation Modeling, 17,* 82–109. https://doi.org/10.1080/10705510903439003.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96,* 201–210. https://doi.org/10.1037/0033-2909.96.1.201.

Lin, G.-C., Wen, Z., Marsh, H. W., & Lin, H.-S. (2010). Structural equation models of latent interactions: Clarification of orthogonalizing and double-mean-centering strategies. *Structural Equation Modeling, 17,* 374–391. https://doi.org/10.1080/10705511.2010.488999.

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling, 13,* 497–519. https://doi.org/10.1207/s15328007sem1304_1.

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9,* 275–300. https://doi.org/10.1037/1082-989X.9.3.275.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143. https://doi.org/10.1016/0883-0355(89)90002-5.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6,* 150–160. https://doi.org/10.1007/s10182-010-0126-1.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5,* 107–124. https://doi.org/10.1080/10705519809540095.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from the comprehensive R archive network (CRAN). https://www.R-project.org/.

Rosseel, Y. (2012). `lavaan`: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514. https://doi.org/10.1007/BF02296192.

Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6,* 461–464. https://doi.org/10.1214/aos/1176344136.

semTools Contributers. (2016). `semTools`: Useful tools for structural equation modeling [Compute software]. Retrieved from https://CRAN.R-project.org/package=semTools.

Umbach, N., Naumann, K., Brandt, H., & Kelava, A. (2017). Fitting nonlinear structural equation models in R with package `nlsem`. *Journal of Statistical Software, 77*(7), 1–20. https://doi.org/10.18637/jss.v077.i07.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70. https://doi.org/10.1177/109442810031002.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35,* 339–361. https://doi.org/10.1177/0146621611405984.

Polychoric Correlations for Ordered Categories Using the EM Algorithm

Kenpei Shiina, Takashi Ueda and Saori Kubo

Abstract A new method for the estimation of polychoric correlations is proposed in this paper, which uses the Expectation-Maximization (EM) algorithm and the Conditional Covariance Formula. Simulation results show that this method attains the same level of accuracy as other methods, and is robust to deteriorated data quality.

Keywords Polychoric correlation · EM algorithm · Conditional covariance formula

1 Correlation Coefficient Computed from Categorical Variables

Despite long-standing warnings by psychometricians, it is still common to use ordered categories (e.g., Likert ratings or verbal labels) as if they were integers. Pearson (1913), the inventor of *r*, did notice that when the number of categories are small, and thus categories are "broad," *r* is biased. This problem has been studied in sociology and psychology. Martin (1978) simulated the broad category problem and concluded: "The findings suggest that the amount of lost information is substantial." Bollen and Barb (1981) performed a similar simulation and arrived at a similar conclusion. Further, they noticed "more complex patterns occurring when the collapsed variables do not have the same number of categories" (1983, p. 286).

K. Shiina (✉) · T. Ueda
Waseda University, Tokyo, Japan
e-mail: shiina@waseda.jp

T. Ueda
e-mail: uedaman@gmail.com

S. Kubo
Tokyo Women's Medical University, Tokyo, Japan
e-mail: kubo.saori@twmu.ac.jp

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_21

Therefore, the use of the polychoric correlation coefficient (Olsson 1979) is generally recommended. The polychoric correlation coefficient was first introduced by Ritchie-Scott (1918) and Pearson and Pearson (1922) in the early 20th century, but it took over half a century before the computationally feasible maximum likelihood procedure was proposed by Olsson (1979).

In this paper, a new computational procedure for polychoric correlation is proposed, based on the Conditional Covariance Formula (Ross 2010) and the Expectation-Maximization (EM) algorithm (Dempster et al. 1977). Its accuracy and robustness are compared to those of other methods.

2 Assumptions on the Data Generating Process

Let x and y be two original, continuous latent variables, with their density function given by a Bivariate Normal Distribution (BND):

$$\phi(x,y\big|\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2}\{\frac{(x-\mu_x)^2}{\sigma_x^2(1-\rho^2)}-\frac{2(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y(1-\rho^2)}\rho+\frac{(y-\mu_y)^2}{\sigma_y^2(1-\rho^2)}\}}. \qquad (1)$$

The original variables are often assumed to be abstract latent variables, especially in psychology. It is further assumed that categorizing the original variables yields manifest variables X and Y, which are integer-valued. In Likert type ratings, for example, it can be postulated that a rater internally categorizes the original variables. In an educational setting, a teacher may categorize original test scores into integer ranks. There are many other empirical settings.

We should consider the number of categories (p for X and q for Y), as well as the arrangement of category boundaries, because the manner by which we partition the original, continuous latent variables into categories will be critical. We can set $\mu_x=\mu_y=0$ and $\sigma_x=\sigma_y=1$ without loss of generality and we can define or assume category boundaries for x and y as:

$$-\infty=\theta_0<\theta_1<\theta_2<\cdots<\theta_p=\infty$$
$$-\infty=\tau_0<\tau_1<\tau_2<\cdots<\tau_q=\infty$$

such that, if $\theta_{i-1}<x<\theta_i$, then $X=i$; if $\tau_{j-1}<y<\tau_j$, then $Y=j$.

The true probability γ_{ij} of each cell in the contingency table (correlation table) corresponds to the rectangular region $[\theta_{i-1},\theta_i]\times[\tau_{j-1},\tau_j]$ in the $x-y$ space (Fig. 1) and is given by:

$$\gamma_{ij}=\int\limits_{\theta_{i-1}}^{\theta_i}\int\limits_{\tau_{j-1}}^{\tau_j}\phi(x,y\big|\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)dydx. \qquad (2)$$

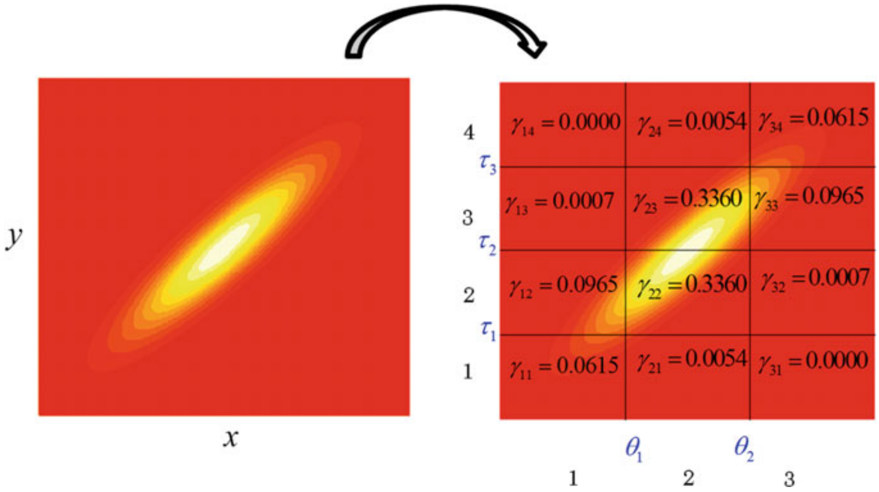


Fig. 1 Left: Heat map of the original distribution $\phi(X, Y|0, 0, 1, 1, 0.9)$. Right: Categorized space with its probability γ_{ij} defined by Eq. (2). $p = 3$, $q = 4$, $\theta_1 = -1, \theta_2 = 1, \tau_1 = -1.5, \tau_2 = 0$, $\tau_3 = -1.5$

Figure 1 shows how the contingency table is generated; let n_{ij} be the cell count of the empirical contingency table $\mathbf{N}$ and let $n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$, then, we can assume $\gamma_{ij} \approx n_{ij}/n$. An example of a contingency table, from which we try to restore the original ρ, is provided in Table 1.

From the rectangle regions in Fig. 1 we can construct a system of Quadruply Truncated Binormal Distributions (QTBDs), $g_{ij}(x, y)$, where:

$$g_{ij}(x, y) \equiv \phi(x, y)/\gamma_{ij}, \quad \theta_{i-1} < x < \theta_i, \quad \tau_{j-1} < y < \tau_j. \tag{3}$$

Notice that a QTBD is defined on a rectangular sub-space $[\theta_{i-1}, \theta_i] \times [\tau_{j-1}, \tau_j]$. Moments of QTBD (Genz and Bretz 2009), as numerically shown in Fig. 2, can easily be computed analytically (Muthén 1990) or by using *R*'s *tmvtnorm* package (Stefan and Manjunath 2015; *R* Core Team 2016).

Table 1 An example of the empirical correlation table (generated from Fig. 1: Right)

		X		
		1	2	3
Y	4	0	53	614
	3	7	3359	964
	2	964	3359	7
	1	614	53	0

Note In ordinary data analysis, we start from this table and try to restore the correlation coefficient

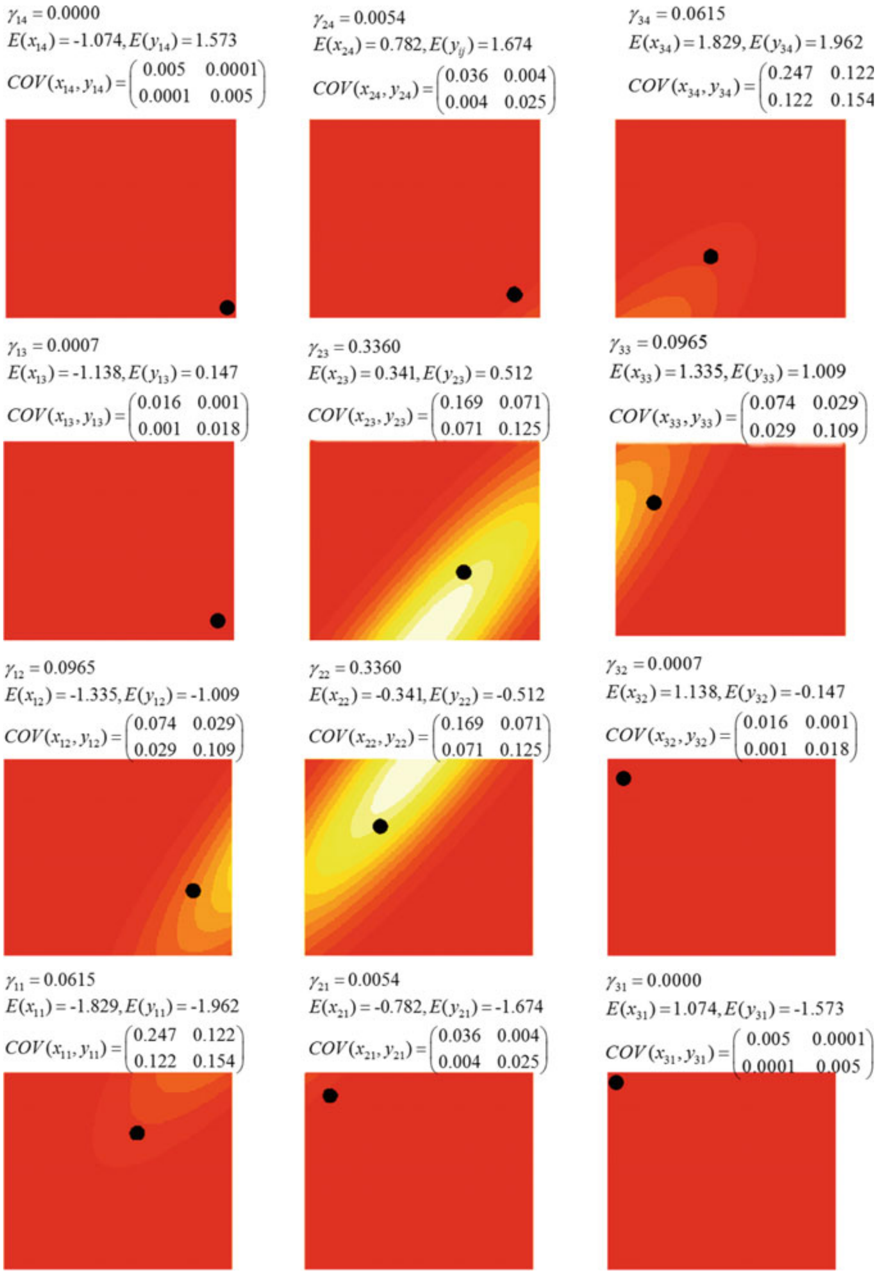


Fig. 2 Estimated moments of 12 QTBDs from Table 1 by using *R* package of *tmvtnorm.* The black blobs designate the mean of each QTBD on the rectangular region

3 Reproducing Original BND Moments from QTBD Using Conditional Covariance Formula

Let U, V, and Z be random variables. The Conditional Covariance Formula (Ross 2010, p. 381) is given by:

$$\begin{aligned} COV(U,V) &= E[COV(U,V|Z)] + COV(E(U|Z), E(V|Z)) \\ &= E[COV(U,V|Z) + E(U|Z)E(V|Z)] - E(U)E(V). \end{aligned} \tag{4}$$

In the present context, under the assumption of BND, $\phi(x, y|0, 0, 1, 1, \rho)$, this formula simplifies to:

$$\rho = \sigma_{xy} = \sum_{i=1}^{p} \sum_{j=1}^{q} \gamma_{ij}(\sigma_{ij} + \bar{x}_{ij}\bar{y}_{ij}) - \underbrace{\cancel{E(x)E(y)}}_{From\ assumption}, \tag{5}$$

where $\bar{x}_{ij}, \bar{y}_{ij}$, and σ_{ij} are the respective means and covariance of the QTBD, $g_{ij}(x, y)$. As mentioned previously, the means and covariance can be computed easily.

Equation (5) indicates that the covariance of the whole can be recovered by aggregating the parts. For example, from Fig. 2 it follows that:

$$\begin{aligned} \sum_{i=1}^{p} \sum_{j=1}^{q} \gamma_{ij}(\sigma_{ij} + \bar{x}_{ij}\bar{y}_{ij}) &= 0.0615 \times \{0.122 + (-1.829)(-1.962)\} \\ &\quad + 0.0965 \times \{0.029 + (-1.335)(-1.009)\} + \cdots + 0.0965 \times \{0.029 + (1.335)(1.009)\} \\ &\quad + 0.0615 \times \{0.122 + (1.829)(1.962)\} = 0.90 \end{aligned}$$

which shows a perfect restoration of the original $\rho = 0.90$ in Fig. 1.

If we replace γ_{ij} in Eq. (5) by its empirical counterpart, n_{ij}/n, we have the estimation formula:

$$\hat{\rho} = \hat{\sigma}_{xy} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n_{ij}}{n}(\sigma_{ij} + \bar{x}_{ij}\bar{y}_{ij}) \tag{6}$$

which is valid at least when $\gamma_{ij} \approx n_{ij}/n$.

4 An Iterative Procedure to Estimate ρ

Because $\sigma_{ij}, \bar{x}_{ij}$, and $\bar{y}_{ij}$ in Eq. (6) are functions of ρ, an iterative procedure seems possible with an updating formula:

$$\sigma_{xy}^{(t+1)} \leftarrow \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n_{ij}}{n}\left(\sigma_{ij}^{(t)} + \bar{x}_{ij}^{(t)}\bar{y}_{ij}^{(t)}\right)$$

where the parenthesized t is an index for iteration. The estimation procedure is as follows:

Step 1: Estimate the thresholds. The marginal of the contingency table can be used to estimate the thresholds (Olsson 1979). More precisely:

$$\hat{\theta}_k \approx \Phi^{-1}(\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{q} n_{ij}), \quad \hat{\tau}_k \approx \Phi^{-1}(\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{p} n_{ij})$$

can be used. The method that uses this estimation is often called the two-step procedure (Olsson 1979).

Step 2: Set the Initial value for ρ. The choice of initial value $\rho^{(0)}$ is quite arbitrary. This is based on our own simulation, which is presented in the Numerical Example section.

Step 3: Compute $\sigma_{ij}^{(t)}, \bar{x}_{ij}^{(t)}$ and $\bar{y}_{ij}^{(t)}$. We used *R* package "*tmvtnorm*" to compute these values, as shown in Fig. 2.

Step 4: Compute $\sigma_{xy}^{(t+1)}$ or $\rho_{xy}^{(t+1)}$ [if we assume $\phi(x, y|0, 0, 1, 1, \rho)$]. Use the values computed in Step 3 along with the recursive formula mentioned above.

Step 5: If there is no convergence, go back to Step 3.

5 EM Algorithm

The above computational procedure yielded satisfactory results in our simulation, as shown in the Numerical Example section. The derivation of Eq. (6) is, however, rather heuristic, and remains at a level of descriptive statistics. In this section, we show that Eq. (6) can be derived from a stationary equation within the framework of the EM Algorithm (Dempster et al. 1977).

The EM algorithm comprises the Expectation step (E-step), where the expectation of the likelihood is calculated by taking the missing variables into account, and the Maximization step (M-step), where the parameters are estimated by maximizing the likelihood function found in the E-step. The parameters found in the M-step are then used as the starting point of a new E-step phase, and the process is iterated until convergence.

As shown in the Appendix, the conditional expected log likelihood Q of the current problem can be written as:

$$Q = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} \iint \log \phi(x_{ij}, y_{ij}|\rho) \times g_{ij}\left(x_{ij}, y_{ij}|\rho^{(t)}\right) dx_{ij} dy_{ij}$$

where $(x_{ij}, y_{ij}) \in [\theta_{i-1}, \theta_i] \times [\tau_{j-1}, \tau_j]$. After some operations (see the Appendix), we have a stationary equation that needs to be solved for ρ:

$$\begin{aligned} &\rho(1-\rho^2) + (\rho^2+1)\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}\left[COV(x_{ij}, y_{ij}) + E(x_{ij})E(y_{ij})\right] \\ &-\rho\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}\left(V(x_{ij}) + V(y_{ij}) + E(x_{ij})^2 + E(y_{ij})^2\right) = 0. \end{aligned} \tag{7}$$

This is a cubic equation in variable ρ, and thus generally difficult to solve. Fortunately, with the help of Eq. (6) and the Conditional Covariance Formula, we can find an immediate solution to the stationary equation:

$$\rho = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}\left[COV\left(x_{ij}, y_{ij}\right) + E\left(x_{ij}\right)E\left(y_{ij}\right)\right] \tag{8}$$

with

$$2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}\left(V(x_{ij}) + V(y_{ij}) + E(x_{ij})^2 + E(y_{ij})^2\right) \approx V(x) + V(y). \tag{9}$$

This is because the insertion of Eqs. (8) and (9) into Eq. (7) yields the following result: $\rho(1-\rho^2) + (\rho^2+1)\rho - 2\rho = 0$. Note that under the assumption that the BND is $\phi(x, y|0, 0, 1, 1, \rho)$, Eq. (9) is naturally satisfied by using Eq. (4). Therefore, we have proved that the iterative procedure could be interpreted as a type of EM algorithm.

6 Numerical Example

A total of 10,000 contingency tables (p = 2, q = 3) with 1024 entries were generated, as shown in Table 2. The original BND was $\phi(x, y|0, 0, 1, 1, 0.9)$ and the thresholds were $\theta_1 = 0, \tau_1 = -1$, and $\tau_2 = -1$.

We compared four types of correlation coefficients $\hat{\rho}$ from the following:

1. Original continuous variables,
2. Integer-valued (Likert) variables,

Table 2 An example of a 2 by 3 correlation table

		X	
		1	2
Y	3	4	181
	2	353	324
	1	162	0

3. Olsson's polychoric correlation method (with two-step estimation), and
4. The new method proposed in the current study.

The results of the numerical simulation are depicted in Fig. 3 and Table 3. Not surprisingly, the original continuous variables fared the best amongst the other methods, and the estimation from integer-valued variables (Likert rating) was very poor. A comparison between the approach by Olsson (1979) and the present method shows that both methods demonstrated a similar level of accuracy. However, Olsson's polychoric correlation yielded a large number of overestimations.

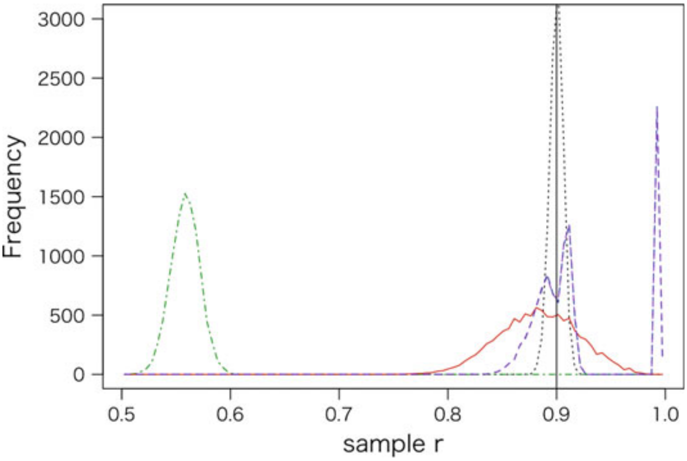


Fig. 3 The results of numerical simulation. Left most distribution is from Likert-type categories. Dotted line around 0.9 is from the original continuous variable. Solid line designates the results obtained by using the new method. Results from Olsson's polychoric correlation are shown in long dash: the distribution shows three peaks

Table 3 Summary statistics of the simulation

	Original continuous variable	Categorized (Likert) variable	Polycholic (two-step method)	New method
N	10000	10000	10000	10000
Mean	0.8998	0.5584	0.9180	0.8832
RMS	0.8999	0.5585	0.9191	0.8839
SD	0.0059	0.0133	0.0439	0.0367
RMS error	0.0047	0.3416	0.0323	0.0323

Note RMS = Root Mean Square, RMS error = Root Mean Square from true value ($\rho = 0.90$)

7 Discussion

Apparently, the new method shows satisfactory robustness in our simulation. Of course, in order to further validate our new procedure, a more substantive numerical check is needed; we are currently working on this. A major concern we have is that this method is rather slow, as is often the case with EM algorithms. On the other hand, a clear merit of this approach is its simplicity and theoretical clarity. We can observe that Eq. (6) enjoys a direct connection to the ordinary definition of Pearson's correlation coefficient: $r=\frac{1}{n}\sum_{k=1}^{n}x_k y_k$. The reason for this is that, if $p\to\infty, q\to\infty$ in such a way that $\max_{i=1,p}[\theta_i-\theta_{i-1}]\to 0$ and $\max_{j=1,q}[\tau_j-\tau_{j-1}]\to 0$, then the cells become very small; consequently $n_{ij}/n\to 1/n, \sigma_{ij}\to 0, \bar{x}_{ij}\approx x'_{ij}$, and $\bar{y}_{ij}\approx y'_{ij}$ where $(x'_{ij}, y'_{ij})\in[\theta_{i-1},\theta_i]\times[\tau_{j-1},\tau_j]$ is any vector within the rectangle. Therefore in Eq. (6), we have, as a rough approximation:

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}(\sigma_{ij}+\bar{x}_{ij}\bar{y}_{ij})\to\sum_{\substack{i=1\\n_{ij}\neq 0}}^{p}\sum_{j=1}^{q}\frac{1}{n}x'_{ij}y'_{ij},$$

which means that, when categories are very fine, we have a scatter plot and the ordinary formula for the correlation coefficient.

Acknowledgements This work was supported by JSPS KAKENHI Grant Numbers 23530871 and 16H02050. Correspondence concerning this article should be addressed to Kenpei Shiina, Department of Educational Psychology, School of Education, Waseda University, Shinjuku-ku, Tokyo, 169-8050, JAPAN.

Appendix

This Appendix shows the derivation of the stationary Eq. (7) within the framework of the EM algorithm.

Let the observed integer-valued data pair (e.g., pair of Likert ratings) be:

$$(X_k, Y_k), k=1,2,3,\ldots,n\ ,\ \text{where}\ X_k\in\{1,2,\ldots,p\}, Y_k\in\{1,2,\ldots,q\}.$$

from which we can construct a correlation table. According to the data generation process in the main text, n sample pairs:

$$(x_k, y_k),\quad k=1,2,3,\ldots,n$$

are first extracted from the standard BND and then categorized into $p\times q$ rectangular regions, which are defined by the thresholds.

The Likelihood of Complete Data

The complete data is the concatenation of the missing data (x_k, y_k) and the incomplete data (X_k, Y_k). The likelihood L of complete data is written as:

$$L = f(x_1, y_1, X_1, Y_1, x_2, y_2, X_2, Y_2, \ldots, x_n, y_n, X_n, Y_n|\rho).$$

The likelihood of data pair k is written as:

$$L_k = f_k(x_k, y_k, X_k, Y_k|\rho).$$

Assuming the independence of data pairs, we have:

$$L = f(x_1, y_1, X_1, Y_1, x_2, y_2, X_2, Y_2, \ldots, x_n, y_n, X_n, Y_n|\rho) = \prod_{k=1}^{n} f_k(x_k, y_k, X_k, Y_k|\rho)$$
$$= \prod_{k=1}^{n} f_k(X_k, Y_k|x_k, y_k, \rho)\phi(x_k, y_k|\rho).$$

Notice that:

$$f_k(X_k, Y_k|x_k, y_k, \rho) = \begin{cases} 1 & if\,(x_k, y_k) \in [\theta_{X_k-1}, \theta_{X_k}] \times [\tau_{Y_k-1}, \tau_{Y_k}] \\ 0 & otherwise \end{cases}$$

Therefore, we have:

$$L = f(x_1, y_1, X_1, Y_1, x_2, y_2, X_2, Y_2, \ldots, x_n, y_n, X_n, Y_n|\rho) = \prod_{k=1}^{n} \phi^*(x_k, y_k|\rho)$$

where ϕ^* is a BND under the restriction: $(x_k, y_k) \in [\theta_{X_k-1}, \theta_{X_k}] \times [\tau_{Y_k-1}, \tau_{Y_k}]$.

The Incomplete Data

Using QTBD, the probability density of missing data, given a data pair (X_k, Y_k) is:

$$f_k(x_k, y_k|X_k, Y_k, \rho) = \phi^*(x_k, y_k|\rho)/\gamma_{X_k Y_k} = g_{X_k Y_k}(x_k, y_k|\rho)$$

and thus, the probability density of incomplete data is given as:

$$\prod_{k=1}^{n} f_k(x_k, y_k|X_k, Y_k, \rho) = \prod_{k=1}^{n} \phi^*(x_k, y_k|\rho)/\gamma_{X_k Y_k} = \prod_{k=1}^{n} g_{X_k Y_k}(x_k, y_k|\rho)$$

where $(x_k, y_k) \in [\theta_{X_k-1}, \theta_{X_k}] \times [\tau_{Y_k-1}, \tau_{Y_k}]$.

Expected Log Likelihood **(E-step)**

Using the above results, the conditional expected log likelihood Q can be written as:

$$
\begin{aligned}
Q &= \int_{\theta_{X_1-1}}^{\theta_{X_1}}\int_{\tau_{Y_1-1}}^{\tau_{Y_1}}\int_{\theta_{X_2-1}}^{\theta_{X_2}}\int_{\tau_{Y_2-1}}^{\tau_{Y_2}}\cdots\int_{\theta_{X_n-1}}^{\theta_{X_n}}\int_{\tau_{Y_n-1}}^{\tau_{Y_2}} \{\log\prod_{k=1}^{n}\phi^{*}(x_k,y_k|\rho)\prod_{k=1}^{n} g_{X_kY_k}(x_k,y_k\big|\rho^{(t)})d\mathbf{x}d\mathbf{y} \\
&= \int_{\theta_{X_1-1}}^{\theta_{X_1}}\int_{\tau_{Y_1-1}}^{\tau_{Y_1}}\int_{\theta_{X_2-1}}^{\theta_{X_2}}\int_{\tau_{Y_2-1}}^{\tau_{Y_2}}\cdots\int_{\theta_{X_n-1}}^{\theta_{X_n}}\int_{\tau_{Y_n-1}}^{\tau_{Y_n}} \{\sum_{k=1}^{n}\log\phi^{*}(x_k,y_k|\rho)\prod_{k=1}^{n} g_{X_kY_k}(x_k,y_k\big|\rho^{(t)})d\mathbf{x}d\mathbf{y} \\
&= \sum_{k=1}^{n} Q_k
\end{aligned}
$$

where $Q_k=\int_{\theta_{X_k-1}}^{\theta_{X_k}}\int_{\tau_{Y_k-1}}^{\tau_{Y_k}}\log\phi^{*}(x_k,y_k|\rho)\times g_{X_kY_k}(x_k,y_k\big|\rho^{(t)})dx_kdy_k$.

This simplification is possible because any variables other than (x_k,y_k) are integrated out.

There are some Q_k's that are defined in the same region, $[\theta_{i-1},\theta_i]\times[\tau_{j-1},\tau_j]$, and thus have the same value. We can categorize these Q_k's and arrive at the final expression:

$$
Q=\sum_{k=1}^{n} Q_k=\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}Q_{ij}
$$

where $Q_{ij}=\int_{\theta_{i-1}}^{\theta_i}\int_{\tau_{j-1}}^{\tau_j}\log\phi^{*}(x_{ij},y_{ij}|\rho)\times g_{ij}(x_{ij},y_{ij}\big|\rho^{(t)})dx_{ij}dy_{ij}$ with the understanding that (x_{ij},y_{ij}) are variables confined to $[\theta_{i-1},\theta_i]\times[\tau_{j-1},\tau_j]$.

The Derivation of the Stationary Equation **(M-step)**

Because

$$
\begin{aligned}
Q_{ij} &= \int_{\theta_{i-1}}^{\theta_i}\int_{\tau_{j-1}}^{\tau_j}\log\phi^{*}(x_{ij},y_{ij}|\rho)\times g_{ij}(x_{ij},y_{ij}\big|\rho^{(t)})dx_{ij}dy_{ij} \\
&= \int_{\theta_{i-1}}^{\theta_i}\int_{\tau_{j-1}}^{\tau_j}\left[-\log(2\pi\sqrt{1-\rho^2})-\frac{1}{2(1-\rho^2)}(x_{ij}^2-2\rho x_{ij}y_{ij}+y_{ij}^2)\right]\times g_{ij}(x_{ij},y_{ij}\big|\rho^{(t)})dx_{ij}dy_{ij} \\
&= -\log(2\pi\sqrt{1-\rho^2})-\frac{1}{2(1-\rho^2)}\{E(x_{ij}^2)-2\rho E(x_{ij}y_{ij})+E(y_{ij}^2)\}
\end{aligned}
$$

and

$$\frac{\partial Q_{ij}}{\partial \rho} = \frac{\rho}{1-\rho^2} + \frac{(\rho^2+1)E(x_{ij}y_{ij}) - \rho(E(x_{ij}^2) + E(y_{ij}^2))}{(1-\rho^2)^2},$$

the derivative of Q with respect to ρ is:

$$\begin{aligned}\frac{\partial Q}{\partial \rho} &= \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\frac{\partial Q_{ij}}{\partial \rho} = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\left[\frac{\rho}{1-\rho^2} + \frac{(\rho^2+1)E(x_{ij}y_{ij}) - \rho(E(x_{ij}^2) + E(y_{ij}^2))}{(1-\rho^2)^2}\right] \\ &= \frac{n\rho}{1-\rho^2} + \frac{(\rho^2+1)\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}E(x_{ij}y_{ij}) - \rho\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(E(x_{ij}^2)+E(y_{ij}^2))}{(1-\rho^2)^2}.\end{aligned}$$

Setting this to 0, and rearranging the terms, we have:

$$\rho(1-\rho^2) + (\rho^2+1)\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}E(x_{ij}y_{ij}) - \rho\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}(E(x_{ij}^2)+E(y_{ij}^2)) = 0$$

and further:

$$\begin{aligned}&\rho(1-\rho^2) + (\rho^2+1)\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}\left[E(x_{ij}y_{ij}) - E(x_{ij})E(y_{ij}) + E(x_{ij})E(y_{ij})\right] \\ &\quad -\rho\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}(E(x_{ij}^2)+E(y_{ij}^2) - E(x_{ij})^2 - E(y_{ij})^2 + E(x_{ij})^2 + E(y_{ij})^2) \\ &= \rho(1-\rho^2) + (\rho^2+1)\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}\left[COV(x_{ij},y_{ij}) + E(x_{ij})E(y_{ij})\right] \\ &\quad -\rho\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}}{n}(V(x_{ij}) + V(y_{ij}) + E(x_{ij})^2 + E(y_{ij})^2) = 0.\end{aligned}$$

The last line is the stationary Eq. (7) in the main text.

References

Bollen, K. A., & Barb, K. H. (1981). Pearson's R and coarsely categorized measures. *American Sociological Review, 46,* 232–239.

Bollen, K. A., & Barb, K. H. (1983). Collapsing variables and validity coefficient (Reply to O'Brien). *American Sociological Review, 48,* 286.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39,* 1–38.

Genz, A., & Bretz, F. (2009). Computation of multivariate normal and *t* probabilities. In *Lecture notes in statistics* (Vol. 195). New York: Springer.

Martin, W. S. (1978). Effects of scaling on the correlation coefficient: Additional considerations. *Journal of Marketing Research, 15,* 304–308.

Muthén, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology, 43,* 131–143.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44,* 443–460.

Pearson, K. (1913). On the measurement of the influence of "Broad Categories" on correlation. *Biometrika, 9,* 116–139.

Pearson, K., & Pearson, E. S. (1922). On polychoric coefficients of correlation. *Biometrika, 14,* 127–156.

R Core Team. (2016). *R*: *A language and environment for statistical computing*. Vienna, Austria: *R* Foundation for Statistical Computing. https://www.R-project.org/.

Ritchie-Scott, A. (1918). The correlation coefficient of a polychoric table. *Biometrika, 12,* 93–133.

Ross, S. M. (2010). *A first course in probability* (8th ed.). New Jersey: Pearson.

Stefan, W., & Manjunath, B. G. (2015). *tmvtnorm*: Truncated multivariate normal and Student t distribution. *R package version 1.4-10.*

A Structural Equation Modeling Approach to Canonical Correlation Analysis

Zhenqiu (Laura) Lu and Fei Gu

Abstract Canonical Correlation Analysis (CCA) is a generalization of multiple correlation that examines the relationship between two sets of variables. Spectral decomposition can be applied and canonical correlations and canonical weights are obtained. Anderson (2003) also provided the asymptotic distribution of the canonical weights under normality assumption. In this article, we propose a Structural Equation Modeling (SEM) approach to CCA. Mathematical forms are presented to show the equivalence among these models. The weight matrix is obtained as the inverse of the loading matrix and the variance or standard errors of weights are calculated through the Delta method. Different popular SEM software such as Lavaan, Mplus, EQS are demonstrated to illustrate the application, and the results are compared with those obtained from Anderson's (2003) formula. Related issues are also discussed in the last section.

Keywords Canonical correlation analysis · Structural equation modeling

1 Introduction

Canonical Correlation Analysis (CCA), first introduced by Hotelling (1936), is a generalization of multiple correlation analysis that examines the relationship between two sets of variables. Following a stepwise procedure, pairs of linear combinations of original variables are derived successively, one from each set, such that the linear combinations of the current pair have maximal correlation and are uncorrelated with all linear combinations of previously derived pairs (Anderson 2003). This stepwise procedure can be continued to derive as many pairs of linear

Z. (Laura) Lu (✉)
The University of Georgia, 325 Aderhold Hall, Athens, GA 30602, USA
e-mail: zlu@uga.edu

F. Gu
McGill University, Montreal, QC H3A 0G4, Canada
e-mail: fei.gu@mcgill.ca

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_22

combinations as the number of variables in the smaller set. In the CCA terminology, these newly created linear combinations of original variables are called *canonical variates*. Those coefficients used in linear combinations to create canonical variates are called *canonical weight coefficients* or just *weights*. The maximal correlations between a pair of canonical variates are called *canonical correlations*, which provide a concise summary of the relationship between the two sets. The correlations between a canonical variate and its original variables are called *canonical loadings*. Finally, the correlations between a canonical variable and the other original variables are called *index coefficients*, or *cross loadings*. From the above description, it is clear that the canonical correlations are exclusively determined by the canonical weight coefficients. Thus, the goal of CCA is essentially to find the optimal weights that maximize these canonical correlations. For normalization purposes, the weight coefficients must also satisfy the unit-variance restriction on each canonical variate, in addition to the bi-orthogonality restrictions on the canonical variates (within-set orthogonality and between-set orthogonality). This stepwise procedure is easy to understand conceptually but not computationally effective. Actual implementations, however, are replaced by a mathematically equivalent spectral decomposition of some quadruple product of covariance/correlation matrices so that all canonical correlations and the associated weight coefficients can be obtained simultaneously. For the set having more variables, additional canonical variates may be derived (which are orthogonal to all existing ones). However, more constraints are needed to uniquely determine these additional weight coefficients (Anderson 2003, p. 499).

Structural Equation Modeling (SEM), evolved from the earlier methods in genetic path analysis of Wright (1918, 1921, 1934; see Bollen 1989), is a generalization of multivariate linear models. SEM includes a very broad set of models such as path analysis models, measurement models, factor models, structural relation models, and latent growth models. It is now being widely used in the social sciences, educational sciences, business, and other fields. Usually, a SEM model includes a measurement model for exogenous variables, a measurement model for endogenous variables, and an overarching structural model for relationships among exogenous and endogenous variables. Unlike CCA, which only investigates the relationship between two sets of observed variables, SEM examines the underlying relationship among many variables, including latent variables in addition to observed variables.

There are connections between CCA and SEM. However, the statistical relation is less obvious and consequently less well known. We found two articles on this topic, Bagozzi et al. (1981), and Fan (1997). Based on the insightful discussion by Bagozzi et al. (1981), canonical correlation analysis could be treated as a special case of a structural relations model. Following this idea, an innovative approach from SEM to CCA was developed by Fan in 1997 when a Multiple Indicators and Multiple Causes (MIMIC) model in SEM was used to analyze CCA. They had tried several SEM models, “but the current example seems to be the only plausible one” (Fan 1997, p. 77). The current example here means the MIMIC model. However, as mentioned in Fan (1997)’s paper, “the representation of CCA using SEM is not straightforward” (Fan 1997, p. 69).

In this article, we will propose a new and a straightforward SEM approach to CCA. The main advantages of the new approach over the approach by Fan include that it is a one-stage procedure, instead of a two-stage procedure, and it can more easily be applied using various existing SEM software packages. Other advantages can be thought of, for instance, missing data of the observed variables can more easily be handled and various robust estimation methods of the canonical correlation coefficient are readily available.

2 A New SEM Approach to CCA

In this section we introduce the new SEM representation of CCA. Before that, we first briefly review the conventional CCA and the SEM model from mathematical forms.

2.1 Conventional CCA

Let $\mathbf{X}$ be a p-variate ($p \geq 1$) zero-mean vector of p random variables in the first variable set, and $\mathbf{Y}$ be a ($p + d$)-variate ($d \geq 0$) zero-mean vector of ($p + d$) random variables in the second variable set. We assume that $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$ are the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21}'$ is the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$. So if we use $\mathbf{Z} = \left[\mathbf{X}' \quad \mathbf{Y}'\right]'$, then the covariance matrix of $\mathbf{Z}$ is

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

Let $\mathbf{a}_{1i}$ and $\mathbf{a}_{2i}$ ($i = 1, 2, \ldots, p$) be the canonical weight vectors for the ith pair of canonical variates (V_{1i}, V_{2i}) of $\mathbf{Z}$, respectively, such that $V_{1i} = \mathbf{a}_{1i}'\mathbf{X}$ and $V_{2i} = \mathbf{a}_{2i}'\mathbf{Y}$. The goal of conventional CCA is to maximize

$$E(V_{1i}V_{2i}) = E\left(\mathbf{a}_{1i}'\mathbf{X}\mathbf{Y}'\mathbf{a}_{2i}\right) = \mathbf{a}_{1i}'\mathbf{\Sigma}_{12}\mathbf{a}_{2i} \ (i = 1, 2, \ \ldots, p),$$

subject to the following unit-variance and orthogonality constraints:

$$\mathbf{a}_{1i}'\mathbf{\Sigma}_{11}\mathbf{a}_{1i} = \mathbf{a}_{2i}'\mathbf{\Sigma}_{22}\mathbf{a}_{2i} = 1, \quad (i = 1, 2, \ \ldots, p), \tag{1}$$

$$\mathbf{a}_{1i}'\mathbf{\Sigma}_{11}\mathbf{a}_{1j} = \mathbf{a}_{2i}'\mathbf{\Sigma}_{22}\mathbf{a}_{2j} = 0, \quad (i \neq j \text{ and } i,j = 1, 2, \ \ldots, p), \tag{2}$$

$$\mathbf{a}_{1i}'\mathbf{\Sigma}_{12}\mathbf{a}_{2j} = 0, \quad (i \neq j \text{ and } i,j = 1, 2, \ \ldots, p), \tag{3}$$

The constraints in (1) restrict each canonical variate to have unit variance, the constraints in (2) require within-set orthogonality, and the constraints in (3) require between-set orthogonality. Taken together, the constraints in (2) and (3) are the so-called bi-orthogonality constraints of conventional CCA.

Computationally, the eigenvectors of the $p \times p$ matrix $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ contain p weight coefficients $\mathbf{a}_{1i}$ ($i = 1, 2, \ldots, p$), and the eigenvectors of the $(p + d) \times (p + d)$ matrix $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ contain $(p + d)$ weight coefficients $\mathbf{a}_{2i}$ ($i = 1, 2, \ldots, p$). The biggest p eigenvalues of $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ are the same as the p eigenvalues of $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. They are equal to the square of canonical correlations of $\mathbf{X}$ and $\mathbf{Y}$.

2.2 SEM

In SEM, there are two types of variables, exogenous variables and endogenous variables. An exogenous variable is a variable whose variability is assumed to be determined by causes outside of the causal model under consideration. And an endogenous variable is a variable whose variability is to be explained by exogenous and other endogenous variables inside of the causal model.

Visually, a SEM model can be commonly depicted as what is called a *path diagram* to show the causality and relationship among variables, in which squares indicate observed variables, circles indicate latent variables, curved lines with arrowheads at both ends show correlated variables, and straight lines with arrowhead at one end tell the causal paths from the end without arrowhead to the end with arrowhead.

Mathematically, a general SEM model without mean can be presented as the following form with three equations,

$$\mathbf{X} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{4}$$

$$\mathbf{Y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{5}$$

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{6}$$

where model (4) is called the X measurement model, in which $\mathbf{X}$ is the vector of observed indicators of $\boldsymbol{\xi}$, a vector of the latent independent (exogenous) variable, and $\boldsymbol{\Lambda}_{\mathrm{x}}$ is a factor loading matrix for $\mathbf{X}$; model (5) is called the Y measurement model, in which $\mathbf{Y}$ is the vector of observed indicators of $\boldsymbol{\eta}$, a vector of latent dependent (endogenous) variables, and $\boldsymbol{\Lambda}_{\mathrm{y}}$ is a factor loading matrix $\mathbf{Y}$; and model (6) is called the structural model, $\mathbf{B}$ is a coefficient matrix for endogenous variables, and $\boldsymbol{\Gamma}$ is a coefficient matrix for exogenous variables; and $\boldsymbol{\delta}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\zeta}$ are measurement error for $\mathbf{X}$ and $\mathbf{Y}$, and residual part for $\boldsymbol{\eta}$, respectively.

2.3 SEM Representation of CCA

In order to derive the new SEM approach to CCA, we adopt the matrix form of conventional CCA. By concatenating all canonical weight vectors horizontally for each set, we obtain $\mathbf{A}_1 = \begin{pmatrix} \mathbf{a}_{11} & \cdots & \mathbf{a}_{1p} \end{pmatrix}$ of order $p \times p$, and $\mathbf{A}_2 = \begin{pmatrix} \mathbf{a}_{21} & \cdots & \mathbf{a}_{2p} \end{pmatrix}$ of order $(p + d) \times p$. We also consider d additional canonical weight vectors $\mathbf{A}_3 = \begin{pmatrix} \mathbf{a}_{2,p+1} & \cdots & \mathbf{a}_{2,p+d} \end{pmatrix}$, and the associated d canonical variates that can be derived from the second variable set $\mathbf{Y}_2$. Anderson (2003, p. 499) discussed that the columns of $\mathbf{A}_3$ can be uniquely determined "by various types of requirements, for example, that the submatrix formed by the lower" d "rows be upper or lower triangular with positive diagonal elements." For the canonical variates generated by vector $\mathbf{A}_3$, we further assume the following unit-variance and orthogonality constraints:

$$\mathbf{a}_{2j}^{'}\mathbf{\Sigma}_{22}\mathbf{a}_{2j} = 1, \quad (j = p+1, p+2, \ldots, p+d) \tag{7}$$

$$\mathbf{a}_{2i}^{'}\mathbf{\Sigma}_{22}\mathbf{a}_{2j} = 0, \quad (i = 1, 2, \ldots, p, \text{ and } j = p+1, p+2, \ldots, p+d) \tag{8}$$

$$\mathbf{a}_{1i}^{'}\mathbf{\Sigma}_{12}\mathbf{a}_{2j} = 0, \quad (i = 1, 2, \ldots, p, \text{ and } j = p+1, p+2, \ldots, p+d) \tag{9}$$

Now let be a $(2p + d) \times (2p + d)$ block-diagonal matrix, in which $\mathbf{A}_1$ is a block of $p \times p$ and $(\mathbf{A}_2\ \mathbf{A}_3)$ is another block of $(p + d) \times (p + d)$. Because of the constraints (1)–(3) and (7)–(9), algebraically the conventional CCA states that

$$\mathbf{A}'\mathbf{\Sigma}\mathbf{A} = \begin{pmatrix} \mathbf{A}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2' \\ \mathbf{0} & \mathbf{A}_3' \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_3 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{A}_1'\mathbf{\Sigma}_{11}\mathbf{A}_1 & \mathbf{A}_1'\mathbf{\Sigma}_{12}\mathbf{A}_2 & \mathbf{A}_1'\mathbf{\Sigma}_{12}\mathbf{A}_3 \\ \mathbf{A}_2'\mathbf{\Sigma}_{21}\mathbf{A}_1 & \mathbf{A}_2'\mathbf{\Sigma}_{22}\mathbf{A}_2 & \mathbf{A}_2'\mathbf{\Sigma}_{22}\mathbf{A}_3 \\ \mathbf{A}_3'\mathbf{\Sigma}_{21}\mathbf{A}_1 & \mathbf{A}_3'\mathbf{\Sigma}_{22}\mathbf{A}_2 & \mathbf{A}_3'\mathbf{\Sigma}_{22}\mathbf{A}_3 \end{pmatrix} = \left(\begin{array}{c:c:c} \mathbf{I}_p & \mathbf{R} & \mathbf{0} \\ \hdashline \mathbf{R} & \mathbf{I}_p & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} & \mathbf{I}_d \end{array}\right)$$

where $\mathbf{0}$ denotes a matrix of zeros of proper dimension, $\mathbf{I}_p$ and $\mathbf{I}_d$ are identity matrices of dimensions p and d, separately, and $\mathbf{R}$ is a diagonal matrix whose diagonal elements are canonical correlations, which are often sorted in descending order.

In other words, the vector of canonical variates of $\mathbf{X}$ and $\mathbf{Y}$ is

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} = \mathbf{A}' \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{pmatrix} \mathbf{A}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2' \\ \mathbf{0} & \mathbf{A}_3' \end{pmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \mathbf{A}'\mathbf{Z} \tag{10}$$

where $\mathbf{V}_1 = \left(V_{11} \cdots V_{1p}\right)'$, $\mathbf{V}_2 = \left(V_{21} \cdots V_{p+d}\right)'$, and its covariance matrix is

$$Cov(\mathbf{V}) = \mathbf{A}'\mathbf{\Sigma}\mathbf{A} = \left(\begin{array}{c:c:c} \mathbf{I}_p & \mathbf{R} & \mathbf{0} \\ \hdashline \mathbf{R} & \mathbf{I}_p & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} & \mathbf{I}_d \end{array}\right) \tag{11}$$

in which **R** is a diagonal matrix whose diagonal elements are canonical correlations, which are often sorted in descending order. In order to maximize the above expectation subject to those constraints, Lagrange multipliers are used. Next, the function is differentiated with respect to the elements of $\mathbf{a}_{1i}$ and $\mathbf{a}_{2i}$.

Based on (10), we can transform it to an equivalent form

$$\mathbf{Z} = \left(\mathbf{A}'\right)^{-1}\mathbf{V}, \tag{12}$$

with Cov(**V**) having the form of (11). Equation (12) is just a simplified Y measurement model in SEM when $\mathbf{\Lambda}_{\mathrm{y}} = (\mathbf{A}')^{-1}$, $\boldsymbol{\eta} = \mathbf{V}$, and $\boldsymbol{\varepsilon} = \mathbf{0}$. We also have a simplified SEM structural model if we assume $\boldsymbol{\eta} = \mathbf{V}$, $\mathbf{B} = \mathbf{I}$, $\mathbf{\Gamma} = \mathbf{0}$, and $\boldsymbol{\zeta} = \mathbf{0}$.

In short, the matrix form of conventional CCA has been represented by

$$\mathbf{Y} = \left(\mathbf{A}'\right)^{-1}\boldsymbol{\eta}, \tag{13}$$

which is a simplified SEM model

$$\mathbf{Y} = \mathbf{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

where $\mathbf{\Lambda}_{\mathrm{y}} = (\mathbf{A}')^{-1}$, $\boldsymbol{\varepsilon} = \mathbf{0}$, $\boldsymbol{\eta} = \mathbf{V}$ with

$$Cov(\boldsymbol{\eta}) = \left(\begin{array}{c:c:c} \mathbf{I}_p & \mathbf{R} & \mathbf{0} \\ \hdashline \mathbf{R} & \mathbf{I}_p & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} & \mathbf{I}_d \end{array}\right).$$

Therefore, the weight matrix **A** in CCA can be obtained as the transpose of the inverse of the loading matrix in SEM, $\mathbf{A}' = \mathbf{\Lambda}_{\mathrm{y}}^{-1}$.

In order to estimate the standard errors of canonical weight coefficients SE(**A**), the Delta method is adopted. Because in the SEM approach we have $\mathbf{A}' = \mathbf{\Lambda}_{\mathrm{y}}^{-1}$, the weight matrix is a function of the loading matrix. So we have

$$\begin{aligned}
\mathrm{var}\left(\mathbf{A}'\right) &= \mathrm{var}\left(\mathbf{\Lambda}_y^{-1}\right) = \mathrm{var}\left(f\left(\mathbf{\Lambda}_y\right)\right)\\
&= \frac{\partial f}{\partial \lambda}\mathrm{var}\left(\mathbf{\Lambda}_y\right)\left(\frac{\partial f}{\partial \lambda}\right)'\\
&= \left[-\mathbf{\Lambda}_y^{-1}\left(\frac{\partial \mathbf{\Lambda}_y}{\partial \lambda}\right)\mathbf{\Lambda}_y^{-1}\right]\mathrm{var}\left(\mathbf{\Lambda}_y\right)\left[-\mathbf{\Lambda}_y^{-1}\left(\frac{\partial \mathbf{\Lambda}_y}{\partial \lambda}\right)\mathbf{\Lambda}_y^{-1}\right]'\\
&= \mathbf{\Lambda}_y^{-1}\left(\frac{\partial \mathbf{\Lambda}_y}{\partial \lambda}\right)\mathbf{\Lambda}_y^{-1}\mathrm{var}\left(\mathbf{\Lambda}_y\right)\left[\mathbf{\Lambda}_y^{-1}\right]'\left(\frac{\partial \mathbf{\Lambda}_y}{\partial \lambda}\right)'\left[\mathbf{\Lambda}_y^{-1}\right]'
\end{aligned}$$

Regarding the asymptotic distribution of the loadings, Anderson (1999, 2003) provided formulas (3.25)–(3.29) to calculate the asymptotic standard errors of canonical weight coefficients for conventional CCA.

2.4 Software Implementation

Convectional CCA can be implemented by using software packages such as Proc CANCORR in SAS/STAT (SAS Institute Inc. 1993), the CCA package in R (R Core Team 2013), the MANOVA command in IBM-SPSS (SPSS 2012), the algebraic function for eigen-analysis in R, MATLAB (MathWorks, Inc. 2012), and SAS/IML.

If we use the SEM approach, then CCA can be implemented in existing SEM software packages, such as the Lavaan package (Rosseel et al. 2013) in R, the SEM package (Fox 2006) in R, Mplus (Muthén and Muthén 2012), EQS (Bentler 1995), the OpenMx package (Boker et al. 2011) in R, and LISREL (Jöreskog and Sörbom 2006).

In this article, we compare the results from the traditional CCA approach to those from the SEM approach by employing the software packages, such as Lavaan, Mplus, EQS and the others. Standard errors are compared between those obtained from the Delta method and those derived from the Anderson's formulas (1999, 2003).

3 Real Data Analysis

In this section, we illustrate the SEM approach to canonical correlation analysis by using a real data example and by comparing the results to those obtained from a regular CCA analysis. In the SEM approach, the loading matrix was first estimated using existing SEM software packages. Next, the inverse of the loading matrix is taken to be the CCA weight matrix.

We analyze the relationships between various kinds of food intake and the mortality rate by various kinds of cancer. There are two sets of variables: variables on food supplies and cancer variables. The data on food supplies are from 34 countries of the world from FAOSTAT (Food and Agriculture Organization of the United Nations 1998). The cancer variables, X variables, consisted of the following four cancer sites: (x1) esophagus, (x2) stomach, (x3) pancreas, and (x4) liver. Seven Y variables were included: (y1) alcohol, (y2) meat, (y3) fish, (y4) cereal, (y5) vegetable, (y6) milk products, and (y7) the total calorie per day. In this case, we have

$$
\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \hline y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \left(\begin{array}{cccc|ccccccc}
\lambda_{11} & & sym & & & & & & & & \\
\lambda_{21} & \lambda_{22} & & & & & & & & & \\
\lambda_{31} & \lambda_{32} & \lambda_{33} & & & & & & & & \\
\lambda_{41} & \lambda_{42} & \lambda_{43} & \lambda_{44} & & & & & & & \\
\hline
& & & & \lambda_{55} & & & & & & \\
& & & & \lambda_{65} & \lambda_{66} & & & sym & & \\
& & & & \lambda_{75} & \lambda_{76} & \lambda_{77} & & & & \\
& & & & \lambda_{85} & \lambda_{86} & \lambda_{87} & \lambda_{88} & & & \\
& & & & \lambda_{95} & \lambda_{96} & \lambda_{97} & \lambda_{98} & \lambda_{99} & & \\
& & & & \lambda_{105} & \lambda_{106} & \lambda_{107} & \lambda_{108} & \lambda_{109} & \lambda_{1010} & \\
& & & & \lambda_{115} & \lambda_{116} & \lambda_{117} & \lambda_{118} & \lambda_{119} & \lambda_{1110} & \lambda_{1111}
\end{array} \right) \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \hline \eta_5 \\ \eta_6 \\ \eta_7 \\ \eta_8 \\ \eta_9 \\ \eta_{10} \\ \eta_{11} \end{bmatrix}
$$

$$
= \left(\begin{array}{c|c} \mathbf{\Lambda}_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{\Lambda}_2 \end{array} \right) \boldsymbol{\eta} = (\mathbf{A}')^{-1} \boldsymbol{\eta}
$$

where

$$
COV(\boldsymbol{\eta}) = \left(\begin{array}{cccc|cccc|ccc}
1 & & & & \rho_1 & & & & & & \\
0 & 1 & & & 0 & \rho_2 & & & & 0 & \\
0 & 0 & 1 & & 0 & 0 & \rho_3 & & & & \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & \rho_4 & & & \\
\hline
\rho_1 & & & & 1 & & & & & & \\
0 & \rho_2 & & & 0 & 1 & & & & 0 & \\
0 & 0 & \rho_3 & & 0 & 0 & 1 & & & & \\
0 & 0 & 0 & \rho_4 & 0 & 0 & 0 & 1 & & & \\
\hline
& & & & & & & & 1 & & \\
& & 0 & & & 0 & & & 0 & 1 & \\
& & & & & & & & 0 & 0 & 1
\end{array} \right) = \begin{pmatrix} \mathbf{I}_4 & \mathbf{R} & \mathbf{0} \\ \mathbf{R} & \mathbf{I}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_3 \end{pmatrix},
$$

We use the full information maximum likelihood (FIML) estimation method, and the −2*loglikelihood value is 3224.4953. First, the point estimates of the four canonical correlations are identical to those obtained from any traditional CCA approach. Second, Table 1 in the Appendix shows the estimates of the loading matrices and their corresponding standard errors (SE) for the CCA analysis using the Lavaan package in R (as the results for the other SEM packages are highly similar, we only focus on the results obtained using Lavaan). Table 2 in the Appendix compares the weight matrices and their standard errors (SE) obtained from (1) the (asymptotic) Anderson formulas, and from (2) the inverse of the loading matrices obtained via Lavaan and the Delta method. We can see both point estimates and SEs are almost the same for the different approaches.

4 Conclusions and Discussion

There are many scholarly significances and advantages of this new SEM approach to CCA. First of all, this approach is very simple and it can be easily applied in various SEM packages. There are many other advantages to be explored. For example, it treats the observed variables as endogenous variables by estimating the loading parameters. Consequently, future research may focus on handling missing data of the observed variables by using FIML. That is, estimating weight coefficients using the other methods, including regular CCA and the Goria and Flury's model (Goria and Flury 1996), will treat the observed variables as exogenous variables. Also, future studies may consider the effectiveness of various robust estimation methods to minimize the contaminating influences of outliers.

Appendix

See Tables 1 and 2.

Table 1 Loading matrices and their standard errors estimated via the Lavann package

est method:	FIML	−2loglik:	3224.496			
Loading1						
−1.836	0.916	−3.074	−2.485			
−13.043	0.462	7.134	5.244			
−2.308	2.109	−3.991	4.071			
−3.385	−3.207	−12.236	3.163			
STDERR1						
5.132	10.035	0.773	0.662			
3.541	71.012	3.327	3.445			
11.717	12.643	1.173	1.011			
17.792	18.515	1.935	1.641			
Loading2						
−23.883	−7.307	−26.061	−9.992	−9.991	−5.611	−4.853
1.021	−7.736	−19.459	−6.402	4.176	5.595	−6.842
−3.436	−6.755	1.842	12.896	1.243	5.476	0.129
−2.228	2.931	22.282	−6.224	15.675	−8.323	−9.280
4.043	−32.032	8.784	−9.451	33.259	10.941	16.060
−6.005	29.614	−7.207	−10.736	5.185	23.932	0
−57.688	−74.403	−197.238	10.947	218.822	0	0
STDERR2						
40.813	130.111	7.297	11.997	9.91	10.389	11.525
42.879	7.046	3.709	8.202	5.817	6.473	4.701
37.518	19.255	3.229	5.117	10.162	10.05	12.244
17.450	13.836	4.888	15.98	6.138	7.991	9.923
177.045	25.116	10.633	31.658	11.89	13.377	10.543
163.53	33.807	8.663	21.006	11.952	11.519	
415.745	321.176	52.665	175.423	38.471		

Note

1. Results from the other SEM software packages, such as Mplus and EQS, are highly similar
2. Loading columns may have different signs for different software packages

Table 2 The weight matrices and their standard errors (SE) (1) from the Anderson's formulas (asymptotic) and (2) from the Delta method (columns are sorted according to the absolute values of canonical correlations)

(1) Weight matrices of the traditional CCA approach and the asymptotic SEs from the Anderson's formulas			
mat1			
0.059	0.105	−0.146	−0.193
−0.025	0.061	0.027	0.004
0.038	−0.030	−0.221	0.107
0.039	0.019	0.124	0.016
SE1			
0.028	0.794	0.579	0.063
0.009	0.149	0.337	0.022
0.029	1.201	0.169	0.074
0.016	0.677	0.107	0.042
mat2			
−0.002	0.033	0.008	−0.012
0.008	−0.030	0.022	−0.013
−0.014	0.028	0.018	0.041
−0.017	0.013	0.002	−0.007
−0.007	0.004	0.013	−0.013
−0.002	0.010	−0.013	−0.005
0.002	0	−0.002	0.002
SE2			
0.004	0.044	0.180	0.011
0.005	0.120	0.163	0.012
0.006	0.100	0.157	0.012
0.003	0.009	0.069	0.005
0.002	0.068	0.024	0.005
0.002	0.069	0.058	0.005
0.000	0.010	0.001	0.001
(2) Weight matrices from the Lavaan SEM package and the SEs from the Delta method			
mat1			
0.060	0.106	0.148	−0.195
−0.025	0.062	−0.028	0.004
0.039	−0.030	0.224	0.108
0.039	0.019	−0.126	0.016
SE1			
0.029	0.745	0.542	0.064
0.009	0.139	0.316	0.022
0.030	1.125	0.160	0.075
0.016	0.634	0.101	0.042

(continued)

Table 2 (continued)

(2) Weight matrices from the Lavaan SEM package and the SEs from the Delta method			
mat2			
−0.002	0.033	−0.008	0.012
0.008	−0.030	−0.022	0.013
−0.014	0.029	−0.018	−0.042
−0.017	0.013	−0.002	0.007
−0.007	0.004	−0.013	0.014
−0.002	0.011	0.013	0.005
0.002	0.000	0.002	−0.002
SE2			
0.004	0.041	0.169	0.013
0.007	0.114	0.153	0.054
0.007	0.094	0.148	0.042
0.004	0.012	0.065	0.031
0.003	0.064	0.023	0.016
0.003	0.065	0.054	0.017
0.001	0.010	0.001	0.004

Note Columns are sorted by the absolute values of canonical correlations

References

Anderson, T. W. (1999). Asymptotic theory for canonical correlation analysis. *Journal of Multivariate Analysis, 70,* 1–29.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York, NY: Wiley.

Bagozzi, R. P., Fomell, C., & Larcker, D. F. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivariate Behavioral Research, 16,* 437–454.

Bentler, P. M. (1995). EQS structural equations program manual. Multivariate Software.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76*(2), 306–317.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do they have in common? *Structural Equation Modeling, 4*(1), 65–79.

Food and Agriculture Organization of the United Nations. (1998). *FAOSTAT statistics database*. http://www.fao.org/faostat/en/#data.

Fox, J. (2006). Teacher's corner: Structural equation modeling with the sem package in R. *Structural Equation Modeling, 13*(3), 465–486.

Goria, M. N., & Flury, B. D. (1996). Common canonical variates in k independent groups. *Journal of the American Statistical Association, 91,* 1735–1742.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 28,* 321–377.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80 for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International.

MathWorks, Inc. (2012). MATLAB and Statistics Toolbox, Natick, Massachusetts, United States.

Muthén, B. O., & Muthén, L. K. (2012). Software Mplus Version 7.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., & Savalei, V. (2013). *lavaan: Latent variable analysis [Software]*. http://CRAN.R-project.org/package=lavaan (R package version 0.5-14).

SAS Institute Inc. (1993). SAS/STAT Software.

SPSS, I. (2012). Statistics for windows, version 20.0. IBM Corp., Armonk, NY.

Wright, S. (1918). On the nature of size factors. *Genetics, 3*(4), 367.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*(7), 557–585.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics, 5*(3), 161–215.

Dealing with Person Differential Item Functioning in Social-Emotional Skill Assessment Using Anchoring Vignettes

Ricardo Primi, Daniel Santos, Oliver P. John, Filip De Fruyt and Nelson Hauck-Filho

Abstract When analyzed via item response theory, Likert-type items are modeled by estimating a set of thresholds (i.e., parameters that inform on the latent trait level required for endorsing a given scale option) that are assumed to be invariant across the population of individuals. If persons vary in response styles this assumption may not hold. This is called person differential functioning (PDIF). Anchoring vignettes offer an approach to learn how individuals translate the latent trait into Likert responses, and a method to assess potential variability in item thresholds across individuals. A vignette presents hypothetical persons differing on the attribute of interest (usually low, medium and high), and asks respondents to rate the hypothetical persons in the same Likert scale used in self-assessment. This can then be used to resolve PDIF, potentially producing measures that are more comparable.

R. Primi (✉) · N. Hauck-Filho
Universidade São Francisco, Campinas, Brazil
e-mail: rprimi@mac.com

N. Hauck-Filho
e-mail: hauck.nf@gmail.com

R. Primi · D. Santos · O. P. John · F. De Fruyt · N. Hauck-Filho
EduLab21, Ayrton Senna Institute, São Paulo, Brazil
e-mail: daniel.ddsantos@gmail.com

O. P. John
e-mail: o_johnx5@berkeley.edu

F. De Fruyt
e-mail: Filip.DeFruyt@UGent.be

D. Santos
Universidade de São Paulo, Ribeirão Preto, Brazil

O. P. John
University of California, Berkeley, USA

F. De Fruyt
Ghent University, Ghent, Belgium

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_23

We investigated if the patters of responses to vignettes have a developmental trend and if they are related to cognitive capacity, using data from a large-scale educational assessment. We then investigated if anchor-adjusted scores produce more reliable and valid measures.

Keywords Anchoring vignettes · PDIF · Social-emotional skill assessment

1 Introduction

1.1 Response Bias in Self Reports

Socio-emotional skills are most frequently assessed in education with self-ratings using rating scales with a Likert format. One issue in these assessments is that the effects of response bias are much more pronounced in children under age 16 and addressing response bias is thus critically important for good measurement (Primi et al. 2016; Soto et al. 2008).

Response bias distorts item responses and leads to construct irrelevant variance, that is, variance due to factors other than the construct one intends to measure. Well-known response biases include (a) *acquiescence*: tendency to choose responses stating agreement regardless of the content of the item, (b) *disacquiescence*: tendency to choose responses stating disagreement regardless of the content of the item, (c) *extreme response bias*: tendency to use the end points of a scale regardless of the content, (d) *middle response bias*: tendency to use the midpoint of the Likert scale regardless of the content, (e) *social desirability bias*: the tendency to answer questions in a way to present oneself in a positive way, and (f) *group reference bias*: systematic differences across respondents regarding internalized group/culture frames of reference to make relative judgments about themselves (e.g., Duckworth and Yeager 2015; He et al. 2014; Wetzel and Carstensen 2015).

Socio-emotional skill assessment through self-reports is popular because such data are easy and inexpensive to collect, especially in large-scale assessments for low-stake purposes (Kyllonen et al. 2014). Despite advantages, self-rating methods assume that participants interpret and use response categories in the same way, and that response styles do not meaningfully affect item responses. The previously described response distortions may operate not only at the level of the individual but may also affect variance at the aggregated level, such as the class, school, or region. Two schools, for example, may have identical latent trait means, but their observed aggregated means may vary because of differential acquiescence. Alternatively, students at different schools may use different frames of reference reflecting socio-economic, cultural, or developmental differences, making comparisons among schools or grades difficult. This multitude of problems encouraged researchers to think about methods to account for response styles and group-reference effects.

1.2 Responses Bias as Differential Person Functioning

In Item Response Theory (IRT), bias in test *items* is generally conceptualized as *Differential Item Functioning* (DIF). DIF occurs when subjects with the same level on the construct but from different groups have different probabilities of choosing a particular answer. The underlying cause is an additional dimension that differs across groups and affects item responses beyond the dimension of interest. Therefore, *group* is a proxy variable for this second dimension that interacts with a particular item, and will hence be indexed as an item-by-group interaction effect on item endorsing. When not accounted for, these differences will be confounded with latent scores, making the groups appear more different than they really are on the dimension of interest.

Individual differences in response styles are better conceptualized as a form of *Differential Person Functioning* (PDIF) (Johanson and Osborn 2004). PDIF occurs when a person differently rates two types of items that measure the same trait with equivalent difficulty but that differ in some irrelevant feature, such as keying direction (e.g., true vs. reverse-keyed). Alternatively, and more generally, we can say that PDIF occurs when two persons with the same level on the construct endorse differently the same item or set of items that have similar difficulty. Again, the underlying cause is a second dimension that varies across persons and affects item endorsement beyond the main dimension of interest. Items that share a feature will be more prone to elicit this particular bias. Being a person variable, it can be modeled as a second latent dimension indexed by a person-by-item-group interaction effect. When not accounted for, it will be confounded with latent scores potentially compromising test validity.

Solving DIF and PDIF involves modeling interaction effects (i.e., item-by-group or person-by-item-group effects). Whereas solving DIF is relatively simple (i.e., by estimating different difficulties for each group), solving PDIF is more challenging. Let's consider the example of group reference bias. Imagine an item designed to measure self-management (conscientiousness), such as '*I'm a careful and dedicated student*; *I always keep my things organized*,' and students are asked to respond on a scale with '1' (not at all like me), '2' (little like me), '3' (moderately like me), '4' (a lot like me) and '5' (completely like me). Imagine two groups A and B with very different reference standards of what is considered an acceptable level of organization and dedication. Imagine that group *A* has a higher standard for organization and dedication in mind than group *B*. Imagine two persons *a* and *b* from these two groups. On the latent level, they could have the same average level on self-management but according to their internalized standards, person *a* will tend to choose a lower value on the item response scale (e.g., a '2') than person *b* (e.g., a '3'). This difference reflects different group standards, rather than a real difference between the two persons. If all items in a scale are affected by these different internalized standards in the same way, then bias will be unidentifiable and confounded with the latent score, such that person *a* will end up with a lower on self-management.

Solving PDIF relies on the existence of item features that are (a) correlated with this second dimension that we want to control for and, simultaneously, (b) are not related to the construct of interest. Consider acquiescence bias as an example where solving PDIF is doable. For instance, we can measure extraversion with items representing *high* levels of sociability such as *i1: 'I talk a lot'* as well as *low* levels such as *i2: 'I tend to be quiet'*. For any person, after controlling for general item difficulties, the expected response on these two items will be the same. As an example, assume that *i1* and *i2* have the same level of difficulty. If a person responds with a 4 on *i1,* the expected response on *i2* would be 2 (because reflected $6 - 2 = 4$). The difference between actual *vs* expected responses can be used to estimate PDIF related to acquiescence (Primi et al. 2017; Soto et al. 2008).

In summary, to solve PDIF using IRT methodology, there is a need for item groups that instantiate features related to the second dimension and that can then be used as contrasts to estimate person-by-item effects. In the case of acquiescence, scores on true and false keyed items can be used to estimate this bias and then correct for it. But what item features could be used to instantiate group reference bias? There is no easy solution to this problem since this type of bias tends to affect all items in the same way (Mõttus et al. 2012). One candidate method to help correct for reference-group bias is using anchoring vignettes.

1.3 Anchoring Vignettes

Anchoring vignettes, initially used in political science, have been suggested as an effective means to control for group reference bias (King et al. 2004; Primi et al. 2016; Wand and King 2008). Specifically, respondents are asked to rate hypothetical persons described in different vignettes; these vignettes vary systematically on the attributes to be assessed, and ratings are obtained using the kinds of items that will also be used for respondents' self-descriptions, with the same rating scale and response format. For instance, at the beginning of the questionnaire, three vignettes are presented describing persons with respectively low, medium and high scores on the skill of negative emotional regulation: (a) **low:** *Beto gets irritated, and he gets easily grumpy. He is always worried about everything, and it is difficult for him to make decisions*, (b) **average:** *Fabiana deals well with stress, and she trusts on her own abilities, but sometimes she gets sad and anxious,* and (c) **high:** *Pedro is calm, and he copes well with tense and stressful situations. He hardly ever feels sad.* The respondent is asked to rate "*How calm and confident do you think is Beto/Fabiana/Pedro?*", providing response options in Likert scale format similar to the self-report items: '1' (not at all like him/her), '2' (a little like him/her), '3' (moderately like him/her), '4' (a lot like him/her), and '5' (completely like him/her). Later in the questionnaire respondents respond to similar items about themselves on the same construct.

If we assume that all the fictitious characters on the vignettes have the intended levels on the construct (the concept of anchors) and the process of rating others is equivalent to ratings of self (according to the consistency principle; see King et al. 2004), respondents should give the same expected ordered ratings for these three vignettes. Any between-person variance in these ratings is a direct indicator of bias. For instance, extreme response bias will be related to greater likelihood of giving extreme scores of ‘1’ to Beto and ‘5’ to Pedro, as compared to more moderate scores of ‘2’ and ‘4’. In contrast, group reference bias, reflecting systematic group differences in benchmarks of what constitutes an average level of negative emotional regulation, would be shown by average differences in Fabiana’s ratings by group A and B.

In summary, responses on vignettes can potentially be used to estimate group-reference bias and then be used as control variables accounting for bias when correlating test scores with external variables. In recent research, anchoring vignettes have been used in the Programme for International Student Assessment (PISA) to correct for some paradoxical findings regarding reversed relationships from the individual to the country level: specific traits were positively related with an outcome at the individual level, though when scores were aggregated at the cultural level, the same trait turned out to be unrelated or even negatively related to this outcome (see: Kyllonen and Bertling 2013; Stankov et al. 2017; von Davier et al. 2017).

1.4 Research Questions

Despite its use in large scale assessments like PISA, little is known about the impact of correcting scale scores using anchoring vignettes and how this affects their psychometric properties and validity. The goals of the present research are four-fold: (a) What is the relationship between original scores and recoded scores using a non-parametric correction relying on vignettes? (b) Is there a developmental trend in response patterns to vignettes? (c) Are response patterns related to indices of cognitive capacity such as achievement tests? (d) Do recoded scores have improved reliability and validity in predicting standardized achievement?

2 Method

2.1 Data

We collected data on vignettes in two representative samples of students attending public schools in Brazil: Sample 1 (Rio de Janeiro): N = 23,133 students from 430 schools attending grades 10th and 12th, and Sample 2 (São Paulo): N = 42,845 students from 500 schools from grades 6th to 12th. The number of students varied across grades in both samples. Total number of students broken down by grades are 6th N = 6,720, 7th N = 4,623, 8th N = 5,855, 9th N = 6,566, 10th N = 21,322, 11th N = 6,015, 12th N = 14,685. For further age comparisons, we also collected data from a small sample of 192 undergraduate students. All data were collected in the course of social-emotional skill assessments conducted by researchers from Edulab21 (Ayrton Senna Institute).

2.2 Instruments

Social-emotional skills were assessed with two self-report versions of SENNA (Primi et al. 2016) that measure five broad social-emotional skill domains, conceptually akin to the dimensions in the Big Five model of personality (e.g., John et al. 2008): **E**: Engaging with others, **A**: Amity, **N**: Negative-emotion regulation, **C**: Conscientious Self-management, and **O**: Open-mindedness.

Across samples, we used ten anchoring vignette sets, two per skill domain. Each set was composed of three descriptions of hypothetical persons representing a high, medium, and low position on each of the five domains. Following each description, participants were asked to rate the socio-emotional skill levels of the various characters using marker items defining either the high pole or the low pole on each construct (5 domains × 2 poles = 10 sets total). High and low pole marker items were 'sociable/outgoing' and 'shy/introverted' for E; 'kind' and 'quarrelsome/selfish' for A; 'calm/confident' and 'nervous/insecure' for N; 'organized' vs 'messy/disorganized' for C; and 'imaginative/creative' and 'little imagination/difficulty to be creative' for O. The complete set of vignettes can be downloaded from http://www.labape.com.br/rprimi/ias/dic_vignettes_v2.xlsx.

As criterion variables we used official standardized achievement test scores (in Portuguese and Math), which we obtained for each student from the education authorities in the State.

3 Results

We first examined the relationship between the original scores and the scores recoded according to the individual's vignettes responses (research question *a*). A non-parametric method is the simplest way to use vignettes to recode scores (King et al. 2004). The procedure is as follows: let y be the subject's response, ranging from 1 to 5 on a self-rating item and v_{lo}, v_{av}, v_{hi} the subject's responses on three vignettes of respectively low, average and high levels on the same domain as the self-rating item. Let z be the recoded response. The rules of recoding are: $z = 1$ if $y < v_{lo}$, $z = 2$ if $y = v_{lo}$, $z = 3$ if $v_{lo} < y < v_{av}$, $z = 4$ if $y = v_{av}$, $z = 5$ if $v_{av} < y < v_{hi}$e, $z = 6$ if $y = v_{hv}$ and $z = 7$ if $y > v_{hv}$. In summary, each vignette response is an anchor point on the Likert scale that informs what level the subject considers low, average and high. Then the respondent's self-rating is compared to the responses to the anchor points and transformed into a new scale from 1 to 7, indicating whether a response is below low, equal to low, above low but below average, equal to average,above average but below high, equal to high and above high. We run these transformations on all items of a domain using a vignette set of the same domain. Original scores are calculated as average endorsements on the items of a domain. Recoded scores are calculated as average endorsements on the recoded items. Therefore, the metric of original scores is on a 1 to 5 scale, whereas recoded scores range from 1 to 7.

Things get more complicated when an individual's responses to vignettes does not follow the normative ordering. For instance, a subject may answer $v_{lo} = 2$, $v_{av} = 1$ and $v_{hi} = 4$. This pattern exhibits a reversion of low with average vignettes (coded as 2, 1, 3). More commonly, subjects tie some vignettes like $v_{lo} = 2$, $v_{av} = 2$ and $v_{hi} = 4$ (coded as {1,2}, 3). For these patterns, more than one recoding is possible. For instance, a self-report response of 2 could be recoded as 2 or 5 in the former and 2 or 4 in the latter case. There is no consensus on how to treat these anomalies. One pragmatic solution used in PISA is to tie reversals and pick up the lowest value when two recoding options are encountered (Kyllonen and Bertling 2013). We followed this procedure.

Correlations between original and recoded scores for E, A, N, C and O were $r = 0.59$, 0.71, 0.93, 0.84 and 0.66, respectively (M = 0.75). Figure 1 shows a scatter plot of original (x-axis) and recoded (y-axis) scores on C in Sample 1 broken

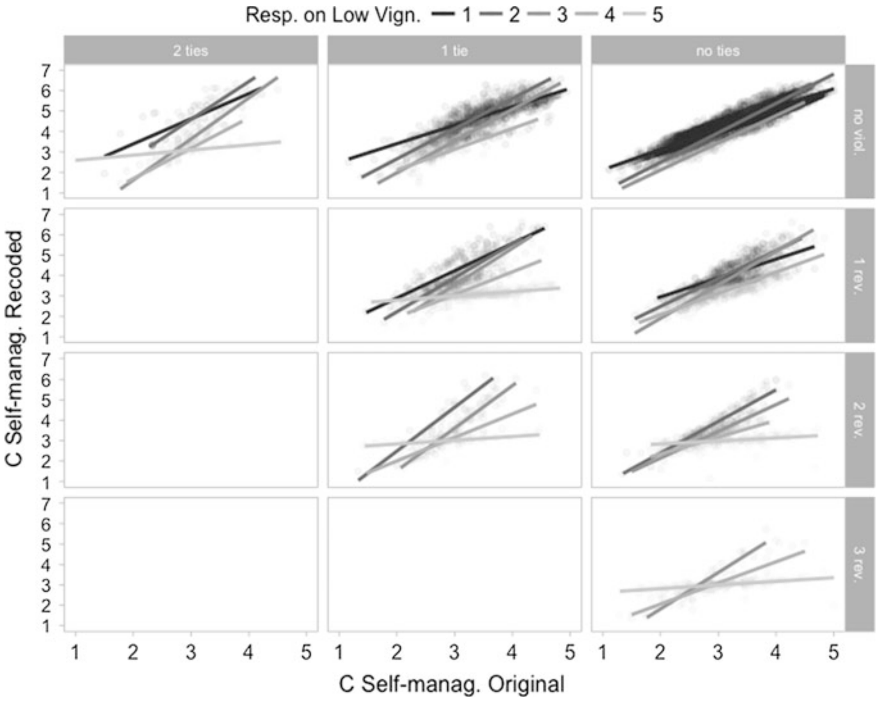


Fig. 1 Original (X) versus recoded (Y) scores, violations (row facet being 0 no reversals, 1, 2 and 3 reversals), ties (column facet being 2 ties, 1 one tie and no ties) and response to low vignette on self-management (color)

down by rows on number of violations (0, 1, 2, 3 reversals) and by column by number of ties (2 ties, 1 tie and no ties). Five different colors of the regression lines indicate the response on the vignette representing the low level of conscientious self-management. Therefore, we expect responses 1 or 2 since this vignette was intended to represent a hypothetical person low in self-management.

Two patterns are important to notice. First, the upper right quadrant shows the relationship of subjects that ordered vignettes perfectly. It can be seen that original *vs* recoded scores are very similar. But slopes of the relationship changes. Those who gave 2 or high responses to the lowest vignettes had a steeper slope than those who gave 1. This a desirable change since subjects who use a restricted range on the rating scale will have their scale stretched from 2–5 to 1–7 after recoding. This is an interesting mechanism of equating extreme and middle responders. Second, when a subject answers 5, a very unexpected response to the lowest vignette, the method lowers the recoded scores. In this case any response from 1 to 4 will be recoded to 1 when using the lowest value rule. Therefore, subjects that gave this unexpected response will have almost all items recoded to 1 or 2 (if they give 5) and therefore will have a lower score on self-management.

Next we examined developmental patterns on vignettes (research question *b*). We calculated the proportion of vignettes correctly ordered. We consider an order correct if vignettes ratings are ordered as the expected pattern of low, average, and high (1,2,3) or if they tie adjacent vignettes, namely low with average or average with high ({1,2},3 or 1, {2,3}). Each subject obtained a score ‘1’ if their ratings showed a correct pattern and ‘0’ otherwise. Figure 2 shows the proportions of this variable by vignette set (for the five domains) and for each grade from 6 to 12 (where 12 is the last year of high school) and “grade 16” refers to undergraduate students; these values are from Sample 2 plus the sample of undergraduate students.

Figure 2 shows clear evidence for the expected developmental pattern: the proportion of correctly ordered responses increased with the grade level of the students. Older students (i.e., in the higher grades) were more likely correct, whereas younger students made more mistakes. Indeed, more than 80% of undergraduate university students (Grade 16 in Fig. 2) ordered all the vignettes correctly. But the percentage of correct orderings was as low as 30% in 6th grade students for the most difficult vignette set, Open-mindedness rated on the reverse-keyed item. In general, we found a linear increase associated with school grades, topped by the much-better educated university students. Younger students in grades 6 and 7 made more errors and showed more unexpected patterns. The vignettes for Open-mindedness and Negative Emotion Regulation were generally more difficult than the other three sets.

Research question *c* examined how performance on the vignettes is related with standard indicators of cognitive development in school, namely scores on academic achievement tests. For this analysis, we first calculated a Global Consistency Index (GCI) from vignettes’ responses. Since vignettes have a predefined, normative order, the expected order of the response for each vignette was 1 for the low vignettes, 2 for the average vignettes and 3 for the high vignettes. For each subject, we paired these expected vectors with their responses and calculated a correlation

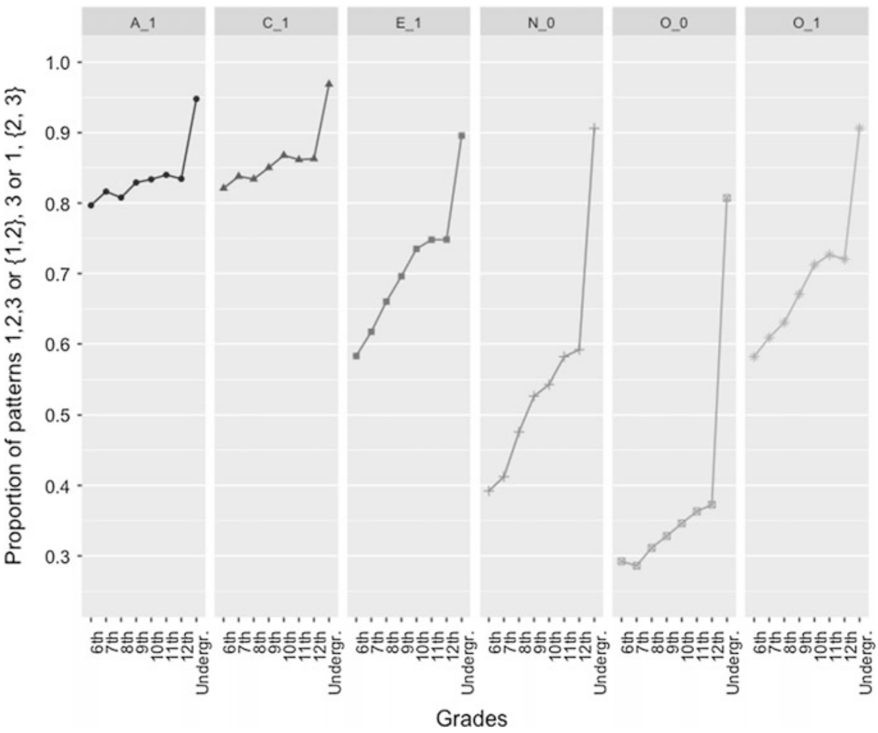


Fig. 2 Proportion of correctly ordered patterns (y-axis) by grade (x-axis) on six vignettes sets (A: Amity, C: Conscientious Self-management, E: Engaging with others, N: Negative Emotion Regulation, and O: Open-mindedness (in addition, codes "_1" or "_0" after each domain code indicate whether the questions about the persons in the vignettes were true keyed (e.g., 'imaginative/creative' for O) or false keyed (e.g., 'little imagination/difficulty in being creative' for O)

coefficient between the two vectors. Therefore, the GCI is a within-subject correlation of expected order of vignettes and subjects' actual responses. It is a metric going from a minimum of −1 (completely reversed order in all vignette sets) to +1 (perfectly correct order in all vignette sets). In Sample 1, for example, the mean of this consistency index was 0.61 (*SD* = 0.32, skew = −1.04). That is, this index has generally positive values and its distribution is skewed negatively, indicating that most students give responses in line with the normative order. Consistent with the age trends in Fig. 2, we found that vignette consistency was positively correlated with school achievement; in Sample 1, the consistency index correlated 0.45 with achievement test scores in Portuguese and 0.32 with achievement in Math. These are substantial correlations and indicate that students with greater reading-writing skills and quantitative knowledge also gave more consistent responses on vignettes.

Finally, we explored whether recoded Big Five scores were more reliably and valid than original scores (research question *d*). Table 1 shows internal consistency coefficients of the Big Five scales and their correlations with standardized

Table 1 Reliability indices and criterion validity (correlation with standardized achievement in Portuguese and math) of original and recoded scores in Sample 1

Domains		A	C	E	N	O
Reliability						
Original		0.78	**0.87**	0.79	**0.87**	0.84
Recoded		**0.85**	0.82	**0.82**	0.77	**0.91**
Criterion validity						
Original	Port	0.11	0.13	0.02	0.01	0.19
Recoded	Port	0.09	0.13	0.08	0.00	0.20
Original	Math	0.07	0.10	0.00	0.05	0.10
Recoded	Math	0.06	0.10	0.03	0.04	0.12

achievement. Recoding increased the reliability estimates for three of the five dimensions but decreased reliability for the two other domains (for C and O). Validity coefficients (i.e., correlations with scores on the Portuguese and Math achievement tests) stayed largely unchanged.

4 Discussion

Together, these findings indicate three main points about what vignettes measure. First, rating vignettes correctly (i.e., consistent with normative expectations) was related to age and to achievement scores in Portuguese and mathematics. These findings suggest that vignette performance is more strongly related to aspects of cognitive development than previously realized. Vignette performance showed similar patterns of associations with external variables—such as grade level in school and cognitive measures of school achievement—as traditional intelligence measures do (see Stankov et al. 2017 for similar results). At the same time, vignette performance seems to tap specific knowledge about people and their typical socio-emotional functioning (e.g., engaging with others; self-management in task contexts) and how to express this knowledge in numerical scales. The index we have proposed to capture this complex set of social-cognitive skills (GCI) could be used as a measure of children's psycho-social maturity or readiness to provide self-ratings on psychological characteristics.

Second, vignettes can be a potential way to measure and solve response bias via recoding. However, when anomalies in the vignette rating process are present, recoding can change scores in an undesirable way. Students who make a lot of errors in vignette ratings also had lower school achievement scores. Recoding using the lowest value will hence lower the socio-emotional skills' scores of students making order violations. A confounder is introduced when assessing the relationship of recoded scores with achievement. Now, the recoded socio-emotional skill scores are contaminated with the criterion (a cognitively affected variable), leading to spuriously higher associations of recoded scores with achievement (Primi et al. 2016).

Finally, recoded scores seemed to be more reliable in some cases, though this is again a spurious increase due to a method artifact of the non-parametric recoding.

If these increases were real increments in true variance related to total variance we should have seen improved validity. Recoding introduces a dependency of item to vignette responses since they become a function of the responses on vignettes, thus increasing the correlation among items due to this common source variance. Covariances among recoded items are no longer covariances among independent observations, but covariances among items dependent on common vignettes items. On top of that, if a subject violates the normative vignette ordering, their recoded responses will be even more similar. Individual differences in the knowledge necessary to order vignettes properly will further affect item responses, changing responses in an undesirable way since it introduces other information, rather than the desired correction for bias. Von Davier et al. (2017) proposed a mathematical formulation of this problem and presented a simulation study that suggests that violations are responsible for pseudo-increases in reliability.

Several limitations of the present research need to be considered. First, the specific vignettes used here might not be valid measures of the type of biases we intended to assess and control. Second, there is increasing evidence that the non-parametric method to recode responses is not ideal (see von Davier et al. 2017). Other methods have been based on item response modeling (see Bolt et al. 2014), and their use may have achieved different results here. Finally, vignettes have shown promise primarily in studies like PISA that aim to compare scores of samples from different cultures or languages. The present samples come from a single culture. More research is needed to address whether the utility of vignettes may be limited to cross-cultural research contexts.

In conclusion, the present results suggest a cautiously optimistic stance regarding the utility of vignettes and their use for correcting response biases in mono-cultural research. After positive initial findings in PISA, further research is needed to elaborate what vignettes really measure and to test whether they are useful for capturing other kinds of response styles. For example, Mõttus et al. (2012) suggested that vignettes response patterns may capture extreme response style bias. Acquiescence is another response style that has recently attracted renewed attention (e.g., Soto et al. 2008); we suggest that future research examine whether vignettes may offer a novel way to assesses acquiescence and to correct for that response style. More generally, more work is needed to better distinguish among different kinds of response biases and to test whether and how vignettes can be employed to improve the quality of measurement in psychological and educational contexts. Multidimensional item response models will likely prove the most productive way forward for this important line of research.

References

Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods, 19*(4), 528–541. https://doi.org/10.1037/met0000016.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters. *Educational Researcher*. https://doi.org/10.3102/0013189X15584327.

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. J. R. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology, 45*(7), 1028–1045. https://doi.org/10.1177/0022022114534773.

Johanson, G. A., & Osborn, C. J. (2004). Acquiescence as differential person functioning. *Assessment & Evaluation in Higher Education, 29*(5), 535–548. https://doi.org/10.1080/0260293042000160401.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). New York, NY: Guilford Press.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 191–207. https://doi.org/10.1017/S000305540400108X.

Kyllonen, P., & Bertling, J. (2013). Innovative questionnaire assessment methods to increase cross country comparability. In *Handbook of international large-scale assessment* (Vols. 1–0, pp. 277–285). Chapman and Hall/CRC. https://doi.org/10.1201/b16061-15.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2014). Personality, motivation, and college readiness: A prospectus for assessment and development: Personality, motivation, and college readiness. *ETS Research Report Series, 2014*(1), 1–48. https://doi.org/10.1002/ets2.12004.

Mõttus, R., Allik, J., Realo, A., Pullmann, H., Rossier, J., Zecca, G., et al. (2012). Comparability of self-reported conscientiousness across 21 countries. *European Journal of Personality, 26*(3), 303–317. https://doi.org/10.1002/per.840.

Primi, R., Santos, D., John, O. P., De Fruyt, F., & Filho, N. H. (2017). *Acquiescence and person differential functioning (DIF): Solving person DIF with balanced scales*. Paper submitted for publication.

Primi, R., Santos, D., John, O. P., & De Fruyt, F. D. (2016a). Development of an inventory assessing social and emotional skills in Brazilian youth. *European Journal of Psychological Assessment, 32*(1), 5–16. https://doi.org/10.1027/1015-5759/a000343.

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016b). Anchoring Vignettes. *European Journal of Psychological Assessment, 32*(1), 39–51. https://doi.org/10.1027/1015-5759/a000336.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology, 94*(4), 718–737. https://doi.org/10.1037/0022-3514.94.4.718.

Stankov, L., Lee, J., & von Davier, M. (2017). A note on construct validity of the anchoring method in PISA 2012. *Journal of Psychoeducational Assessment*, 0734282917702270. https://doi.org/10.1177/0734282917702270.

von Davier, M., Shin, H.-J., Khorramdel, L., & Stankov, L. (2017). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement*, 0146621617730389. https://doi.org/10.1177/0146621617730389.

Wand, J., & King, G. (2008). *Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes* (SSRN Scholarly Paper No. ID 1082000). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=1082000.

Wetzel, E., & Carstensen, C. H. (2015). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, 1–13. https://doi.org/10.1027/1015-5759/a000291.

Random Permutation Tests of Nonuniform Differential Item Functioning in Multigroup Item Factor Analysis

Benjamin A. Kite, Terrence D. Jorgensen and Po-Yi Chen

Abstract The purpose of the present research was to introduce and evaluate random permutation testing applied to measurement invariance testing with ordered-categorical data. The random permutation test builds a reference distribution from the observed data that is used to calculate a p value for the observed $(\Delta)\chi^2$ statistic. The reference distribution is built by repeatedly shuffling the grouping variable and then saving the $\Delta\chi^2$ statistic between the two models fitted to the resulting data. The present research consisted of two Monte Carlo simulations. The first simulation was designed to evaluate random permutation testing across a variety of conditions with scalar invariance testing in comparison to an existing analytical solution: the robust mean- and variance-adjusted $\Delta\chi^2$ test. The second simulation was designed to evaluate the random permutation test applied to testing configural invariance by evaluating overall model fit (the χ^2 fit statistic). Simulation results and suggestions for the use of the random permutation test are provided.

Keywords Measurement invariance · Differential item functioning
Ordered-categorical data · Permutation · Multiple group confirmatory
factor analysis

B. A. Kite · P.-Y. Chen
University of Kansas, Lawrence, KS, USA
e-mail: bakite@ku.edu

P.-Y. Chen
e-mail: p090c021@ku.edu

T. D. Jorgensen (✉)
University of Amsterdam, Nieuwe Achtergracht, 127 (Room D7.17),
1018 WS Amsterdam, The Netherlands
e-mail: T.D.Jorgensen@uva.nl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_24

1 Introduction

Behavioral researchers often use multiple-group confirmatory analysis (MG-CFA) to test measurement invariance (MI) with indicator variables on a Likert-type scale. The procedure of testing MI can be seen as a procedure of finding items with differential item functioning (DIF). In a MG-CFA framework, testing MI with ordinal data usually involves comparing nested invariance models. To test hypotheses about different levels of invariance, researchers could first use the ordinal estimators based on polychoric correlations from software such as M*plus* and `lavaan`, which employ diagonally weighted least squares (DWLS) estimation. A robust a mean- and variance-adjusted test statistic can be requested in M*plus* using the command "ESTIMATOR = WLSMV" or from `lavaan` using the argument `estimator` = "`WLSMV`", where the "MV" stands for the mean and variance adjustment to the chi-squared test statistic. MI testing can be conducted by comparing the global fit indices such as chi-squared statistic (χ^2) or alternative fit indices (AFI) between invariance models. Among different criteria developed for MI testing, researchers have found that the chi-squared difference ($\Delta\chi^2$) test substantially outperforms other fixed cutoffs based on change in AFI (e.g., change in CFI) by showing greater power and a better ability to control Type I error rate across different scenarios (Sass et al. 2014).

The $\Delta\chi^2$ tests of ordinal estimators in MG-CFA usually require researchers to apply robust corrections during the testing procedures to mitigate the influences of not using consistent estimators for the weight matrix in fit function (Savalei 2014). Software such as M*plus* (Muthén and Muthén 2015) and `lavaan` (Rosseel 2012) both provide robust $\Delta\chi^2$ tests for researchers to compare invariance models estimated by DWLS. Unfortunately, even though robust $\Delta\chi^2$ tests are considered best practice for testing MI with ordinal data in MG-CFA, there are some important issues that warrant further attention.

Most simulation research of the mean- and variance-adjusted $\Delta\chi^2$ test utilizes the implementation provided by M*plus* with the DIFFTEST command when using ESTIMATOR = WLSMV. Researchers have found contradictory conclusions during simulations about its ability to control Type I error rate (see the following sections for details). The mean- and variance-adjusted $\Delta\chi^2$ test is also implemented in `lavaan` via the `lavTestLRT` function, but it has not been examined in a published Monte Carlo simulation. Furthermore, the corrected χ^2 statistic obtained through WLSMV also has been shown to be inappropriate to test the configural invariance assumption (whether the item-factor configurations are identical across groups) when the model is only an approximation of the true population model (Jorgensen et al. 2017), but evidence of inflated Type I error rates under certain conditions (Bandalos 2014) suggests that a test of overall model fit could yield inflated Type I errors even when models fit perfectly.

To address these issues, in the current study, we propose a nonparametric method for testing MI based on the permutation test. We compare the robust $(\Delta)\chi^2$

tests provided by M*plus* and lavaan with two simulation studies. Through these simulations, we provide researchers (a) explanations about contradictory conclusions in previous studies about the robust $\Delta\chi^2$ test in M*plus*, (b) systematic evaluations of the robust $\Delta\chi^2$ test provided by lavaan, and (c) a new solution that can outperform robust $\Delta\chi^2$ test under conditions when it fails to yield nominal error rates. The rest of this article is organized as follows. We first briefly introduce the robust $(\Delta)\chi^2$ tests provided by M*plus* and lavaan, then explain their problems in MI testing. After that, we illustrate the rationale of the permutation test we propose and explain its theoretical advantages. Lastly, we investigate the relative performances between methods through our simulations and provide recommendations for researchers.

2 The Robust $\Delta\chi^2$ Test in M*plus* for Testing MI with Ordinal Data

The robust $\Delta\chi^2$ test provided by M*plus* is a widely used implementation for MI testing with ordinal data in MG-CFA recommended by popular structural equation modeling textbooks (e.g., Kline 2016; Little 2013). Muthén and Muthén (2015) suggested that researchers use the DIFFTEST command in M*plus* in order to correctly scale $\Delta\chi^2$. The DIFFTEST command in M*plus* applies the mean and variance adjustment to the $\Delta\chi^2$ statistic between nested models, as discussed by Asparouhov and Muthén (2006; see also Satorra 2000). The parent model (e.g., a configural model) is fitted to the data, and matrices containing information about the model are saved in a separate output file. When the nested model (e.g., a scalar invariance model) is fitted and the text file containing matrices from the parent model is provided, DIFFTEST uses information from both models to compute a "scaled and shifted" $\Delta\chi^2$ statistic that asymptotically yields nominal Type I error rates. A more detailed explanation of the computation involved with the DIFFTEST command can be found in Asparouhov and Muthén (2006).

3 The Robust $\Delta\chi^2$ Test in lavaan for Testing MI with Ordinal Data

Besides M*plus*, empirical researchers could also use the "lavTestLRT" function provided by lavaan for MI testing (Rosseel 2012). When two nested models are supplied to the lavTestLRT function, the correction outlined by (Satorra 2000) is applied to produce a mean- and variance-adjusted $\Delta\chi^2$ statistic. Within the lavTestLRT function in lavaan, there are two options for how to compute the

Jacobian of the constraint function. The first option (`method = "exact"`) is to calculate an exact solution from a constraint function applied to the full parameter vector, which requires that the two models are nested in the parameter sense, not the more flexible sense of nested covariance structures (Bentler and Satorra 2010). The second option (`method = "delta"`) provides an approximation to the Jacobian and only requires models to be nested in covariance sense, such that the set of predictions that could possibly be made by the parent model include all possible predictions made by the nested model. In the present research, we used the second option, which is `lavaan`'s default method beginning with version 0.6-1.1109.

4 Problems with Currently Available Methods

Asparouhov and Muthén (2006) conducted a small simulation to show that their robust $\Delta\chi^2$ test effectively controls the Type I error rate when the total sample sizes are asymptotically large: 1100 and 2200. A follow-up study conducted by Sass et al. (2014) found contradicting results when sample sizes were more realistically small or moderate. Specifically, Sass et al. found that the Type I error rate of the robust $\Delta\chi^2$ test provided by M*plus* was always substantially inflated in all of their conditions with symmetrically distributed thresholds (range from 7–9%), and 6–9% in asymmetric conditions. One explanation to these contradicting results could be that the sample sizes that Sass et al. examined are in general smaller than the sample sizes in Asparouhov and Muthén (2006), and small samples are inconsistent with the derivation of the robust test statistic, which relies on asymptotic theory. However, if the $\Delta\chi^2$ statistic obtained from WLSMV requires more than 1000 observations, then its applicability will be severely limited, considering most of MI studies in psychology won't have this large of sample size (Putnick and Bornstein 2016).

After thoroughly examining the results in Sass et al. (2014), we found another possible explanation. That is, in their simulations the scalar invariance model was different from the ordinary settings by unnecessarily constraining two additional parameters. Specifically, to make sure the configural model was identified, Sass et al. fixed the mean and variance of latent factor to 0 and 1 in both groups. When estimating the scalar invariance model, Sass et al. did not release these two constraints in the second group as suggested in literature, which resulted in an overly stringent scalar invariance model (Kline 2016; Little 2013). We believe this could be another reason that caused their inflated Type I error rates.

According to our knowledge, there is still no study evaluating the performance of the `lavTestLRT` function in `lavaan`, despite its use by empirical researchers (e.g., Antoniadou et al. 2016). Note that Satorra (2000) originally proposed the adjustment for the $\Delta\chi^2$ statistic to correct for continuous non-normal data, not categorical data. The utility of this $\Delta\chi^2$ correction with ordinal estimators like

WLSMV seems to rest quite heavily on the asymptotic assumption. We therefore think it is worthwhile to conduct a simulation to compare different implementations of the correction that might not be equivalent in small to moderate samples, such as the DIFFTEST procedure in M*plus* and the `lavTestLRT` procedure in `lavaan`.

Finally, besides the unsolved issues we mentioned for the robust $\Delta\chi^2$ test in M*plus* and `lavaan`, we believed there is also a common limitation shared by the robust χ^2 obtained from the WLSMV estimator in both software packages. Specifically, we believe the χ^2 obtained from WLSMV estimator might not be a valid statistic for evaluating the configural invariance in small to moderate samples because it is derived from asymptotic theory. Bandalos (2014) found inflated Type I error rates for the robust χ^2 statistic when the sample size is small, especially when thresholds are asymmetrically distributed.

5 Permutation Tests of MI with Ordinal Data

To solve the problem of $(\Delta)\chi^2$ test statistics mentioned above, we proposed a permutation test of MI with ordinal data, which would be free from asymptotic theory and should be able to control the Type I error rate reasonably well regardless of the sample size and distribution of the thresholds. Specifically, we propose to apply the random permutation testing to $(\Delta)\chi^2$ with ordered-categorical data to overcome the issue of the difference statistic not following a central χ^2 distribution. The focus of the present research is demonstrating how this approach works and evaluating its performance. The proposed random permutation test is a nonparametric method based on the idea of building an empirical reference distribution reflecting the null hypothesis that groups have the same model configuration and measurement parameters. In other words, the reference distribution is built under the assumption of a true null hypothesis that there is no effect of group membership on measurement properties (e.g., configuration, parameter values). This reference distribution is used to calculate a *p* value when testing the null hypothesis of invariance. The benefit of permutation testing is that building a nonparametric reference distribution alleviates many of the assumptions of standard parametric hypothesis tests. When testing for the effect of group membership on a test statistic, a null distribution can be built by randomly shuffling the grouping variable and saving the resulting test statistic after each shuffle. If there is no difference in measurement-model configurations or parameters between groups, the observed test statistic (calculated from the original data) should be consistent with the values created by randomly shuffling the grouping variable; that is, the observed value would only exceed the upper 95th percentile of the permuted values 5% of the time. This should keep the Type I error rate of the test procedure nominal (i.e., at 5% when using $\alpha = 0.05$). Building a null distribution this way is especially useful when the distribution of the test statistic is unknown.

6 Method

To address the issues of the currently available two methods mentioned in the introduction, we conducted two Monte Carlo studies. Study 1 is designed to compare the relative performances between the robust $\Delta\chi^2$ test provided by M*plus*, the robust $\Delta\chi^2$ test provided by `lavaan`, and our new proposed permutation method on detecting DIF. In Study 1, based on the assumption that researchers have confirmed configural invariance hypothesis, we conducted the $\Delta\chi^2$ tests between scalar and configural invariance model with the three methods above. The relative performances between methods were evaluated in terms of Type I error rate and power across 1000 replications within each condition. In simulation Study 2 we focused on the performance of the Type I error rate the χ^2 obtained from the three methods. In Study 2 we examined whether the corrected χ^2 provided in M*plus* and `lavaan` would reject the configural invariance model too often in comparison to the permutation method we proposed. In both simulations, we follow Sass et al. (2014) and used (0.036, 0.064) as the acceptable range for observed Type I error rates, In both simulations, data were generated in R using the `simulateData` function in `lavaan`. A two-group, single-factor, model with eight indicator variables was used as the population model. The factor loadings were fixed at 0.6 except in conditions when loadings were not invariant (i.e., when the loadings of first two items in Group 2 were different from Group 1). Residual variances for indicator variables were always set at $1 - \lambda^2$ so that latent item responses would have unit variance. The number of shuffling with each permutation test was set to be 500. The design factors we manipulated in the two simulations (i.e., sample size, distribution of thresholds, the number of categories per item, and the presence of measurement non-invariance) are illustrated as follows.

Study 1 evaluated the random permutation $\Delta\chi^2$ against analytically derived robust $\Delta\chi^2$ test statistics. The simulation design was a fully crossed 2 (response categories) × 2 (threshold symmetry) × 2 (sample size) × 2 (factor loading invariance) design resulting in 16 between-replication conditions used to generate data, each having 1000 replications. In each replication, four different $\Delta\chi^2$ tests were conducted: robust $\Delta\chi^2$ tests in M*plus* and `lavaan`, our permutation test for $\Delta\chi^2$, and an unadjusted $\Delta\chi^2$ test as a reference.

In Study 1, we set the sample size as 300 (150 per group) or 600 (300 per group). These settings are similar to the small and medium sample sizes Sass et al. (2014) used. The number of categories per item was set to be 2 or 5 to represent the dichotomous and ordinal scales that researchers frequently used in practice. In addition, we also simulated either symmetrically or asymmetrically distributed thresholds, given that previous studies have found that he distribution of thresholds could affect the results of $\Delta\chi^2$ related tests (e.g., Sass et al. 2014). Specifically, in conditions with ordinal items, the symmetric and asymmetric thresholds are set to be (−1.30, −0.47, 0.47, 1.30) and (−0.25, 0.38, 0.84, 1.28) as used by Sass et al. (2014). Threshold values for symmetrically and asymmetrically dichotomous items are set to be 0 and 0.7 respectively, as the average of the conditions manipulated in previous research (Beauducel and Herzberg 2006; Rhemtulla et al. 2012). The non-invariance we manipulated in the current study is limited to factor loadings. Specifically, in Study 1 we created non-invariance by subtracting 0.25 (Sass et al. 2014) from the factor loadings for Items 1 and 2 in the population model in the focal group. Specifically, in non-invariant conditions, the factor loadings of Items 1 and 2 in the model will be 0.60 in the reference group but were $0.6 - 0.25 = 0.35$ in the focal group. In contrast, Items 3–8 in both groups always had factor loadings of 0.60 in all conditions.

There were two models compared in each replication: a configural invariance model and a scalar invariance model. The configural model had the factor loadings and thresholds freely estimated for both groups, whereas the latent variable in each group had its estimated mean and variance fixed to be 0 and 1, respectively. Further, in the configural model, the variances of the latent response variables (i.e., scales of normally distributed responses assumed to underlie observed discrete item responses) were fixed to 1 in both groups (i.e., we used the so-called "delta" method of identification available in M*plus* and `lavaan`). The scalar invariance model had the factor loadings and thresholds constrained to equality across groups. Constraining the measurement parameters across groups allowed the latent variable mean and variance to be estimated in the focal group rather than fixed to 0 and 1.

The simulation conditions of Study 2 are almost identical to those of Study 1 except we removed the non-invariant conditions and the estimation of scalar invariance, given the exclusive focus on Type I error rates of the χ^2 statistic for the configural invariance model. Additionally, in order to increase the magnitude of asymmetry in our data to better match the work of Bandalos (2014), we changed the distribution of asymmetric thresholds to (1.198) and (0.85, 1.10, 1.45, and 2.00).

7 Results

Type I error rates for tests of scalar invariance are shown in Table 1. Results showed that random permutation testing and `lavTestLRT` had reasonable Type I error control. The random permutation test had Type I errors within the nominal range of 0.036–0.064 in all eight equal measurement parameter conditions, whereas the M*plus* DIFFTEST procedure had inflated error rates in the two conditions where there were two response options with asymmetric thresholds, even though the inflation is not as severe as Sass et al. (2014) found with ordinal data.

Power for scalar invariance tests are shown in Table 2. The M*plus* DIFFTEST procedure consistently showed the highest power, with `lavTestLRT` showing power equal to or greater than the random permutation test (see Table 2). All testing procedures showed higher power in conditions higher group sizes, more response categories, and symmetric thresholds.

The results of simulation Study 2 in Table 3 showed that the random permutation test of configural invariance had acceptable Type I error control in all eight study conditions. The mean- and variance-adjusted χ^2 tests provided by M*plus* and

Table 1 Type I error rates for $\Delta\chi^2$ tests

N	# Categories	Thresholds	Permutation	M*plus*	`lavaan`	Unadjusted
150	2	Symmetric	0.050	0.060	0.056	0.143
300			0.043	0.052	0.050	0.128
150	5		0.053	0.062	0.054	0.131
300			0.053	0.057	0.053	0.098
150	2	Asymmetric	0.053	0.065	0.054	0.135
300			0.056	0.078	0.065	0.139
150	5		0.050	0.053	0.047	0.131
300			0.054	0.062	0.056	0.128

Table 2 Power for $\Delta\chi^2$ tests

N	# Categories	Thresholds	Permutation	M*plus*	`lavaan`	Unadjusted
150	2	Symmetric	0.279	0.319	0.292	0.452
300			0.543	0.568	0.543	0.703
150	5		0.460	0.504	0.464	0.618
300			0.786	0.811	0.794	0.890
150	2	Asymmetric	0.214	0.258	0.225	0.361
300			0.406	0.457	0.427	0.588
150	5		0.342	0.370	0.335	0.519
300			0.707	0.733	0.712	0.831

Table 3 Type I error rates of χ^2 test in the configural invariance model

N	# Categories	Thresholds	Permutation	M*plus*	`lavaan`	Unadjusted
150	2	Symmetric	0.051	0.049	0.049	0.001
300			0.048	0.052	0.052	0.001
150	5		0.054	0.066	0.066	0.000
300			0.057	0.059	0.059	0.000
150	2	Asymmetric	0.047	0.049	0.049	0.006
300			0.039	0.051	0.052	0.004
150	5		0.049	0.202	0.199	0.014
300			0.035	0.100	0.101	0.002

`lavaan` performed nearly identically and showed inflated Type I errors in conditions with asymmetric thresholds with five response options. The error rates were especially inflated with five response options when the group sizes were 150 (20.2% and 19.9%), and improved but still inflated when the group sizes were 300 (10% and 10.1%). Lastly, the unadjusted χ^2 test provided by `lavaan` showed error rates well below the nominal value of 0.05 in all conditions.

8 Discussion

The purpose of the present research was to evaluate the use of random permutation testing applied to $\Delta\chi^2$ tests with ordered-categorical indicator variables. The research was focused on models estimated with the popular WLSMV estimator. When models with ordered-categorical data are estimated with WLSMV, the $\Delta\chi^2$ related tests require a mean and variance adjustment (Asparouhov and Muthén 2006; Satorra 2000). The random permutation test was introduced as an alternative that is easily implemented in any statistical software, and as a method that should control Type I errors as well or better than existing methods. Study 1 evaluated the random permutation $\Delta\chi^2$ test for measurement invariance in comparison to existing analytical robust solutions, and served as a follow-up to Sass et al. (2014). Study 2 expanded on the work of Jorgensen et al. (2017) and served as a follow-up to Bandalos (2014).

Overall, the random permutation test performed well in both simulations. In Study 1 the random permutation test was the only method that consistently showed Type I errors within the previously defined nominal range of 0.036 and 0.064. Further, the power of the random permutation test was increased in conditions with higher group sizes, more response categories, and symmetric response distributions. As would be expected based on the better error control, the random permutation test showed slightly less power than M*plus* DIFFTEST and `lavTestLRT`. The modification to the design of Sass et al. (2014) in simulation one did result in a better performance of the M*plus* DIFFTEST procedure. When the latent variable mean and variance were freely estimated in the focal group in the scalar invariance model, Type I error rates for the DIFFTEST procedure were closer to $\alpha = 0.05$ than what was reported by Sass and colleagues.

Study 2 replicated the poor Type I error control, previously reported by Bandalos (2014), of the mean- and variance-adjusted χ^2 when data were extremely asymmetric. The random permutation test showed no performance issues with Type I error control. These results show that random permutation testing should be considered an appropriate option for researchers to test DIF using item factor analysis models.

The present research suggests the random permutation testing procedure could be preferable over the parametric approaches in nonideal conditions (small to moderate samples with asymmetric thresholds) because permutation provides better control of the Type I error rate for both χ^2 and $\Delta\chi^2$ than the M*plus* DIFFTEST procedure or `lavaan`’s `lavTestLRT`.

References

Asparouhov, T., & Muthén, B. (2006). *Robust chi square difference testing with mean and variance adjusted test statistics*. M*plus* Web Notes No. 10. Retrieved from www.statmodel.com.

Antoniadou, N., Kokkinos, C. M., & Markos, A. (2016). Development construct validation and measurement invariance of the Greek cyber-bullying/victimization experiences questionnaire (CBVEQ-G). *Computers in Human Behavior, 65,* 380–390.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling, 21*(1), 102–116.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*(2), 186–203.

Bentler, P. M., & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods, 15*(2), 111–123.

Jorgensen, T. D., Kite B. A., Chen P.-Y., & Short S. D. (2017). Finally! A valid test of configural invariance using permutation in multigroup CFA. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology: The 81st annual meeting of the psychometric society, Asheville, North Carolina, 2016* (pp. 93–103). New York, NY: Springer. https://doi.org/10.1007/978-3-319-56294-0_9.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41,* 71–90.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373.

Rosseel, Y. (2012). `lavaan`: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling, 21*(2), 167–180.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233–247). London, England: Kluwer Academic Publishers.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling, 21*(1), 149–160.

Using Credible Intervals to Detect Differential Item Functioning in IRT Models

Ya-Hui Su, Joyce Chang and Henghsiu Tsai

Abstract Differential item functioning (DIF) occurs when individuals from different groups with the same level of ability have different probabilities of answering an item correctly. In this paper, we develop a Bayesian approach to detect DIF based on the credible intervals within the framework of item response theory models. Our method performed well for both uniform and non-uniform DIF conditions in the two-parameter logistic model. The efficacy of the proposed approach is demonstrated through simulation studies and a real data application.

Keywords Credible interval · DIF · Item response model · Markov chain Monte Carlo

1 Introduction

The unidimensional item response theory (IRT) models are statistical models that describe the relationship among a latent trait (intelligence, ability, attitude, etc.), the properties of items, and how respondents answer individual items. Like other statistical models, checking the validity of these models is necessary for the applicability and the success of interpretation. Differential item functioning (DIF) refers to a strong violation of the assumptions in IRT models. More specifically, DIF occurs when individuals from different groups with the same level of ability have different

Y.-H. Su
Department of Psychology, National Chung Cheng University, 168 University Road,
Section 1, Min-Hsiung, Chia-Yi 62102, Taiwan
e-mail: psyyhs@ccu.edu.tw

J. Chang
Department of Economics, University of Texas at Austin, 2225 Speedway,
Austin, TX 78712, USA
e-mail: joyce.chang@utexas.edu

H. Tsai (✉)
Institute of Statistical Science, Academia Sinica, 128 Academia Road,
Section 2, Nankang District, Taipei 11529, Taiwan
e-mail: htsai@stat.sinica.edu.tw

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_25

probabilities of answering an item correctly. Studies of DIF deal with the question of how item scores are affected by external variables that do not belong to the construct to be measured (Glas 1998). Therefore it is important to know which items in a test are subject to DIF.

Many DIF detection methods have been proposed in the literature, including techniques based on the Mantel-Haenszel statistic (Holland and Thayer 1988; Camilli and Penfield 1997; Li 2015), the log-linear models (Kok et al. 1985; Dancer et al. 1994), the IRT models (Hambleton and Rogers 1989; Wang and Woods 2017), and the log-linear IRT models (Kelderman 1989). See Glas (1998) for further discussions. Glas (1998) used the Lagrange multiplier test to evaluate DIF within the framework of several IRT models, including the Rasch model, the one-parameter logistic (1PL), and the two-parameter logistic (2PL) models.

In terms of statistical inference, there are two major approaches: frequentist inference and Bayesian inference. Using the approach of frequentist inference, hypothesis testing and confidence intervals play important roles, and conclusions are drawn based on the frequency or proportion of the observed data. A confidence interval (CI) is a type of interval estimate (of a population parameter) that is computed from the observed data. Confidence intervals (CIs) can be used as a significance test. The simple rule is that if the 95% CI does not include the null value, the null hypothesis is rejected at 0.05 level (e.g., Dahiru 2008, p. 25).

Using the approach of Bayesian inference, a credible interval is an interval in the domain of a posterior probability distribution or a predictive distribution, and is used for interval estimation. See Sect. 7.3 of Garthwaite et al. (2002) for further discussion. So, similar to the frequentist approach, if one uses a Bayesian approach, the null hypothesis is rejected at 0.05 level if the 95% credible interval does not include the null value. Riley and Carle (2012) used 95% credible intervals to assess differences in how respondents answer items administered by computerized adaptive testing versus paper-and-pencil. Nevertheless, their study only focused on uniform DIF without considering non-uniform DIF, and was limited to a small number of replications per experimental condition.

Our goal of this study is to adopt a Bayesian approach to evaluate DIF within the framework of IRT models by using credible intervals. In this paper, we obtained 95% credible intervals to analyze both uniform and non-uniform DIF in the context of 2PL models. The rest of the article is organized as follows. Section 2 introduces our method to detect DIF within the framework of 2PL models. Section 3 describes simulations to investigate the performance of the proposed method in finite samples. Section 4 applies the proposed analysis to the data of the physics examination of the 2010 Department Required Test in Taiwan, and Sect. 5 provides some concluding remarks.

2 Detecting Differential Item Functioning in Two-Parameter Logistic Item Response Model

Let Y_{pj} be the dichotomous response of examinee p on item j, where $p = 1, 2, ..., P$, and $J = 1, 2, ..., J$. Denote b_j and a_j as the location and scale parameters respectively, for item j, and θ_p as the ability parameter for examinee p. In the 2PL model (Birnbaum 1968), the probability of examinee p getting a correct response on item j is given by

$$\pi_{pj} = \Pr(Y_{pj} = 1|\theta_p, a_j, b_j) = \frac{1}{1 + e^{-a_j\theta_p + b_j}}. \tag{1}$$

The parameter a_j is also known as the discrimination parameter (de Ayala 2009), or the slope parameter (Wang 2004), and the parameter b_j is called the difficulty parameter in Embretson and Reise (2000) and Wang and Xu (2015). For more descriptions and discussions of the 2PL model, see Embretson and Reise (2000), Wang (2004), and de Ayala (2009).

An item is said to exhibit DIF if the probability of correctly answering the item differs across separate subgroups after controlling for the underlying ability. Specifically, consider the simplest case of two groups, namely the reference and focal group, and use $g_p = 0$ and $g_p = 1$ to indicate whether the examinee p belongs to the reference group or the focal group. Furthermore, each group has its own difficulty and discrimination parameters. Then, Eq. (1) becomes

$$\pi_{pj} = \Pr(Y_{pj} = 1|g_p, \theta_p, a_j, b_j, c_j, d_j) = \begin{cases} \frac{1}{1+e^{-a_j\theta_p+b_j}}, & g_p = 0, \\ \frac{1}{1+e^{-c_j\theta_p+d_j}}, & g_p = 1, \end{cases} \tag{2}$$

where a_j and c_j are the discrimination parameters and b_j and d_j are the difficulty parameters for the reference and the focal group, respectively. Alternatively, we can adopt the notations of Glas (1998) to write Eq. (2) as

$$\pi_{pj} = \Pr(Y_{pj} = 1|g_p, \theta_p, a_j, b_j, \gamma_j, \delta_j) = \begin{cases} \frac{1}{1+e^{-a_j\theta_p+b_j}}, & g_p = 0, \\ \frac{1}{1+e^{-(a_j+\gamma_j)\theta_p+b_j+\delta_j}}, & g_p = 1. \end{cases} \tag{3}$$

Equation (3) implies that the responses of the reference group are properly described by (1), but that the responses of the focal group need additional difficulty parameters δ_j, additional discrimination parameters γ_j, or both. Therefore, we consider the following two hypotheses:

$$H_{\gamma_j,0} : \gamma_j = 0 \text{ versus } H_{\gamma_j,1} : \gamma_j \neq 0,$$
$$H_{\delta_j,0} : \delta_j = 0 \text{ versus } H_{\delta_j,1} : \delta_j \neq 0.$$

Due to the complexity of the likelihood function, a Bayesian estimation method is often used. Specifically, we follow closely the Bayesian approaches of Chang et al. (2014, 2016). For model identification purpose, the marginal distribution of θ_p is set to be the standard normal.

The procedure for testing the hypotheses runs as follows. Suppose there are J items in the test. For each item, we test $\gamma_j = 0$ and $\delta_j = 0$ separately, and only focus on one item at a time. Let η_j be either γ_j or δ_j. If $\eta_j = \gamma_j$, then $\tilde{\eta}_j = \delta_j$, and vice versa (if $\eta_j = \delta_j$, then $\tilde{\eta}_j = \gamma_j$). Then, a size α test of $\eta_j = 0$ is constructed as follows. First, let item j follow Eq. (3) and set $\tilde{\eta}_j = 0$, whereas the other items follow Eq. (1). In other words, we only focus on testing, if for item j, the responses of the focus group need an additional parameter η_j. Then, we implement the Bayesian analysis via the Markov chain Monte Carlo (MCMC) scheme to construct the equal-tailed $1 - \alpha$ credible interval for the parameter η_j. If the interval includes 0, then we do not reject $\eta_j = 0$. Otherwise, $\eta_j = 0$ is rejected.

3 Simulation Study

In this section, we describe the simulation studies to evaluate the performance of our tests. We fixed the Type-I error of each test (α) to 0.05. All computations were performed using Fortran code with IMSL subroutines. For each p, g_p is randomly assigned to be 0 or 1 with a probability of .50. In each experiment, we simulate a test consisting of 10 items, i.e., $J = 10$. The number of examinees (P) are 200 and 400 students. For the true values of a_j and b_j, for $j = 1, ..., J$, we fit the data of the 26 items of the physics examination (see Sect. 4) to the 2PL model defined in Eq. (1), and use the fitted values of the a_j and the b_j of the first 10 multiple-choice items to be the true values. Regarding the values of γ_j and δ_j, we consider two cases (see Table 1). The first case is that there is only one item with $\gamma_j \neq 0$ or $\delta_j \neq 0$, but not both. The second case is that there are three items of $\gamma_j = 1$ or $\delta_j = 1$, or both. The results are summarized in Table 2.

To construct the credible intervals, we produce 11,000 MCMC draws with the first 1,000 draws as burn-in. For each experiment and each item, we repeat the exercise 1,000 times to create 1,000 credible intervals to get the empirical probability of detecting the DIF. In Table 2, p_η^P is used to denote the probability of rejecting the hypothesis $\eta_j = 0$ for the value of P. Again, η denotes either γ or δ. When a test is used to test $\eta_j = 0$, the probabilities of rejecting the hypothesis $\eta_j = 0$ when it is true and when it is not true are the so-called Type-I error and the power of the test, respectively. In Table 2, the numbers with and without parentheses correspond to power and type-I error, respectively.

As shown in Table 2, it is clear that for DIF items the power increases with the value of P. For non-DIF items the Type-I errors are on average close to the nominal size, although some of them are as large as 0.131 (p_δ^{400} of item 9 for the case of one DIF item) and as small as 0.009 (p_γ^{200} of item 8 for the case of 3 DIF items).

Table 1 Overview of the experiments

Nr. of DIF items	Condition	Test	P
One	1	$\gamma_j = 0$	200, 400
	2	$\delta_j = 0$	200, 400
Three	3	$\gamma_j = 0$; $\delta_j = 0$	200, 400

Table 2 Empirical probabilities of rejecting $\gamma_j = 0$ and those of $\delta_j = 0$

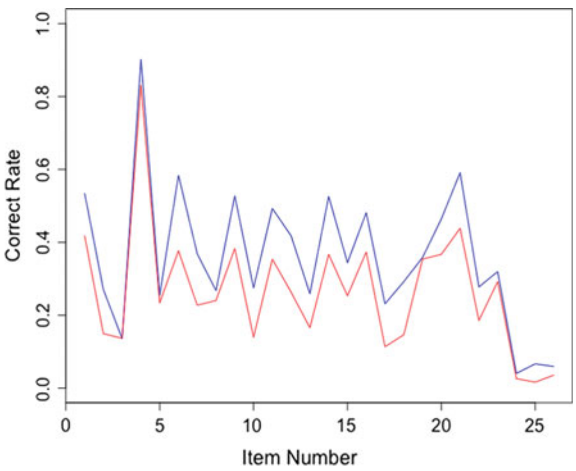
True values			$\gamma_1 = 1$		$\delta_1 = 1$		$\gamma_1 = 1$; $\delta_2 = 1$; $\gamma_3 = 1$ and $\delta_3 = 1$			
Item	a_j	b_j	p_γ^{200}	p_γ^{400}	p_δ^{200}	p_δ^{400}	p_γ^{200}	p_γ^{400}	p_δ^{200}	p_δ^{400}
1	1.195	−0.001	(0.152)	(0.317)	(0.772)	(0.979)	(0.244)	(0.347)	0.117	0.083
2	1.242	1.524	0.042	0.052	0.056	0.058	0.045	0.051	(0.680)	(0.926)
3	0.544	1.955	0.034	0.038	0.058	0.053	(0.121)	(0.222)	(0.149)	(0.249)
4	0.778	−2.195	0.045	0.049	0.103	0.080	0.026	0.048	0.094	0.077
5	0.803	1.254	0.046	0.056	0.039	0.062	0.039	0.051	0.040	0.053
6	0.841	−0.094	0.055	0.053	0.068	0.062	0.053	0.049	0.065	0.067
7	1.011	0.877	0.046	0.056	0.063	0.075	0.053	0.048	0.058	0.068
8	0.082	1.054	0.070	0.012	0.046	0.060	0.009	0.014	0.048	0.061
9	1.444	0.084	0.042	0.052	0.097	0.131	0.060	0.049	0.077	0.105
10	1.934	1.879	0.055	0.080	0.049	0.057	0.023	0.059	0.047	0.043

Moreover, the Type-I error and the power of the test of $\gamma_j = 0$ do not differ much for one or three DIF items. For the test of $\delta_j = 0$, the Type-I error does not change much for one or three DIF items, whereas the power deteriorates from one to three DIF items. It is also interesting to note that the power of detecting DIF on the difficulty parameter is much larger than that on the discrimination parameter.

4 Application

In this section, the proposed procedure described in the previous sections are applied to the data of the physics examination of the 2010 Department Required Test for college entrance in Taiwan provided by the College Entrance Examination Center (CEEC). Examinees have to answer 26 questions in 80 min. The 26 questions are further divided into three parts. The totel score is 100, and the test is administered under formula-scoring directions. For the first part, there are 20 multiple-choice questions, and the examinees have to choose one correct answer out of 5 possible choices. For each correct answer, 3 points are granted, and 3/4 point is deducted from the raw score for each incorrect answer. The second part consists of 4 multiple-response questions, and each question consists of 5 choices, examinees need to select all the answer choices that apply. The choices in each item are knowledge-related, but are

Fig. 1 Plots of the correct rates for male (blue line) and female (red line) for all items in the test



answered and graded separately. For each correct choice, 1 point is earned, and for each incorrect choice 1 point is deducted from the raw score. The adjusted score would only be 0 or above for each of these two parts. The last part consists of 2 calculation problems, and deserves 20 points in total.

The data from 1,000 randomly sampled examinees contains the original responses and nonresponses information, but we treat both nonresponses and incorrect answers the same way and code them as $Y_{pj} = 0$ as suggested by Chang et al. (2014). As for the calculation part, the response Y_{pj} is coded as 1 whenever the original score is more than 7.5 out of 10 points, and zero otherwise (see also Chang et al. 2014). Chang et al. (2016) showed that the 2PL model fits the data well. Here, we consider male as the reference group, and female as the focal group and among the 1,000 examinees, 692 of them are male and 308 are female.

We make more MCMC draws than in Sect. 3. Specifically, we produce 40,000 MCMC draws with the first 10,000 draws as burn-in. Then we test $\gamma_j = 0$ and $\delta_j = 0$, for $j = 1, ..., 26$. Again, we consider $\alpha = 0.05$. The results show that for Item 6, the discrimination and the difficulty parameters are both subject to DIF, whereas for Item 24, only the discrimination parameter is subject to DIF, and for items 7, 17, 18, and 21, only the difficulty parameter is subject to DIF. To further study the testing results, we first note that for each item, and for each examinee, the score can either be 0 or 1. Therefore, for each item, we define the percent of correct rate of each gender to be the percent of scoring 1. The results are summarized in Fig. 1. It is interesting to note that the correct rates for the male are all higher than those for the female, except for items 3 and 19. For these two items, they are almost identical.

Then, we plot the credible intervals for the γ and the δ parameters in Fig. 2. In this figure, the dot in the middle of each interval represents the median of the posterior distribution based on the MCMC draws after burn-in. For the two items the discrimination parameter is subject to DIF: for Item 6, the discrimination parameter is higher for females than for males; for Item 24, the opposite holds.

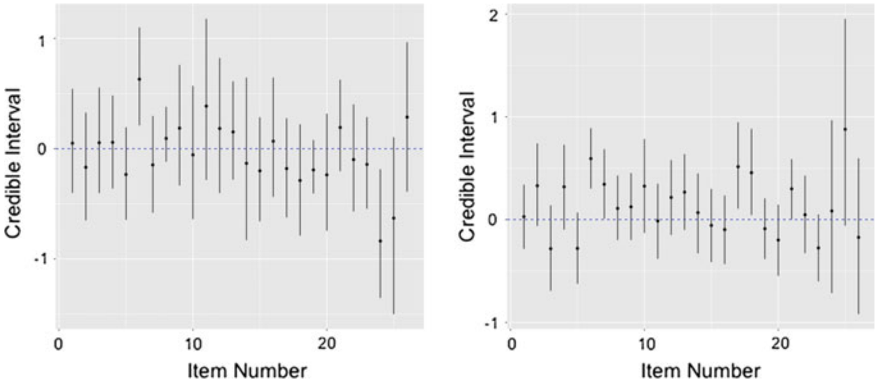


Fig. 2 Plots of the credible intervals for all items for γ_j (left figure) and δ_j (right figure)

From Fig. 1, we know that Item 6 is a relatively easy item and Item 24 is a relatively difficult item. For the 5 items that the difficulty parameter are subject to DIF, it is always that the parameter for the female is higher than that for the male. The results are consistent with Fig. 1.

5 Concluding Remarks

In this article, we propose to use credible intervals to detect DIF in 2PL models. Simulation studies show that the proposed method works reasonably well for detecting the need of an additional difficulty parameter or an discrimination parameter for the responses of the focus group. Applications of the proposed method to other IRT models will be an interesting future line of research. It will also be worthwhile to compare the power of our test with others in the future.

Acknowledgements The research was supported by Academia Sinica and the Ministry of Science and Technology of the Republic of China under grant number MOST 104-2118-M-001-008-MY2. The authors would like to thank Rianne Janssen, the Co-Editor, Dr. Yu-Wei Chang and Ms. Yi-Jhen Wu for their helpful comments and suggestions.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based ore Mantel-Haenszel statistics. *Journal of Educational Measurement*, *34*, 123–139.

Chang, Y.-W., Tsai, R.-C., & Hsu, N.-J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika*, *79*, 255–274.

Chang, J., Tsai, H., Su, Y.-H., & Lin, E. M. H. (2016). A three-parameter speeded item response model: Estimation and application. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (Vol. 167) (pp. 27–38). Switzerland: Springer.

Dahiru, T. (2008). P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan Postgraduate Medicine*, *6*, 21–26.

Dancer, L. S., Anderson, A. J., & Derlin, R. L. (1994). Use of log-linear models for assessing differential item functioning in a measure of psychological functioning. *Journal of Consulting and Clinical Psychology*, *62*, 710–717.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: L. Erlbaum Associates.

Garthwaite, P., Jolliffe, I., & Jones, B. (2002). *Statistical inference*. Oxford: Oxford University Press.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *8*, 647–667.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*, 313–334.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika*, *54*, 681–697.

Kok, F. G., Mellenbergh, G. J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement*, *22*, 295–303.

Li, Z. (2015). A power formula for the Mantel-Haenszel test for differential item functioning. *Applied Psychological Measurement*, *39*, 373–388.

Riley, B. B., & Carle, A. C. (2012). Comparison of two Bayesian methods to detect mode effects between paper-based and computerized adaptive assessments: A preliminary monte carlo study. *BMC Medical Research Methodology*, *12*, 124.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*, 456–477.

Wang, M., & Woods, C. M. (2017). Anchor selection using the Wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, *41*, 17–29.

Wang, W.-C. (2004). Rasch measurement theory and application in education and psychology. *Journal of Education and Psychology*, *27*, 637–694. (in Chinese).

Bayesian Network for Modeling Uncertainty in Attribute Hierarchy

Lihong Song, Wenyi Wang, Haiqi Dai and Shuliang Ding

Abstract In the attribute hierarchy method, cognitive attributes are assumed to be organized hierarchically. Content specialists usually conduct a task analysis on a sample of items to specify the cognitive attributes required by the correctly answered items, and to order these attributes to create an attribute hierarchy. However, the problem-solving performance of experts and novices was almost certain to be different. Additionally, experts' knowledge is highly organized in deeply integrated schemas, while a novice views domain knowledge and problem-solving knowledge separately. Thus, this may bring uncertainty into the attribute hierarchy and lead to different attribute hierarchies for a test. Formally, a Bayesian network is a probabilistic graphical model that represents a set of random latent attributes or variables and their conditional dependencies via a directed acyclic graph. For example, a Bayesian network can be used to represent the probabilistic relationships between latent attributes in the attribute hierarchy. The purpose of this study is to apply Bayesian network for modeling uncertainty in an attribute hierarchy. The Bayesian network created from the attribute hierarchy, which is regarded as a flexible high-order model, is incorporated into three cognitive diagnostic models. The new model has an advantage of taking an account of subjectivity of the attribute hierarchy specified by experts with the uncertainty of

L. Song
Elementary Education College, Jiangxi Normal University,
99 Ziyang Road, Nanchang, Jiangxi, People's Republic of China
e-mail: viviansong1981@163.com

W. Wang (✉) · S. Ding
School of Computer and Information Engineering, Jiangxi Normal University,
99 Ziyang Road, Nanchang, Jiangxi, People's Republic of China
e-mail: wenyiwang@jxnu.edu.cn

S. Ding
e-mail: ding06026@163.com

H. Dai
School of Psychology, Jiangxi Normal University, 99 Ziyang Road,
Nanchang, Jiangxi, People's Republic of China
e-mail: daihaiqi@aliyun.com

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_26

item responses. Fraction subtraction data were analyzed to evaluate the performance of the new model.

Keywords Cognitive diagnosis · Attribute hierarchy · Bayesian network
The reduced reparameterized unified model

1 Introduction

More and more researchers are interested in combining psychometrics and cognitive science to a new psychometric area. In an educational assessment, cognitive diagnostic assessment (CDA) that combines psychometrics and cognitive science has received increased attention recently (Leighton and Gierl 2007; Nichols et al. 1995; Rupp et al. 2010; Tatsuoka 2009). This approach potentially provides useful diagnostic information regarding students' strengths and weaknesses, and can facilitate individualized learning (Chang 2015; Chang and Wang 2016). However, how to incorporate these two fields into all aspects of the development of CDA calls for more study to explore.

In the attribute hierarchy method (AHM), cognitive attributes are assumed to be organized hierarchically (Leighton et al. 2004). Content specialists usually conduct a task analysis on a sample of items to specify the cognitive attributes required by the correctly answered items, and to order these attributes to create an attribute hierarchy. However, the problem-solving performance of experts and novices was almost certain to be different. Additionally, experts' knowledge is highly organized in deeply integrated schemas, while a novice views domain knowledge and problem-solving knowledge separately. Thus, this may bring uncertainty into the attribute hierarchy and lead to different attribute hierarchies for a test (Wang and Gierl 2011).

Formally, a Bayesian network is a probabilistic graphical model that represents a set of random latent attributes or variables and their conditional dependencies via a directed acyclic graph. For example, a Bayesian network can be used to represent the probabilistic relationships between latent attributes in the attribute hierarchy. Moreover, mixing the Bayesian network proficiency model with the fusion evidence model would produce a very attractive class of models (Yan et al. 2004).

The purpose of this study is to apply Bayesian network for modeling uncertainty in an attribute hierarchy. The Bayesian network created from the attribute hierarchy, which is regarded as a flexible high-order model, is incorporated into three cognitive diagnostic models, including the deterministic-inputs, noisy ‘‘and’’ gate (DINA) model (Haertel 1989; Junker and Sijtsma 2001), the revised DINA (rDINA) model (Song et al. 2012), and the reduced reparameterized unified model (rRUM; Hartz 2002). The new model has an advantage of taking an account of subjectivity of the attribute hierarchy specified by experts along with the uncertainty of item responses. Fraction subtraction data were analyzed to evaluate the performance of the new model.

2 Method

2.1 Cognitive Diagnostic Models

Let X_{ij} be the response of examinee i to item j, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, J$. Let $\boldsymbol{\alpha}_\mathbf{i}$ be examinee i attribute pattern. Let $\boldsymbol{\beta}$ be a vector of item parameters. Cognitive diagnostic models often utilize a Q-matrix (Embretson 1984; Tatsuoka 1990, 1995, 2009). The entries of a Q-matrix are 1 or 0, in which $q_{jk} = 1$ means that attribute k is involved in correctly answering item j, otherwise, $q_{jk} = 0$. Let q_j be the q-vector of item j in the Q-matrix.

The item response function for the DINA model is as follows

$$P_j(\boldsymbol{\alpha}_\mathbf{i}) = P(X_{ij} = 1|\boldsymbol{\alpha}_\mathbf{i}, \boldsymbol{\beta}_j) = g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}}, \tag{1}$$

where $\boldsymbol{\beta}_j = (s_j, g_j)$, $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ is an ideal latent response, and s_j and g_j are the slipping and guessing parameters of item j.

The item response function for the rRUM is as follows

$$P_j(\boldsymbol{\alpha}_\mathbf{i}) = P(X_{ij} = 1|\boldsymbol{\alpha}_\mathbf{i}, \boldsymbol{\beta}_j) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{ik})q_{jk}}, \tag{2}$$

where $\boldsymbol{\beta}_j = (\pi_j^*, \mathbf{r}_j^*)$, the baseline parameter π_j^* is the probability of correct response to item j given that an examinee has mastered all the required attributes for the item, and the probability of correct response to item j is proportional to the penalty parameters r_{jk}^* when an examinee has not mastered attribute k.

The item response function for the rDINA model is as follows

$$P_j(\boldsymbol{\alpha}_\mathbf{i}) = P(X_{ij} = 1|\boldsymbol{\alpha}_\mathbf{i}, \boldsymbol{\beta}_j) = (1-s_j)^{w_{ij}} g_j^{1-w_{ij}}, \tag{3}$$

where $w_{ij} = \boldsymbol{\alpha}_i'\mathbf{q}_j/\mathbf{q}_j'\mathbf{q}_j$ is a latent response variable, and $\boldsymbol{\beta}_j = (s_j, g_j)$. As in the DINA model, s_j and g_j are the slipping and guessing parameters of item j. The latent response variable describes the proportion of attribute mastery of the examinee i on item j. It can be 0, 1, or a fraction between 0 and 1. For example, for an item j with $q_j = (1, 1)$, if $\boldsymbol{\alpha}_\mathbf{i} = (0, 1)$ or $\boldsymbol{\alpha}_\mathbf{i} = (1, 0)$, $w_{ij} = 0.5$; if $\boldsymbol{\alpha}_\mathbf{i} = (0, 0)$, $w_{ij} = 0$, otherwise, $w_{ij} = 1$.

The rDINA model relaxes the DINA model assumption of equal probabilities of success for examinees lacking some attributes for an item. In fact, it assumes that if an examinee has not mastered all the required skills for an item, the probability of success varies depends on how many required attributes have been mastered. This is to say that a high probability of success at the item level will occur so long as the examinee has been mastered a larger number of the required skills. The rDINA model, which is similar to the DINA model, is a parsimonious model. It can also be considered as an alternative simple model of the rRUM in some situations.

2.2 Bayesian Network for Cognitive Diagnosis

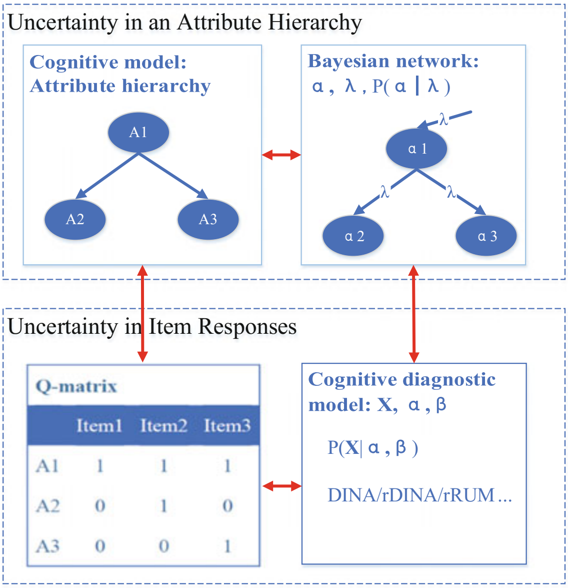
In CDA, a critical issue is model determination. An attribute hierarchy can be viewed as a representation of a cognitive model of task performance (Leighton and Gierl 2007). The attribute hierarchy is designed to describe relationships among the attributes required to solve a set of test items. In addition, a psychometric model is needed to describe relationships between examinees' attribute patterns and item responses. There are three categories of psychometrical models for diagnosis.

(a) One is the high-order DINA model (de la Torre and Douglas 2004). It has a general ability, and considers the relationship between the general ability and each attribute. However, it does not directly take the relationship between latent attributes into account.
(b) The AHM is another widely used model (Gierl et al. 2010; Wang and Gierl 2011). It has given a framework for incorporating the logic hierarchy of attributes (Leighton et al. 2004) and provided a way for knowledge representation.
(c) The Yan's model employed a BN model to consider the probabilistic relationship of attributes and considered the logical structure, in which attribute 3 is a prerequisite to attribute 4 (Yan et al. 2004). However, the Yan's model did not consider the uncertainty of the attribute hierarchy.

The AHM and the Yan's model only considered the logical hierarchy, and they have not consider the uncertainty in the attribute hierarchy. We focus on this question in this study. Next, we build a Bayesian network model, which combines an attribute hierarchy with a psychometric model. Figure 1 displays a framework for modeling the uncertainty in an attribute hierarchy and item responses. In this framework, considering the attribute hierarchy as a directed acyclic graph, we created a BN model for specifying a joint distribution of attributes, denoted by $P(\boldsymbol{\alpha}|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is probabilistic parameters and $P(\boldsymbol{\lambda})$ is a prior distribution of $\boldsymbol{\lambda}$ in the BN model. In the BN model, we employ a directed acyclic graph to model an attributes hierarchy and use the BN parameters to describe the quantitative relationship between attributes.

For example, we assume an attribute hierarchy that contains three attributes. In the attribute hierarchy, attribute A1 is considered to be a prerequisite to attributes A2 and A3. Here, parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_{20}, \lambda_{21}, \lambda_{30}, \lambda_{31})$ are added to the model to describe the quantitative relationship between these three attributes. For attribute A1, λ_1 is the probability of mastering attribute A1, and $1 - \lambda_1$ is the probability of not mastering attribute A1. λ_{20} is the probability of mastering attribute A2 given that attribute A1 is not mastered, while λ_{21} is the probability of mastering attribute A2 when attribute A1 is mastered. Similarly, λ_{30} is the probability of mastering attribute A2 when attribute A1 is not mastered, and λ_{31} is the probability of mastering attribute A2 when attribute A1 is mastered.

The BN model can be used to provide a probabilistic relationship between attributes, which can be combined into different cognitive diagnostic models. As to the uncertainty of item response, we integrated the BN model to psychometric



	Item1	Item2	Item3
A1	1	1	1
A2	0	1	0
A3	0	0	1

Fig. 1 The framework for modeling the uncertainty of an attribute hierarchy and item responses

models, such as the DINA model, the rDINA model, and the rRUM. This process makes the BN model more flexible, and can help with model determination.

2.3 *Estimation*

Using the condition independence of $\mathbf{X}$ given $\boldsymbol{\alpha}$, the joint posterior distribution of parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\lambda}$ given $\mathbf{X}$ is as follows

$$\mathrm{P}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{X}) \propto \mathrm{L}(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\beta})\mathrm{p}(\boldsymbol{\alpha}|\boldsymbol{\lambda})\mathrm{p}(\boldsymbol{\lambda})\mathrm{p}(\boldsymbol{\beta}), \tag{4}$$

where $\mathrm{L}(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\beta})$, $\mathrm{p}(\boldsymbol{\alpha}|\boldsymbol{\lambda})$, $\mathrm{p}(\boldsymbol{\lambda})$, and $\mathrm{p}(\boldsymbol{\beta})$ are respectively the likelihood function $\mathrm{L}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ based on formula (1), (2), or (3), the joint distribution of attributes, the prior distribution of $\boldsymbol{\lambda}$, and the prior distribution of $\boldsymbol{\beta}$. For using Metropolis-Hastings within Gibbs sampling, the full conditional distributions of the parameters given the data and the rest of parameter are as follows

$$\mathrm{p}(\boldsymbol{\lambda}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{X}) \propto \mathrm{p}(\boldsymbol{\alpha}|\boldsymbol{\lambda})\mathrm{p}(\boldsymbol{\lambda}), \tag{5}$$

$$\mathrm{p}(\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{X}) \propto \mathrm{L}(\boldsymbol{\alpha}, \boldsymbol{\beta})\mathrm{p}(\boldsymbol{\alpha}|\boldsymbol{\lambda}), \tag{6}$$

$$\mathrm{p}(\boldsymbol{\beta}|\boldsymbol{\alpha},\boldsymbol{\lambda},\mathbf{X}) \propto \mathrm{L}(\boldsymbol{\alpha},\boldsymbol{\beta})\mathrm{p}(\boldsymbol{\beta}). \tag{7}$$

We estimate or simulate observations $\boldsymbol{\lambda}, \boldsymbol{\alpha},$ and $\boldsymbol{\beta}$ from the Gibbs sampler by repeatedly drawing from the full conditional distributions at iteration t using the following steps:

Step 1: Estimate the parameter $\boldsymbol{\lambda}^{(t)}$ at iteration *t*. Given a set of attributes or nodes $\mathrm{A} = \{\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_\mathrm{K}\}$, let $\mathrm{r_k}$ be the number of levels of node $\mathrm{A_k}$. Let $\pi(\mathrm{A_k})$ be the parent nodes of node $\mathrm{A_k}$ and let q_k be the number of levels of $\pi(\mathrm{A_k})$. Let $\lambda_{\mathrm{ijk}} = \mathrm{P}(\mathrm{A_i} = \mathrm{k}|\pi(\mathrm{A_i}) = \mathrm{j})$ be a conditional probability such that $\sum_{\mathrm{k}=1}^{\mathrm{r_k}} \lambda_{\mathrm{ijk}} = \sum_{\mathrm{k}=1}^{\mathrm{r_k}} \mathrm{P}(\mathrm{A_i} = \mathrm{k}|\pi(\mathrm{A_i}) = \mathrm{j}) = 1$. For sample data $\mathrm{D} = \left[\boldsymbol{\alpha}_1^{(t-1)}, \boldsymbol{\alpha}_2^{(t-1)}, \ldots, \boldsymbol{\alpha}_N^{(t-1)}\right]$, where $\boldsymbol{\alpha}_i^{(t-1)}$ represents a realization or observed value of A, then let

$$m_{ijk}^{(t-1)} = \sum_{1=1}^{\mathrm{N}} \chi\left(\mathrm{i}, \mathrm{j}, \mathrm{k}\colon \boldsymbol{\alpha}_l^{(t-1)}\right), \tag{8}$$

where $\mathrm{m_{ijk}}$ is the frequency of event $\{\mathrm{A_i} = \mathrm{k}, \pi(\mathrm{A_i}) = \mathrm{j}\}$ in sample data. The prior distribution of λ_{ij} follows a Dirichlet distribution, denoted by $\mathrm{Dir}\left[\mathrm{a_{ij1},} \mathrm{a_{ij2}}, \cdots, \mathrm{a_{ijr_k}}\right]$, then the estimate of λ_{ijk} can be written as (Zhang and Guo 2006)

$$\boldsymbol{\lambda}_{ijk}^{(t)} = \frac{m_{ijk}^{(t-1)} + \mathrm{a_{ijk}}}{\sum_{\mathrm{k}=1}^{\mathrm{r_k}} \left(m_{ijk}^{(t-1)} + \mathrm{a_{ijk}}\right)}. \tag{9}$$

Step 2: Draw the parameter $\boldsymbol{\alpha}^{(\mathrm{t})}$ at iteration *t*. Assuming a proposed candidate $\boldsymbol{\alpha}_i^{(t)}$, where the entry $\boldsymbol{\alpha}_{ik}^{(t)}$ draw from Bernoulli (0.5), let $\boldsymbol{\alpha}_i^{(t)} = \boldsymbol{\alpha}_i^{(*)}$ with acceptance probability

$$\mathrm{a}\left(\boldsymbol{\alpha}_i^{(t-1)}, \boldsymbol{\alpha}_i^{(*)}\right) = \min\left\{1, \frac{\mathrm{L}\left(\boldsymbol{\alpha}_i^{(*)}, \boldsymbol{\beta}^{(\mathrm{t}-1)}\right)\mathrm{p}\left(\boldsymbol{\alpha}_i^{(*)}|\boldsymbol{\lambda}^{(\mathrm{t}-1)}\right)}{\mathrm{L}\left(\boldsymbol{\alpha}_i^{(t-1)}, \boldsymbol{\beta}^{(\mathrm{t}-1)}\right)\mathrm{p}\left(\boldsymbol{\alpha}_i^{(t-1)}|\boldsymbol{\lambda}^{(\mathrm{t}-1)}\right)}\right\}, \tag{10}$$

otherwise, $\boldsymbol{\alpha}_i^{(t)} = \boldsymbol{\alpha}_i^{(t-1)}$.

Step 3: Draw the parameter $\boldsymbol{\beta}$ at iteration *t*. Take the DINA model as an example. The proposed candidates $\mathrm{s_j^{(*)}}$ and $\mathrm{g_j^{(*)}}$ draw from $\mathrm{N}\left(\mathrm{s_j^{(t-1)}}, 0.1\right)$ and $\mathrm{N}\left(\mathrm{g_j^{(t-1)}}, 0.1\right)$. Let $\boldsymbol{\beta}_\mathrm{j}^{(\mathrm{t})} = (\mathrm{s_j^{(t)}}, \mathrm{g_j^{(t)}}) = (\mathrm{s_j^{(*)}}, \mathrm{g_j^{(*)}})$ with acceptance probability

$$\mathrm{a}\left[\left(\mathrm{s}_{\mathrm{j}}^{(\mathrm{t}-1)},\mathrm{g}_{\mathrm{j}}^{(\mathrm{t}-1)}\right),(\mathrm{s}_{\mathrm{j}}^{(*)},\mathrm{g}_{\mathrm{j}}^{(*)})\right]=\min\left\{1,\frac{\mathrm{L}\left(\boldsymbol{\alpha}^{(\mathrm{t})},\mathrm{s}_{\mathrm{j}}^{(\mathrm{t})},\mathrm{g}_{\mathrm{j}}^{(\mathrm{t})}\right)\mathrm{p}\left(\mathrm{s}_{\mathrm{j}}^{(\mathrm{t})}\right)\mathrm{P}\left(\mathrm{g}_{\mathrm{j}}^{(\mathrm{t})}\right)}{\mathrm{L}\left(\boldsymbol{\alpha}^{(\mathrm{t})},\mathrm{s}_{\mathrm{j}}^{(\mathrm{t}-1)},\mathrm{g}_{\mathrm{j}}^{(\mathrm{t}-1)}\right)\mathrm{p}\left(\mathrm{s}_{\mathrm{j}}^{(\mathrm{t}-1)}\right)\mathrm{P}\left(\mathrm{g}_{\mathrm{j}}^{(\mathrm{t}-1)}\right)}\right\}, \tag{11}$$

otherwise, $\boldsymbol{\beta}_{\mathrm{j}}^{(\mathrm{t})}=(\mathrm{s}_{\mathrm{j}}^{(\mathrm{t}-1)},\mathrm{g}_{\mathrm{j}}^{(\mathrm{t}-1)})$.

3 Real Data Analysis

3.1 Fraction Subtraction Data Set

The new model was applied to a widely analyzed fraction subtraction data set (de la Torre 2008; DeCarlo 2012; Tatsuoka 2002; Tatsuoka 1990), which consists of 536 examinees. The Q-matrix, which consists of 15 items, is the same as the one used originally by de la Torre (2008) and DeCarlo (2012). The labels of the attributes are (a) performing a basic fraction-subtraction operation, (b) simplifying/reducing, (c) separating whole numbers from fractions, (d) borrowing one from a whole number to a fraction, and (e) converting whole numbers to fractions.

3.2 Attribute Hierarchy and Bayesian Network

In the analysis of fraction subtraction data, two attribute hierarchies are considered, one (called AH1 in Fig. 2) assumes that the attribute A3 is a prerequisite to attribute A4, and the other (called AH2 in Fig. 3) is derived from the Q-matrix through the pairwise comparison method (Tatsuoka 1995). According to the augment algorithm (Ding et al. 2008), two reduced Q-matrices are obtained with 24 or 9 attribute patterns. Two Bayesian networks (called BN1 and BN2 as shown in Figs. 4 and 5) corresponding to the above two attribute hierarchies are constructed based on the idea of the previous study (Yan et al. 2004), and two joint distributions of attributes are specified, respectively. For the BN1, the parameters for the joint distribution of attributes are specified as follows:

$$\begin{aligned}
&\lambda_1=P(\alpha_1=1),\\
&\lambda_{2,m}=P(\alpha_2=1|\alpha_1=m)\quad \text{for } m=0,1,\\
&\lambda_{5,m}=P(\alpha_5=1|\alpha_1+\alpha_2=m)\quad \text{for } m=0,1,2,\\
&\lambda_{34,m,k}=P((\alpha_3,\alpha_4)=\boldsymbol{\gamma}_k|\alpha_1+\alpha_2+\alpha_4=m)\quad \text{for } m=0,1,2,3,
\end{aligned}$$

where $\boldsymbol{\gamma}_1=(0,0),\boldsymbol{\gamma}_2=(1,0),\boldsymbol{\gamma}_3=(1,1)$, and $\sum_{k=1}^{3}\lambda_{34,m,k}=1$. It should be noted that $\lambda_{34,m,k}$ describes the statistical relationship between $\alpha_1+\alpha_2+\alpha_4$ and (α_3,α_4), which is different from the Yan’s model.

For the BN2, the parameters for the joint distribution of attributes are specified as follows:

$$\begin{aligned}
&\lambda_1 = P(\alpha_1 = 1),\\
&\lambda_{2,m} = P(\alpha_2 = 1 | \alpha_1 = m) \quad \text{for } m = 0, 1,\\
&\lambda_{3,m} = P(\alpha_3 = 1 | \alpha_1 = m) \quad \text{for } m = 0, 1,\\
&\lambda_{4,m} = P(\alpha_4 = 1 | \alpha_1 + \alpha_3 = m) \quad \text{for } m = 0, 1, 2,\\
&\lambda_{5,m} = P(\alpha_5 = 1 | \alpha_1 + \alpha_3 + \alpha_4 = m) \quad \text{for } m = 0, 1, 2, 3.
\end{aligned}$$

Besides the four attribute spaces above, an independent attribute space (called AH0) was also considered. The DINA model, the rDINA model, and the rRUM were used to analyze Tatsuoka's fraction subtraction data by using the Markov Chain Monte Carlo (MCMC) algorithm.

3.3 Evaluation Criteria

To compare these models under different attribute spaces and cognitive diagnostic models, two relative fit statistics are considered in this study: −2log-likelihood (−2LL; Chen et al. 2013) and deviance information criterion (DIC_4; Celeux et al. 2006).

Fig. 2 Attribute hierarchy 1 (AH1)

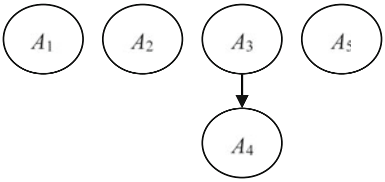
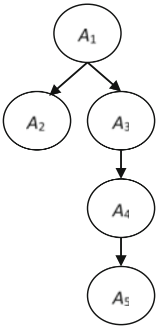


Fig. 3 Attribute hierarchy 2 (AH2)



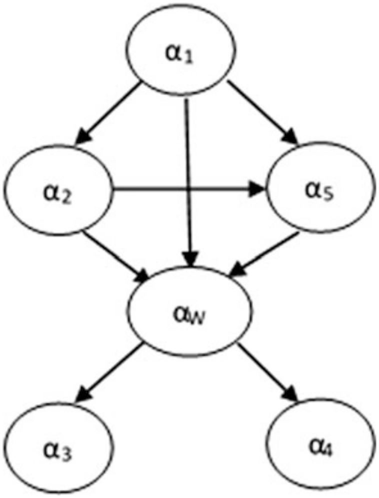


Fig. 4 Bayesian network 1 (BN1)

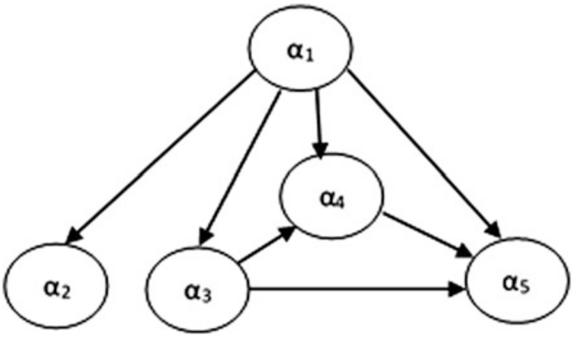


Fig. 5 Bayesian network 2 (BN2)

3.4 Results

Table 1 shows model fit indices across different models. The results indicate that:

(a) The impact of different attribute spaces is very apparent. The BN1 and AH2 with similar results provide better fit than the BN2, AH1, and AH0. The BN2 and AH1 almost provide a better fit than the AH0.

(b) The impact of three cognitive diagnosis models is also apparent. The rRUM model outperforms the other two models. There is an interaction effect between attribute spaces and cognitive diagnosis models.

(c) It is important to note that the impact of different attribute spaces on the rRUM model is relatively smaller than that on the other models.

Table 1 Model fit indices across different models

Criteria	Model	AH0	AH1	AH2	BN1	BN2
−2LL	DINA	7614	7464	7240	6980	7516
	rDINA	7394	7295	7059	6919	7113
	rRUM	6942	6920	6857	6881	7044
DIC_4	DINA	9306	8934	7751	7857	8324
	rDINA	9210	8900	7805	7912	8203
	rRUM	9168	8990	7737	7846	7958

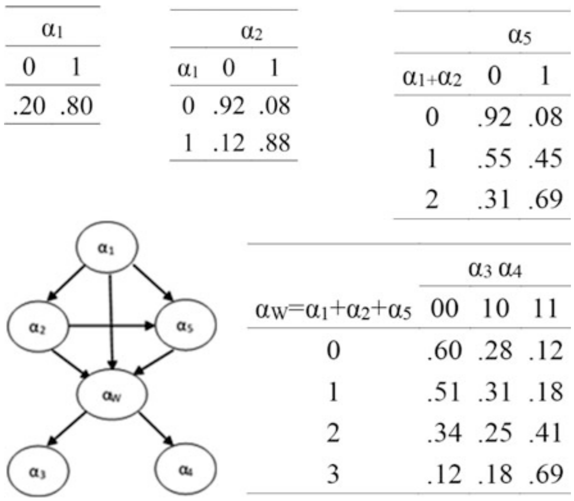
Table 2 The mean of attribute mastery probabilities of examinees with a total score of zero

Attribute hierarchy	Model	A1	A2	A3	A4	A5
AH0	DINA	0.00	0.50	0.35	0.46	0.54
	rDINA	0.00	0.00	0.00	0.00	0.27
	rRUM	0.00	0.27	0.00	0.00	0.54
AH1	DINA	0.00	0.65	0.69	0.46	0.42
	rDINA	0.00	0.00	0.00	0.00	0.15
	rRUM	0.00	0.38	0.00	0.00	0.50

Furthermore, Table 2 shows that the means of attribute mastery probabilities of examinees with a total score of zero. The independent structure of five attributes was also estimated here for comparison. Here we observed that: under the rDINA model, the attribute mastery probabilities are the lowest, and the DINA model shows relatively high attributes mastery probabilities for both structures. For example, for the AH1 under the rDINA model, examinees with a total score of zero are classified as not mastering any of the attributes. While for the AH1 under the DINA model, the means of attribute mastery probabilities for attributes A2 and A3 are greater than 0.50. Because total scores of these examinees are zero, attribute mastery probabilities are theoretically supposed to be zero. Thus, from results of this table, the rDINA and rRUM models perform relatively better than the DINA model.

Figure 6 displays the estimates of the parameters $\boldsymbol{\lambda}$ based on the BN1 and the rRUM. The estimate $\lambda_1 = 0.80$ means that the mastery probability of attribute A1 is 0.8. The estimates $1 - \lambda_{2,0} = 0.92$ and $\lambda_{2,1} = 0.88$ mean that if an examinee has not mastered attribute A1, then the mastery probability of attribute A2 is very low, 0.08; if an examinee has mastered attribute A1, then the mastery probability of attribute A2 is pretty high, 0.88.

Fig. 6 The estimates of the parameters **λ** based on the BN1 and the rRUM



α_1	
0	1
.20	.80

	α_2	
α_1	0	1
0	.92	.08
1	.12	.88

	α_5	
$\alpha_1+\alpha_2$	0	1
0	.92	.08
1	.55	.45
2	.31	.69

	$\alpha_3\ \alpha_4$		
$\alpha_W=\alpha_1+\alpha_2+\alpha_5$	00	10	11
0	.60	.28	.12
1	.51	.31	.18
2	.34	.25	.41
3	.12	.18	.69

4 Conclusion

The study proposed a framework for modeling the uncertainty in both the attribute hierarchy and item responses. Combining the BN model with cognitive diagnostic model, it relaxes restrictions, to some extent, on cognitive models and psychometric models in CDA. The new model is flexible to collect more information about model-data fit. The new model provides a way to assist verifying cognitive models. In conclusion, this study shows that:

(a) Among the five cognitive models, the BN1 fits the fraction subtraction data best.
(b) The rRUM outperforms the other two psychometric models in terms of model-data fit.
(c) The BN parameters, which provides quantitative description on attributes' relationship, can help cognitive model validation.

Some future research directions are also pointed out. One limitation of this study is that the latent structures are fixed in advance and the BN model is only learning parameters. It would be interesting to explore the learning of latent structure from data. More applications deserve to be studied.

Acknowledgements This research is supported by the National Natural Science Foundation of China (Grant No. 31500909, 31360237, and 31160203), the Key Project of National Education Science "Twelfth Five Year Plan" of Ministry of Education of China (Grant No. DHA150285), the National Natural Science Foundation of Jiangxi (Grant No. 20161BAB212044), the Education Science Foundation of Jiangxi (Grant No. 13YB032), the Social Science Foundation of Jiangxi (Grant No. 17JY10), the National Social Science Fund of China (Grant No. 16BYY096), the Humanities and Social Sciences Research Foundation of Ministry of Education of China (Grant No. 12YJA740057), and the Youth Growth Fund and the Doctoral Starting up Foundation of Jiangxi Normal University. The authors would like to thank the editor Steve Culpepper for the valuable comments.

References

Celeux, G., Forbers, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis, 1,* 651–674.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*(1), 1–20.

Chang, H.-H., & Wang, W. Y. (2016). "Internet Plus" measurement and evaluation: A new way for adaptive learning. *Journal of Jiangxi Normal University (Natural Science), 40*(5), 441–455.

Chen, J.-S., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447–468.

Ding, S.-L., Luo, F., Yan, C., Lin, H.-J., & Wang, X.-B. (2008). Complement to Tatsuoka's Q matrix theory. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 417–424). Tokyo: Universal Academy Press.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49* (2), 175–186.

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing, 10*(4), 318–341.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301–321.

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

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

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

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205–237.

Nichols, P., Chipman, S., & Brennan, R. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum.

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

Song, L.-H., Wang, W.-Y., Dai, H., & Ding, S.-L. (2012). The Revised DINA Model Parameter Estimation with EM Algorithm. *International Journal of Digital Content Technology and its Applications, 6*(9), 85–92.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 51,* 337–350.

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

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessments* (pp. 327–359). Hillsdale: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Taylor & Francis Group.

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement, 48*(2), 165–187.

Yan, D. L., Almond, R. G., & Mislevy, R. J. (2004). *A comparison of two models for cognitive diagnosis (ETS RR-04-02)*. Princeton, NJ: Education Testing Service.

Zhang, L., & Guo, H. (2006). *Introduction to Bayesian networks*. Beijing: China Science Publishing & Media Ltd.

A Cognitive Diagnosis Method Based on Mahalanobis Distance

Jianhua Xiong, Fen Luo, Shuliang Ding and Huiqiong Duan

Abstract Cognitive diagnosis methods (CDMs) is very important for cognitive diagnosis, the primary purpose for CDMs is to classify examinees into mutually exclusive categories. Although there exist many CDMs, researchers propose many better new CDMs. Among them, the generalized distance discrimination (GDD) and the Hamming distance discrimination (HDD) receive more and more attention for their advantages of simple and easy to use, high classification accuracy, thus, Mahalanobis distance discrimination (MDD), a generalized CDM is introduced. GDD and HDD are its special cases. Mahalanobis distance (MD) is employed for MDD to calculate the distance between an examinee's observed response pattern (ORP) and all kinds of ideal response pattern (IRP). The Shannon entropy is specified as covariance. According to the principle of minimum distance and designing test blueprint, IRP can be bijection mapped to the state of knowledge. Under dichotomous model, the pattern match ratio and average attribute match ratio are selected as the criteria for evaluating the classification accuracy. The Monte Carlo simulation study shows that the performance of MDD is better than GDD and HDD.

Keywords GDD · HDD · Q-matrix · Shannon entropy · Mahalanobis distance

J. Xiong (✉) · F. Luo · S. Ding
College of Computer Information Engineering, Jiangxi Normal University,
No. 99 Ziyang Road, Nanchang, Jiangxi, People's Republic of China
e-mail: 270281168@qq.com

F. Luo
e-mail: luofen312@163.com

S. Ding
e-mail: ding06026@163.com

H. Duan
School of Foreign Languages, Nanchang Hangkong University,
No. 696 South Fenghe Road, Nanchang, Jiangxi, People's Republic of China
e-mail: englishduan2011@163.com

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_27

1 Introduction

Cognitive diagnosis is the integration of multidisciplinary theory and technology, which can evaluate the individual's cognitive structure and reveal the internal psychological process. The internal mental processing is not directly observed, so it's not easy to measure, diagnose, evaluate. Relevant scholars developed many cognitive diagnosis methods (CDMs) to solve this problem (Tu et al. 2012). According to the recent statistics, there are more than 100 CDMs (Xin et al. 2012). There are some methods attract the researchers' attention: AHM, RSM, DINA (Ding et al. 2012). DINA is a simple and high classification accuracy method, thus it has more research results (de la Torre 2009, 2011; Tu et al. 2010; Zhang et al. 2013). In recent years, a number of implicit cognitive diagnosis models have emerged. For example, under dichotomous model, there are the generalized distance discrimination (GDD, Sun et al. 2011) and the Hamming distance discrimination (HDD, Chiu and Douglas 2013; Luo et al. 2015). Under ploytomous model, there are the generalized distance discrimination based on graded response model (Li et al. 2012; Sun et al. 2013), rule space method built on graded response model (Tian and Xin 2012), a cluster diagnostic method established on grade response items (Kang et al. 2015).

Because there are so many CDMs, the researchers try to integrate some methods. That is to give a general cognitive diagnosis method, and think of a method as its special case. The more abstract a method is, the deeper understanding of its essence is. The GDD and HDD have some advantages, such as, simple and easy to use, high classification accuracy. This paper extracts their essences and proposes a generalized method, that is Mahalanobis distance discrimination (MDD), to adjust the weight matrix of Mahalanobis distance, a new method is presented, the corresponding weight matrix should have better statistical significance and higher classification accuracy. This paper introduces Shannon entropy as the weight matrix of Mahalanobis distance, and discusses its classification performance.

2 Overview of Generalized Distance Discrimination and Hamming Distance Discrimination

In this paper, cognitive diagnosis test include only dichotomous items, the ideal response pattern (IRP) indicates an examinee answering the particular items without slip or guess; the observed response pattern (ORP) indicates the real reaction of an examinee on a set of items, assuming there is no missing data in ORP.

2.1 Generalized Distance Discrimination

Assume N examinees respond to a cognitive diagnostic test with J items, and the total number of IRPs inferred from the Q-matrix is T. The GDD uses the following equation to measure the similarity between ORP and IRP:

$$GD(\mathbf{Y}_i, \mathbf{I}_t) = \sum_{j=1}^{J} GD(Y_{ij}, I_j^{(t)}), \text{and}, GD\left(Y_{ij}, I_j^{(t)}\right) = \left|Y_{ij} - I_j^{(t)}\right| P_j(\theta_i)^{Y_{ij}} Q_j(\theta_i)^{1-Y_{ij}} \tag{1}$$

where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, J$, $t = 1, 2, \ldots, T$, vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$ denotes examinee i's ORP, $\mathbf{I}_t = (I_1^{(t)}, \ldots, I_J^{(t)})$ denotes the tth IRP, the item response for every item of the test is either 0 or 1. $GD(Y_{ij}, I_j^{(t)})$ is the generalized distance for item j, which measures the similarity between the examinee i's ORP and the tth IRP. Where $P_j(\theta_i)$ and $Q_j(\theta_i)$ denote the probability of examinee i getting correct answer and wrong answer on the item j. Under item response theory (IRT), the item characteristic function of two parameters logistic model (2PLM) is adopted to specify $P_j(\theta_i)$ and $Q_j(\theta_i)$ (Qi et al. 2002). $GD(\mathbf{Y}_i, \mathbf{I}_t)$ represents the sum of the generalized distances on all items for examinee i, then choose the corresponding IRP according to the shortest rule, that is $\min\limits_{t=1,\ldots,T}\{GD(\mathbf{Y}_i, \mathbf{I}_t)\}$. Under non-compensatory model, for dichotomous items, if the test blueprint contains the reachability matrix, there exists a bijection mapping between knowledge states and IRP (Ding et al. 2010, 2011). The ORP can be classified into the knowledge state corresponding to this IRP, so that the diagnostic classification can be realized. The GDD method has good performance using simulation data (Sun et al. 2011; Cai et al. 2013; Tu et al. 2013).

2.2 Hamming Distance Discrimination

Luo et al. (2015) surveyed the (Chiu and Douglas 2013) nonparametric cognitive diagnosis method, and put forward the hamming distance discrimination (HDD). HDD uses Hamming distance to define the distance between the examinee of ORP and each IRP. Then classifies the examinees according to the principle of the shortest distance. When there are more than one IRPs with the same minimum Hamming distance for an examinee's ORP, method R and method B are effective auxiliary means. The Hamming distance between examinee i's ORP and t th IRP is defined as

$$HD(\mathbf{Y}_i, \mathbf{I}_t) = \sum_{j=1}^{J} HD\left(Y_{ij}, I_j^{(t)}\right), \text{and}, HD\left(Y_{ij}, I_j^{(t)}\right) = \left|Y_{ij} - I_j^{(t)}\right| \tag{2}$$

$N, J, T, \mathbf{Y}_i$ and $\mathbf{I}_t$ have the same definition as GDD. The test blueprint is also the same as GDD, $HD(Y_{ij}, I_j^{(t)})$ is the Hamming distance for item j, which measures the similarity between the examinee i's ORP and the t th IRP. $HD(\mathbf{Y}_i, \mathbf{I}_t)$ represents the sum of Hamming distance on all J items for examinee i. HDD is a nonparametric CDM, which requires the Q-matrix only. It does not require the estimation of the parameters, so it is simple to operate, and easy to understand. Under the same experimental conditions, it has higher classification accuracy than GDD (Luo et al. 2015).

3 Mahalanobis Distance Discrimination

3.1 The Definition of Mahalanobis Distance

Mahalanobis distance is a weighted distance, which can effectively estimate the similarity between two different samples (Zhang and Fang, 2013). In this study, vector $\mathbf{Y}_i$ denotes examinee i's ORP. Vector $\mathbf{I}_t$ denotes t th IRP, $\mathbf{W}_i$ denotes the weight matrix (diagonal matrix). As long as $\mathbf{W}_i$ is a positive definite matrix, the distance between the ORP and IRP is Mahalanobis distance. The equation is:

$$d^2(\mathbf{Y}_i, \mathbf{I}_t) = (\mathbf{Y}_i - \mathbf{I}_t)^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{I}_t) \tag{3}$$

3.2 Shannon Entropy

Under information theory, entropy is a measure of the uncertainty random event. Suppose an event could have n outcomes, the probability distribution of each result is $X = \{p_1, p_2, \ldots, p_n\}(0 \le p_i \le 1, i = 1, 2, \ldots, n)$ and $\sum_{i=1}^{n} p_i = 1$.

The itself information provided by the ith result is $I_i = -\log p_i$, then the average information on all outcomes of this event is:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{4}$$

where $H(X)$ is Shannon entropy, when the uncertainty of probability distribution X is greater, the value of the corresponding Shannon entropy $H(X)$ is larger; on the contrary, the entropy is smaller. The characteristics are (Wu 2008):

(1) If probability distributions $X = \{\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\}$, then $H(X) = \log(n)$

(2) If probability distributions $X = \{0, \ldots, 0, 1, 0, \ldots, 0\}$, then $H(X) = 0$

(3) For any probability distributions of $X, 0 \leq H(X) \leq \log(n)$.

In Eq. (4), the base of logarithm function can take different values. In this paper, let e be the base of logarithm, that is napierian logarithm, pilot study shows that it has better result than other logarithm.

3.3 Integrated GDD and HDD by Mahalanobis Distance

Let the Eq. (1) in Sect. 2.1 be transformed into the following equation:

$$GD\left(Y_{ij}, I_j^{(t)}\right) = (Y_{ij} - I_j^{(t)})^2 P_j(\theta_i)^{Y_{ij}} Q_j(\theta_i)^{1-Y_{ij}} = (Y_{ij} - I_j^{(t)})^T P_j(\theta_i)^{Y_{ij}} Q_j(\theta_i)^{1-Y_{ij}} (Y_{ij} - I_j^{(t)})$$

GDD can be expressed with Mahalanobis distance, that is:

$$GD(\mathbf{Y}_i, \mathbf{I}_t) = (\mathbf{Y}_i - \mathbf{I}_t)^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{I}_t) \tag{5}$$

The weight matrix in Eq. (5) is defined as follows:

$$\mathbf{W}_i = \begin{pmatrix} P_1(\theta_i)^{Y_{i1}} Q_1(\theta_i)^{1-Y_{i1}} & & 0 \\ & \ddots & \\ 0 & & P_J(\theta_i)^{Y_{iJ}} Q_J(\theta_i)^{1-Y_{iJ}} \end{pmatrix}$$

In the same way, the Eq. (2) in Sect. 2.2 is simply transformed into the following equation:

$$HD\left(Y_{ij}, I_j^{(t)}\right) = (Y_{ij} - I_j^{(t)})^2 = (Y_{ij} - I_j^{(t)})^T (Y_{ij} - I_j^{(t)})$$

The Mahalanobis distance expression of HDD is as shown below:

$$HD(\mathbf{Y}_i, \mathbf{I}_t) = (\mathbf{Y}_i - \mathbf{I}_t)^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{I}_t) \tag{6}$$

The weight matrix in Eq. (6) is an unit matrix, that is $\mathbf{W}_i = \mathbf{E}$. From Eqs. (5) and (6), the essence of the GDD and HDD are Mahalanobis distance. They are just different definition on weight matrix, so Mahalanobis distance discrimination method is a more general cognitive diagnosis method.

3.4 Define the Mahalanobis Distance Between ORP and IRP

From the point of Mahalanobis distance definition, we derive the Mahalanobis distance between ORP and IRP more generally. The Mahalanobis distance between the examinee i's ORP and t th IRP is defined as Eq. (7), and examinee i will be classified into the IRP with minimum Mahalanobis distance.

$$MD(\mathbf{Y}_i, \mathbf{I}_t) = (\mathbf{Y}_i - \mathbf{I}_t)^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{I}_t) \tag{7}$$

$MD(\mathbf{Y}_i, \mathbf{I}_t)$ is the Mahalanobis distance between $\mathbf{Y}_i$ and $\mathbf{I}_t$. The definition of $\mathbf{Y}_i$ and $\mathbf{I}_t$ are consistent with GDD, the Shannon entropy is the diagonal element of the weight matrix (diagonal matrix). There are only two possible outcomes in Shannon entropy for dichotomous items, that is, the probability of correct answer (P) and the probability of error answer $(Q = 1 - P)$. For examinee i and item j, the Shannon entropy using napierian logarithm is expressed as $H(X_{ij}) = -P_j(\theta_i)\ln P_j(\theta_i) - Q_j(\theta_i)\ln Q_j(\theta_i)$. The Shannon entropy of examinee i in all items can be represented by the diagonal matrix $\mathbf{W}_i$:

$$\mathbf{W}_i = \begin{pmatrix} -P_1(\theta_i)\ln P_1(\theta_i) - Q_1(\theta_i)\ln Q_1(\theta_i) & & 0 \\ & \ddots & \\ 0 & & -P_J(\theta_i)\ln P_J(\theta_i) - Q_J(\theta_i)\ln Q_J(\theta_i) \end{pmatrix}$$

For the expression $H(X_{ij}) = -P_j(\theta_i)\ln P_j(\theta_i) - Q_j(\theta_i)\ln Q_j(\theta_i)$, when the value of $P_j(\theta_i)$ is 1 or 0, the value of Shannon entropy is 0, which is the minimum value; when the probability of the correct answer and the wrong answer is equal, that is $P_j(\theta_i) = \frac{1}{2}$, the value of Shannon entropy reaches the maximum; the rest is somewhere in between. Shannon entropy consider not only the proximity degree between ORP and IRP, but also the certainty of ORP. Even if an ORP is close to certain IRP, since the uncertainty of ORP is big, the proximity between them is not reliable, this is different from the Eq. (5). Because the test blueprint is also the same as GDD, IRP can correspond to the state of knowledge, which can achieve the purpose of classification of examinees.

Besides, there are two kinds definition for probability of Shannon Entropy: One is the probability of correct answer based on IRT, which is consistent with the definition of GDD, using $P_j(\theta_i)$ and $Q_j(\theta_i)$ separately representing the probability of examinee i to get correct answer and wrong answer on the item j; the other is the pass rate based on Classical test theory (CTT), this definition is simple to calculate, using P_j and Q_j separately to represent the pass rate and the unpassed rate on the item j.

Compute the Mahalanobis distance between the certain examinee *i*'s ORP and all IRP, find the IRP with the shortest distance. If there exists a bijection mapping between knowledge states and IRP, examinee *i* can be classified to the knowledge state corresponding to IRP.

4 Monte Carlo Simulation Study

This study mainly probes the performance of three CDMs (MDD, GDD and HDD) in different attribute hierarchies (linear, convergent, divergent, unstructured and independent) and slips (0.02, 0.05, 0.10, 0.15 and 0.2), there are 75 experimental conditions. In order to conclude the stability and reduce the experimental error, the simulation number of each experimental condition is 50 times, and the specific experiment design is as follows.

4.1 The Design of Test Q-Matrix

In order to compare MDD method with GDD and HDD method, the experimental conditions are the same as those of Sun et al. (2011) and Luo et al. (2015). The study mainly probes five basic attribute hierarchical structures, they are in sequence: linear, convergence, divergent, unstructured and independent (See Appendix 1), the other more complex attribute hierarchy can be compounded by the five basic hierarchies. Under non-compensatory model, if the test blueprint contains the reachability matrix, there exists a bijection mapping between knowledge states and IRP (Ding et al. 2010, 2011). According to the five hierarchical structures, the typical item assessment patterns under them are obtained, there are 6 items, 7 items, 15 items, 32 items and 64 items respectively. In order to avoid the influence of test length on parameter estimation accuracy, the test length of various hierarchy is roughly the same, the five typical item assessment patterns are repeated 5 times, 5 times, 2 times, 1 time and 1 time in the test, for the independent hierarchy, due to the limitation of the test length, the typical item assessment pattern is sorted by the number of attributes and take the top 30 items. Therefore, in this study, the number of test items for the five attribute hierarchy structures are 30, 35, 30, 32 and 30 respectively (See Appendix 2).

4.2 The Simulation of ORP

In this study, 1000 examinees are adopted under various experimental conditions, and the process of simulating ORP include the following steps:

(1) Calculate the ideal master pattern (IMP) for the five attribute hierarchy (also known as knowledge state) respectively. There are in sequence 7 patterns, 8 patterns, 16 patterns, 33 patterns and 64 patterns.

(2) On the basis of the design of test Q-matrix in Sect. 4.1, the corresponding IRP are fetched through IMP and test Q-matrix. Calculate the total score for each IRP, and sort them by ascending order. Simulate 1000 examinees and distribute them proportionally to each IRP, the IRP with the same test score will assign the same number examinees, then 1000 examinees' IMP will gain under various attribute hierarchy, simulate the examinees' IRP without any slip.

(3) According to the simulated IRP, simulate the examinees' ORP under different slip (such as: 0.02, 0.05, 0.10, 0.15, 0.2) (Leighton et al. 2004).

4.3 Criteria

There are two kinds of criteria to evaluate the discrimination accuracy of different methods, that is, the pattern match ratio (PMR) and average attribute match ratio (AAMR), the equations are as follows:

$$PMR = \frac{\sum_{i=1}^{N} N_{i\_correct}}{N}, \quad AAMR = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} N_{ik\_correct}}{K \times N}$$

where, N is the total number of examinees, K is the number of attributes, $N_{i\_correct}$ represents the agreement between the obtained knowledge state and the known true knowledge state for examinee *i*, if they are agreement entirely, $N_{i\_correct}$ is 1, otherwise 0. $N_{ik\_correct}$ represents the agreement of individual attributes between the obtained knowledge state and the known true knowledge state for examinee *i*, if they are agreement, $N_{ik\_correct}$ is 1, otherwise 0.

4.4 Comparison of Cognitive Diagnostic Methods

In this paper, the differences between MDD, GDD and HDD in PMR and AAMR are compared. For MDD and GDD, the diagonal element of the weight matrix involve the probability *P*, for the acquisition of *P*, there are two methods, one method is based on IRT, combining 1000 examinees' ORP and all IRP to estimate item parameters, using 1000 examinees' ORP to estimate ability parameters, then using 2PLM item characteristic function to calculate the correct answer probability *P* and error answer probability *Q*, another method is based on CTT, which only use 1000 examinees' ORP to calculate the passed rate *P* and the unpassed rate *Q*. In order to probe the effect of different parameter calculation methods on the discriminant result, it is expressed as the following four combinations: MDD-CTT, MDD-IRT, GDD-CTT, and GDD-IRT. In addition, for HDD method, according to the conclusion of Luo et al. (2015) that HDD B method is better than HDD R method, so this research only compares HDD B method (shorthand for HDDB). Therefore the paper compares MDD-CTT, MDD-IRT, GDD-CTT, GDD-IRT and HDDB under various simulation conditions.

5 Results

Tables 1 and 2 show that MDD outperforms GDD and HDDB, while the GDD and HDDB have their own merits. For MDD-CTT and MDD-IRT, their PMR and AAMR are neck and neck. However, using the CTT method to calculate the pass

Table 1 The comparison of each cognitive diagnosis method on pattern match ratio

Attribute hierarchical structure	Slip	Pattern match ratio (PMR)				
		MDD-CTT	MDD-IRT	GDD-CTT	GDD-IRT	HDDB
Linear	0.02	0.9999	0.9999	0.9808	0.9982	0.9999
	0.05	0.9980	0.9980	0.9691	0.9887	0.9980
	0.1	0.9856	0.9857	0.9242	0.9674	0.9852
	0.15	0.9521	0.9534	0.8794	0.9284	0.9502
	0.2	0.8936	0.8959	0.8126	0.8756	0.8869
Convergent	0.02	0.9998	0.9998	0.9808	0.9973	0.9998
	0.05	0.9980	0.9981	0.9667	0.9877	0.9980
	0.1	0.9844	0.9847	0.9167	0.9703	0.9843
	0.15	0.9524	0.9538	0.8702	0.9307	0.9509
	0.2	0.8937	0.8982	0.8179	0.8836	0.8887
Divergent	0.02	0.9863	0.9867	0.9854	0.9707	0.9813
	0.05	0.9633	0.9618	0.9605	0.9282	0.9456
	0.1	0.9083	0.9076	0.9047	0.8700	0.8695
	0.15	0.8416	0.8426	0.8289	0.8096	0.7836
	0.2	0.7494	0.7537	0.7205	0.7348	0.6737
Unstructured	0.02	0.9553	0.955	0.9474	0.9449	0.9457
	0.05	0.8927	0.9007	0.8781	0.8749	0.8803
	0.1	0.8036	0.8223	0.7827	0.7733	0.7928
	0.15	0.7176	0.7367	0.6864	0.6887	0.7109
	0.2	0.6403	0.6582	0.5993	0.6298	0.6353
Independent	0.02	0.9887	0.9786	0.9888	0.9595	0.9844
	0.05	0.9684	0.9564	0.9670	0.8995	0.9566
	0.1	0.9074	0.9023	0.9045	0.8065	0.8803
	0.15	0.8248	0.8193	0.8147	0.7233	0.7777
	0.2	0.7043	0.7025	0.6917	0.6427	0.6431

Table 2 The comparison of each cognitive diagnosis method on average attribute match ratio

Attribute hierarchical structure	Slip	Average attribute match ratio (AAMR)				
		MDD-CTT	MDD-IRT	GDD-CTT	GDD-IRT	HDDB
Linear	0.02	1.000	1.0000	0.9967	0.9997	1.0000
	0.05	0.9997	0.9997	0.9944	0.9981	0.9997
	0.1	0.9975	0.9975	0.9860	0.9944	0.9971
	0.15	0.9911	0.9916	0.9776	0.9872	0.9896
	0.2	0.9790	0.9797	0.9629	0.9762	0.9747
Convergent	0.02	1.0000	1.0000	0.9967	0.9996	1.0000
	0.05	0.9997	0.9997	0.9940	0.9979	0.9996
	0.1	0.9973	0.9974	0.9850	0.9950	0.9971
	0.15	0.9913	0.9918	0.9761	0.9877	0.9898
	0.2	0.9796	0.9811	0.9645	0.9786	0.9749
Divergent	0.02	0.9976	0.9977	0.9972	0.9948	0.9956
	0.05	0.9932	0.9930	0.9923	0.9871	0.9861
	0.1	0.9814	0.9815	0.9805	0.9752	0.9618
	0.15	0.9655	0.9662	0.9638	0.9601	0.9311
	0.2	0.9411	0.9426	0.9390	0.9413	0.8891
Unstructured	0.02	0.9903	0.9916	0.9899	0.9886	0.9905
	0.05	0.9759	0.9802	0.9748	0.9739	0.9772
	0.1	0.9519	0.9614	0.9501	0.9506	0.955
	0.15	0.9257	0.9377	0.9221	0.9270	0.9244
	0.2	0.8993	0.9128	0.8946	0.9074	0.8905
Independent	0.02	0.9976	0.9962	0.9979	0.9927	0.9946
	0.05	0.9931	0.9918	0.9933	0.9811	0.9847
	0.1	0.9781	0.9791	0.9787	0.9599	0.9545
	0.15	0.9558	0.9577	0.9555	0.9361	0.9101
	0.2	0.9216	0.9256	0.9216	0.9084	0.8502

rate is very simple, using the IRT method to calculate the item parameters and ability parameters has preconditions, the estimation method is also more complex, so the MDD-CTT method is simple and feasible.

Figures 1 and 2 show that under the same attribute hierarchical structure, with the increase of slip, the PMR and AAMR have the tendency of decline for all methods. Among them, the MDD approach declines the most slowly under various hierarchical structures, while GDD falls faster than HDDB under the linear and convergence hierarchy. HDDB falls faster than GDD under the divergent and independent hierarchy. For various cognitive diagnosis methods, attribute hierarchical structure affect the accuracy of classification, the linear and convergence hierarchy has the highest accuracy, followed by divergent and independent hierarchy, unstructured hierarchy is the lowest, therefore, the identification of attribute hierarchical structure is very important.

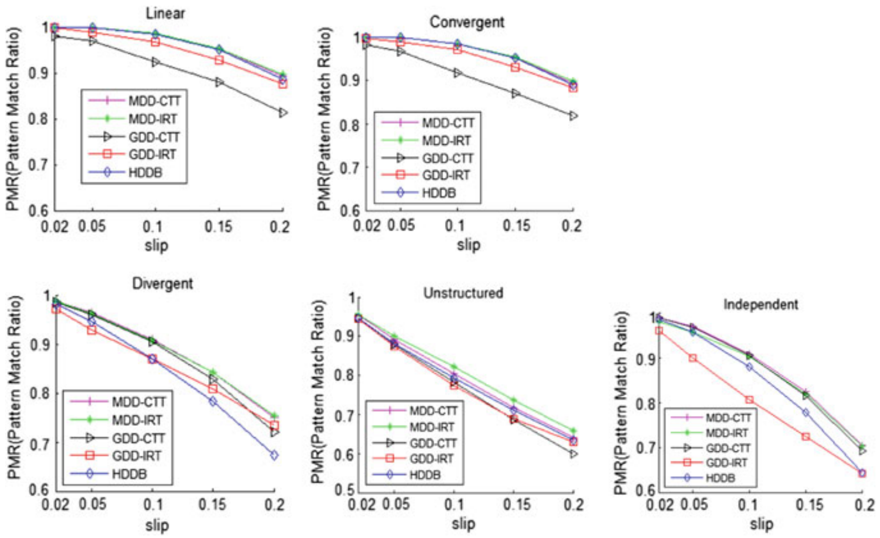


Fig. 1 The comparison of each cognitive diagnosis method on pattern match ratio

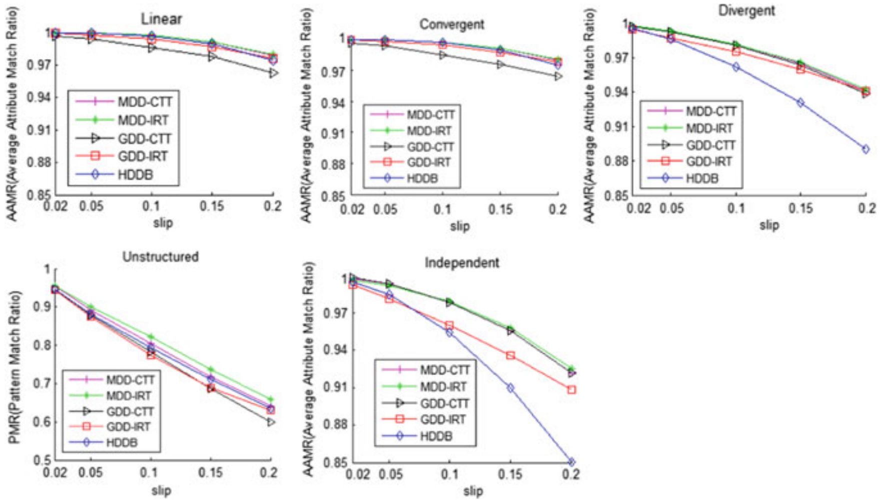


Fig. 2 The comparison of each cognitive diagnosis method on average attribute match ratio

6 Conclusions and Discussions

(1) For all methods, the classification accuracy are lower under independent and unstructured hierarchy, further research should focus on the design of Q-matrix in order to increase the classification accuracy.

(2) The classification accuracy of cognitive diagnosis methods is also affected by slip, the smaller the slip, and the more accurate the classification, likewise, the higher the slip, the lower the classification accuracy. In this study, MDD falls the most slowly, followed by GDD, and HDDB falls the fastest. The results related to the definition of the weighting matrix in different methods, such as HDDB method use unit matrix as weight matrix, although it is simple, it also loses some information of item parameters, thus when slip increases, the more obvious difference between ORP and IRP, and results the decline in classification accuracy rapidly; however, the definitions of weight matrix in MDD method and GDD method are related to the item parameters, therefore the distance between ORP and IRP can be corrected, the decrease is slower.

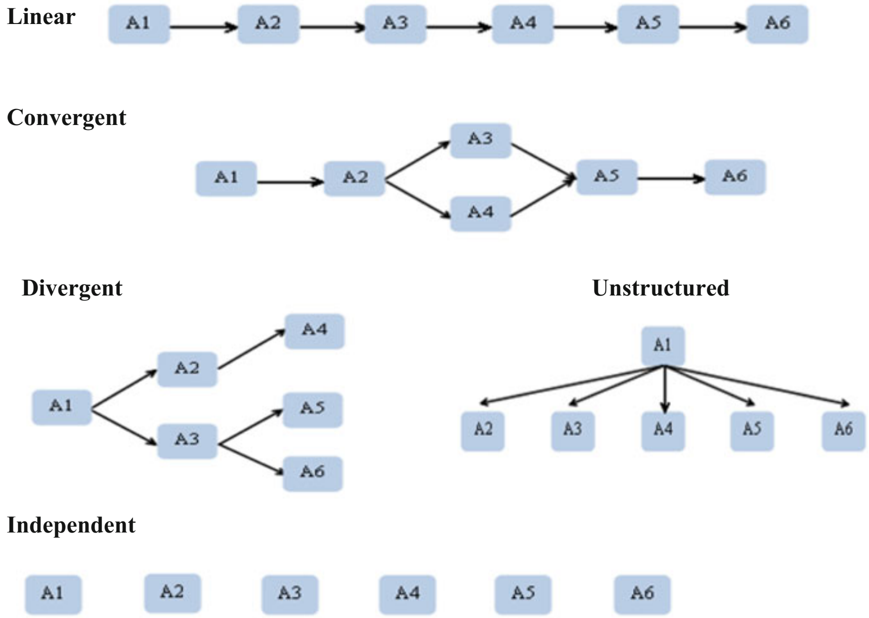
(3) For Mahalanobis distance, this study using Shannon entropy as diagonal elements of weight matrix achieves good performance, as a result, different weight matrixes can be considered, then newer classification methods can be obtained to discuss their classification performance.

(4) This paper extracts the essence of GDD method and HDD method, constructs a more general Mahalanobis distance discrimination method, and compares them under 2PLM. The next step will consider to extend this method to ploytomous model.

(5) The conclusions of this paper are obtained by Monte Carlo simulation study. Under the real situation, there exist a number of influencing factors and missing data, therefore, it is necessary to apply these methods to real data and verify their performance.

Acknowledgements This research is supported by the National Natural Science Foundation of China (Grant No. 31,360,237), Jiangxi Education Science Foundation (Grant No. GJJ160282), Chinese Test International research fund project (Grant No. CTI2017B06)

Appendix 1 Five Basic Attribute Hierarchy Structures



Linear

Convergent

Divergent

Unstructured

Independent

Appendix 2 the Test Q-Matrix of the Five Basic Attribute Hierarchy Structures

Linear

$$Q=\begin{pmatrix} 111111111111111111111111111111 \\ 011111011111011111011111011111 \\ 001111001111001111001111001111 \\ 000111000111000111000111000111 \\ 000011000011000011000011000011 \\ 000001000001000001000001000001 \end{pmatrix}$$

Convergent

$$Q=\begin{pmatrix} 11111111111111111111111111111111111 \\ 01111110111111011111101111110111111 \\ 00101110010111001011100101110010111 \\ 00011110001111000111100011110001111 \\ 00001100000110000011000001100000110 \\ 00000100000010000001000000100000010 \end{pmatrix}$$

Divergent

$$Q=\begin{pmatrix} 11111111111111111111111111111111 \\ 01000011110001110011001101010101 \\ 00100010001111110000111100110011 \\ 00010001001001001111111100001111 \\ 00001000100100101010101011111111 \\ 00000100010010010101010111111111 \end{pmatrix}$$

Unstructured

$$Q=\begin{pmatrix} 111111111111111111111111111111 \\ 010100111111011010100111111011 \\ 001011111111111001011111111111 \\ 000100000111001000100000111001 \\ 000010010010111000010010010111 \\ 000001001001111000001001001111 \end{pmatrix}$$

Independent

$$Q=\begin{pmatrix} 1000001111100000000000011111111 \\ 0100001000011110000000111100000 \\ 001000010001000111000100011100 \\ 000100001000100100110010010011 \\ 000010000100010010101001001010 \\ 000001000010001001011000100101 \end{pmatrix}$$

References

Cai, Y., Tu, D. B., & Ding, S. L. (2013). A simulation study to compare five cognitive diagnostic models. *Acta Psychologica Sinica, 45*(11), 1295–1304.

Chiu, C. Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*(2), 225–250.

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–183.

Ding, S. L., Mao, M. M., Wang, W. Y., Luo, F., & Cui, Y. (2012). Evaluating the consistency of test items relative to the cognitive model for educational cognitive diagnosis. *Acta Psychologica Sinica, 44*(11), 1535–1553.

Ding, S. L., Wang, W. Y., & Yang, S. Q. (2011). The design of cognitive diagnostic test blueprints. *Journal of Psychological Science, 34*(2), 258–265.

Ding, S. L., Yang, S. Q., & Wang, W. Y. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *Journal of Jiangxi Normal University, 34,* 490–495.

Kang, C. H., Ren, P., & Zeng, P. F. (2015). Nonparametric cognitive diagnosis: A cluster diagnostic method based on grade response items. *Acta Psychologica Sinica, 47*(8), 1077–1088.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205–237.

Li, Y., Ding, S. L., & Luo, F. (2012). The generalized distance discrimination based on graded response model. *Journal of Jiangxi Normal University, 36*(6), 636–639.

Luo, Z. S., Li, Y. J., Yu, X. F., Gao, C. L., & Peng, Y. F. (2015). A simple cognitive diagnosis method based on Q-matrix theory. *Acta Psychologica Sinica, 47*(2), 264–272.

Qi, S. Q., Dai, H. Q., & Ding, S. L. (2002). *Principles of modern educational and psychological measurement*. Beijing: Higher Education Press.

Sun, J. N., Xin, T., Zhang, S. M., & de la Torre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement, 37*(7), 503–521.

Sun, J. N., Zhang, S. M., Xin, T., & Bao, Y. (2011). A cognitive diagnosis method based on Q-Matrix and generalized distance. *Acta Psychologica Sinica, 43*(9), 1095–1102.

Tian, W., & Xin, T. (2012). A polytomous extension of rule space method based on graded response model. *Acta Psychologica Sinica, 44*(1), 249–262.

Tu, D. B., Cai, Y., & Dai, H. Q. (2013). Comparison and selection of five noncompensatory cognitive diagnosis models based on attribute hierarchy structure. *Acta Psychologica Sinica, 45*(2), 243–252.

Tu, D. B., Cai, Y., Dai, H. Q., & Ding, S. L. (2010). A polytomous cognitive diagnosis model: P-DINA model. *Acta Psychologica Sinica, 42*(10), 1011–1013.

Tu, D. B., Cai, Y., & Ding, S. L. (2012). *Cognitive diagnosis: Theory, methods and applications*. Beijing Normal University Publishing Group.

Wu, C. M. (2008). Image thresholding based on weighting shannon entropy. *Computer Engineering and Applications, 44*(18), 177–180.

Xin, T., Le, M. L., & Zhang, J. H. (2012). New progress and trends of measurement theory. *China Examinations, 5,* 3–11.

Zhang, S. M., Bao, Y., & Guo, W. H. (2013). A generalized cognitive diagnosis model under a particular polytomous situation. *Psychological Exploration, 33*(5), 444–450.

Zhang, Y. T., & Fang, K. T. (2013). *An introduction to multivariate statistical analysis*. Wuhan: Wuhan University Press.

An Joint Maximum Likelihood Estimation Approach to Cognitive Diagnosis Models

Youn Seon Lim and Fritz Drasgow

Abstract In this study, a simulation-based method for computing joint maximum likelihood estimates of cognitive diagnosis model parameters is proposed. The central theme of the approach is to reduce the complexity of models to focus on their most critical elements. In particular, an approach analogous to joint maximum likelihood estimation is taken, and the latent attribute vectors are regarded as structural parameters, not parameters to be removed by integration with this approach, the joint distribution of the latent attributes does not have to be specified, which reduces the number of parameters in the model. The Markov Chain Monte Carlo algorithm is used to simultaneously evaluate and optimize the likelihood function. This streamlined approach performed as well as more traditional methods for models such as the DINA, and affords the opportunity to fit more complicated models in which other methods may not be feasible.

Keywords Cognitive diagnosis model · Joint maximum likelihood estimation
Simulated annealing

1 Introduction

The cognitive diagnosis model is one of the important psychometrics models because it provides diagnostic information about each individual examinee. An important problem in the application of the model is the estimation of person and item parameters. In most applications, both person and item parameters must be estimated simultaneously. The method of joint maximum likelihood estimation is one procedure that

Y. S. Lim (✉)
Donald and Barbara Zucker School of Medicine at Hofstra/Northwell,
Hofstra University, Hempstead, NY 11549, USA
e-mail: YounSeon.Lim@hofstra.edu

F. Drasgow
University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
e-mail: fdrasgow@illinois.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_28

can be used for this purpose. This paper proposes a method that simultaneously evaluates and optimizes the joint likelihood function.

One classical approach to estimation has treated the person parameters as nuisance parameters and simply integrated them out of the likelihood equation. This approach, called marginal maximum likelihood, is thus a function of only the structural (i.e., item) parameters. In a simulation study, Drasgow (1989) found that marginal maximum likelihood estimates are more accurate than joint maximum likelihood estimates regardless of sample size or test length. However, obtaining marginal maximum likelihood estimates is a complex task because, in some cases, the likelihood function for the structural parameters is not available in closed form and, moreover, may be multimodal (e.g., Doucet et al. 2002). When the marginal likelihood is evaluated, the Expectation-Maximization (EM) algorithm is typically used. However, it is sensitive to initial values and can have local maxima (e.g., Doucet et al. 2002). Furthermore, it can be a computational burden to deal with high-dimensional integration in the EM.

Another approach is Bayesian estimation of the parameters using prior distributions on the person parameter, or on both person and item parameters. This procedure eliminates the problems sometimes encountered in the marginal maximum likelihood estimation (e.g., Hambleton et al. 1991). This approach, however, has its own difficulties. For instance, the prior specification and prior sensitivity are important aspects of Bayesian inferences (e.g., Ghosh et al. 2000). In practice, it can be difficult to give a meaningful full prior specification, especially, for models with many parameters. Furthermore, in the Bayesian framework, the homogenous Markov Chain Monte Carlo (MCMC) methods typically used for the estimation of model parameters are inefficient for maximum a posteriori estimation because a large amount of the computational burden is spent exploring regions of low posterior probability (e.g., Andrieu and Doucet 2000); for complex model estimation (i.e., the reparameterized unified model) MCMC may be prohibitively slow to converge. Finally, MCMC methods are often more suited for integration, not optimization problems (e.g., Jacquier et al. 2007).

Joint Maximum Likelihood Estimation (JMLE), in which item parameters are estimated at the same time as person parameters, is straightforward. The maximum likelihood estimates of the person and item parameters can be obtained from this likelihood function by standard procedures (e.g., Lord 1974). Neyman and Scott (1948) showed that when the number of structural parameters increases with the number of incidental parameters, estimates may not be consistent. Even when the estimates of structural parameters are consistent, the property of efficiency may not hold. Lord (1968) JMLE procedure is an example of the situation dealt with by Neyman and Scott, and the consistency of the structural (i.e., item) parameter estimates has been questionable.

However, in the context of one IRT model, Haberman (1977) proved the joint consistency of maximum likelihood estimates of item and person parameters for the Rasch model. He obtained strong consistent estimates of the parameters as the number of items and examinees go to infinity. Douglas (1997) was also able to prove uniform asymptotic consistency in a unidimensional class of kernel-smoothing-based nonparametric IRT item response function estimation procedures under less restrictive assumptions than Haberman's. Empirical results obtained by Lord (1975) and by Swaminathan and Gifford (1983), for example, showed that the JMLE procedure can give accurate results with as few examinees as $I = 200$ provided item $J \geq 60$. Hulin et al. (1982) conducted a Monte Carlo study to investigate the effects of four sample sizes ($I = 200$, 500, 1000, or 2000) and three test lengths (J =15, 30, or 60 items) on the accuracy of joint parameter estimation. They found that, for a two-parameter model, there must be at least $J = 30$ and $I = 500$, and for a 3-parameter model, there must be at least $J = 60$ and $I = 1000$.

The JMLE method proposed in this study is carried out by means of a combination of the simulated annealing algorithm and stochastic simulation of the hidden Markov chain. The central theme of the approach is to omit variables related to the joint distribution of latent attributes to trim back model complexity. This algorithm is shown to converge for the set of joint maximum likelihood parameter estimates under suitable regularity conditions. Note that in this study, we assume that the Q-matrix is known and is not estimated.

2 Algorithm and Properties

Let Y_{ij} denote the observed response of examinee i to item j, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$. For examinee i, let $\boldsymbol{\alpha}_i = \{\alpha_{ik}\}$ denote the latent binary attribute vector, $k = 1, 2, \ldots, K$, where $\alpha_{ik} = 1$ indicates mastery of the kth skill attribute and $\alpha_{ik} = 0$ indicates nonmastery of the attribute. Under the assumption of conditional independence, the joint likelihood L for the item responses is

$$L = L(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{I}\prod_{j=1}^{J} P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \boldsymbol{\beta}_j)^{Y_{ij}}[1 - P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \boldsymbol{\beta}_j)]^{1-Y_{ij}}, \quad (1)$$

where $\boldsymbol{\beta}_j = \{\beta_{jk}\}$ denotes the item parameters for item j.

The item parameters $\boldsymbol{\beta}_j$ as well as the person parameters $\boldsymbol{\alpha}_i$ are required to be estimated at the same time. The values of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_j$ that maximize the likelihood,

$$\arg\max_{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_j} L \quad (2)$$

are the joint maximum likelihood estimates. One approach is to estimate the values of parameters directly by iteratively setting $\frac{\partial L}{\partial \boldsymbol{\alpha}_i} = 0$, and $\frac{\partial L}{\partial \boldsymbol{\beta}_j} = 0$ (Lord 1968). However, some difficulties can be encountered. First, there are some cases or models in which the maximum likelihood estimates or the likelihood in closed form do not exist (e.g., Hambleton et al. 1991). Second, it is a computational burden to iterate between the two sets of partial derivatives; moreover, the numerical optimization on very high dimensional models is time consuming. Finally, it is challenging to estimate the standard errors of the maximum likelihood estimates based on the second order derivatives (e.g., Jacquier et al. 2007).

To avoid these potential problems, in this study the approach is modified in three different ways. One is to implement a regularization term for all model parameters. This is accomplished by establishing uniform (flat) prior distributions, and then obtaining the maximum a posteriori values of the parameters. The assumption of flat priors for the parameters means that the prior terms for those parameters can be set to unity, and therefore the maximum a posteriori updates for $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\boldsymbol{\beta}}_j$ are identical to the maximum likelihood updates for the parameters (e.g., Ghosh et al. 2000; Patz and Junker 1999) on bounded intervals.

Second, rather than estimate the distribution of $\boldsymbol{\alpha}_i$, each α_{ik} is treated as a parameter to be estimated. This has been problematic in IRT models in which the latent variables are continuous because something must be done to fix the scale. However, for cognitive diagnosis models in which the latent attributes are binary, the scale is solidly pinned down between the two possible values, 0 or 1 in the parameter space. This results in a more streamlined model and yields simpler Markov chains and consistent results, as shown in a later part of this paper.

Third, we propose an algorithm that is a combination of the insights of standard MCMC algorithms and simulated annealing algorithms in the Bayesian framework. The initial value of this algorithm is obtained from the nonparametric estimator of latent attribute variables (Lim and Drasgow 2017). Given the estimates of $\boldsymbol{\alpha}$, the item parameters are estimated, and then the estimates of item parameters are used to update the estimates of $\boldsymbol{\alpha}$. This procedure is repeated until the convergence criterion is satisfied.

Simulated annealing is an inhomogeneous variant of MCMC used to perform combinational optimization. This method samples from a sequence of density functions whose support concentrates itself on the set of maximum likelihood estimates. The power $\gamma(t)$, $t = 1, \ldots, T$, which is termed the temperature, makes it possible to explore the entire search space systematically by being increased simultaneously as the Markov chain increases (e.g., van Laarhoven and Aarts 1989). As in simulated annealing, this proposed algorithm replaces the target joint density $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta})$ as

$$\pi_{\gamma(t)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto P(\boldsymbol{\alpha}, \boldsymbol{\beta})^{\gamma(t)} P(\boldsymbol{\alpha}) P(\boldsymbol{\beta}), \tag{3}$$

where $\lim_{t\to+\infty} \gamma(t) = \infty$. When $\gamma(t) > 1$, $P(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is raised to the $\gamma(t)$ power and the effects of the priors $P(\boldsymbol{\alpha})$ and $P(\boldsymbol{\beta})$ disappear on the range of values (e.g., Jacquier et al. 2007). Nonetheless, they are necessary to ensure their integrability without affecting the maximum joint likelihood estimates.

In simulated annealing, convergence to the set of global maxima is ensured for a sequence $\gamma(t)$ growing logarithmically (e.g., van Laarhoven and Aarts 1989). However, it is difficult to concentrate on the global modes because the logarithmic function increases too slowly (e.g., Doucet et al. 2002). To solve this problem, it is assumed that the $\gamma(t)$ are fixed and do not depend on the iteration.

Formally, the proposed algorithm seeks to maximize the joint likelihood in the Bayesian framework with a fixed temperature $\gamma(T)$,

$$\pi_{\gamma(T)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto L^{\gamma(T)} P(\boldsymbol{\alpha}) P(\boldsymbol{\beta}). \tag{4}$$

The likelihood term L reappears in this Bayesian formulation but is now accompanied by the uninformative prior distributions of the parameters. As it is usually impossible to sample from the density directly, MCMC methods are used to simulate samples from a sequence of joint densities, $\pi_{\gamma(T),n}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, where n indexes the length of the Markov Chain.

It is important to compare and contrast this algorithm with the marginal maximum likelihood (or marginal maximum a posterior) estimation methods related to simulated annealing. The basic idea is to generate a sequence of artificial distributions from a density in which the latent variables are replicated temperature $\gamma(t)$ times by data argumentation. Then the sequence concentrates itself on the set of marginal maxima. For generation, non-homogenous MCMC algorithms (Andrieu and Doucet 2000; Doucet et al. 2002), original sequential monte carlo methods (Johannes et al. 2008), and a standard evolutionary MCMC method (Jacquier et al. 2007) have been employed.

These researchers advocate that as the chain goes to infinity, the sequence of density concentrates itself upon the marginal maximum of structural parameters. Then the estimates of structural parameters are obtained without resorting to a gradient based method. Temperature $\gamma(t)$ is assumed to be increased as the chain increases, especially in terms of the theoretical foundation. In contrast, our algorithm estimates the joint maximum likelihood in the Bayesian framework. The joint density is alternately raised to $\gamma(T)$ as in simulated annealing while the priors are not exponentiated unlike simulated annealing. The initial values of this algorithm is obtained from a nonparametric approach. The estimates for the values of parameters are obtained given the estimates of the other parameters. Furthermore, the joint distribution of latent variables does not need to be estimated because each component is regarded as an individual parameter. Unlike the algorithm presented here, the methodologies for marginal maximum likelihood (or marginal maximum a posterior) require or are suitable for continuous latent variable models.

This algorithm has several practical advantages. First, this approach does not require estimating the distribution of $\boldsymbol{\alpha}$. Second, unlike a Bayesian approach, informative prior distributions for the parameters are not necessary. Third, the MAP (= ML) estimates are obtained neither exploring regions of low posterior probability nor integrating over the incidental parameters. Finally, in combination with standard MCMC algorithm, simulated annealing maintains the speed and reliability of gradient descent algorithms while at the same time avoiding local minima (Zomaya and Kazman 2010). This approach can also handle models whose closed form expressions are unknown like the marginal maximum likelihood (or marginal maximum a posterior) estimation methods.

2.1 MCMC Algorithm

For the proposed approach, the Metropolis-Hastings algorithm with simulated annealing is used for sampling from $\pi_{\gamma(T)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto P(\boldsymbol{\alpha}, \boldsymbol{\beta})^{\gamma(T)} P(\boldsymbol{\alpha}) P(\boldsymbol{\beta})$. Like Birnbaum's two stage paradigm (Birnbaum 1968), this algorithm starts with the estimated initial values of latent attribute variable $\boldsymbol{\alpha}$ by using a nonparametric technique proposed by Lim and Drasgow (2017). In this approach, the uniform prior distributions are established over the parameters.

Step 1. Estimate the initial value of this algorithm: person parameter $\boldsymbol{\alpha}$
A nonparametric method (Lim and Drasgow 2017) is used to estimate the initial value of this algorithm. The method estimates the person parameter $\boldsymbol{\alpha}$ based on the Hamming distance between ideal and observed response patterns. This approach consists of two phases, the computation of all possible ideal response vectors and the classification phase.

The ideal responses $\eta_{ij\text{noncompensatory}}$ are defined as $\prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}}$, $\eta_{ij\text{disjunctive}}$ are defined as $1 - \prod_{k=1}^{K}(1 - \alpha_{ik})^{q_{jk}}$, and $\eta_{ij\text{compensatory}}$ are defined as rounding of $(\sum_{k=1}^{K} (\hat{\alpha}_{ik} \times q_{jk})/K)$ for examinee i and assessment item j. All possible ideal response vectors $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_{2^K}$ are constructed from all 2^K possible patterns for $\boldsymbol{\alpha}_i$. In the classification stage, the Hamming distances between $\boldsymbol{Y}_i$ and each of $\boldsymbol{\eta}_m$, for $m = 1, 2, \ldots, 2^K$, are computed by simply counting the number of times two vectors disagree as given by

$$D(\boldsymbol{Y}_i, \boldsymbol{\alpha}_m) = \sum_{j=1}^{J} \mid Y_{ij} - \eta_{mj} \mid . \tag{5}$$

The estimator is obtained by minimizing this distance over all possible attribute patterns,

$$\hat{\boldsymbol{\alpha}}_i = \arg \min_{m \in \{1,2,\ldots,2^K\}} D(\boldsymbol{Y}_i, \boldsymbol{\alpha}_m). \tag{6}$$

The theoretical justification is that the true attribute vector minimizes the expected distance between $\boldsymbol{Y}_i$ and $\boldsymbol{\eta}_m$, under some general conditions on the underlying model. Unlike the other nonparametric approaches (e.g., Chiu and Douglas 2013), the estimator is applicable for noncompensatory, disjunctive as well as compensatory models.

Step 2. Draw $\boldsymbol{\beta}^{\gamma(T)}_{\{n=1,2,\ldots,N\}} | \boldsymbol{\alpha}_{(0)} \sim P(\boldsymbol{\beta}^{\gamma(T)} | \boldsymbol{\alpha}_{(0)}, \mathbf{Y}) \propto P(\mathbf{Y} | \boldsymbol{\alpha}, \boldsymbol{\beta}^{\gamma(T)}) P(\boldsymbol{\beta})$.
Given the estimates of person parameter $\boldsymbol{\alpha}$, the Markov chains of item parameter $\boldsymbol{\beta}$ are obtained until meet the convergence criteria. Tests for normality of the draws such as Jarque-Bera test and Shapiro-Wilk goodness-of-fit test is used as the criteria (e.g., Chauveau and Diebolt 1998; Jacquier et al. 2007). Here $\boldsymbol{\beta}^{\gamma}(T)$ is considered as the $\gamma(t), t = 1, \ldots, T$ independent copies of $\boldsymbol{\beta}$. That is,

$$P(\boldsymbol{\beta}^{\gamma(T)} | \boldsymbol{\alpha}, \mathbf{Y}) \propto \prod_{t=1}^{T} (\mathbf{Y} | \boldsymbol{\alpha}, \boldsymbol{\beta}^{\gamma(t)}) P(\boldsymbol{\beta}). \tag{7}$$

Instead of generating $\gamma(T)$ copies, simulated annealing is used for this algorithm. The simulated annealing provides additional flexibility and efficiency in generating $\gamma(T)$ copies of item parameters $\boldsymbol{\beta}$. More specifically, (9) is obtained by

$$(\boldsymbol{\beta}_{(n+1)}^{\gamma(T)}, \boldsymbol{\beta}_{(n)}^{\gamma(T)}) = \min\{1, \exp(\gamma(T) \times (\log P(\boldsymbol{\beta}_{(n+1)}|\boldsymbol{\alpha}, \mathbf{Y})) - \log(P(\boldsymbol{\beta}_{(n)}|\boldsymbol{\alpha}, \mathbf{Y}))))\} \tag{8}$$

Step 3. Draw $\alpha_{ik,\{1,2,\ldots,N\}}^{\gamma(T)}|\boldsymbol{\beta}_{(0)} \sim P(\alpha_{ik}^{\gamma(T)}|\boldsymbol{\beta}_{(0)}, \mathbf{Y}) \propto P(\mathbf{Y}|\alpha_{ik}^{\gamma(T)}, \boldsymbol{\beta})P(\alpha_{ik})$.
Now given the estimated item parameters $\boldsymbol{\beta}$ from the previous step, the estimates of person parameter $\boldsymbol{\alpha}$ are updated. The draws of the person parameter $\boldsymbol{\alpha}$ are generated until no values are updated during an iteration like Hartz (2002). The independent draws of each $\gamma(t), t = 1, \ldots, T$ α_{ik} are,

$$P(\alpha_{ik}^{\gamma(T)}|\boldsymbol{\beta}, \mathbf{Y}) \propto \prod_{t=1}^{T}(\mathbf{Y}|\alpha_{ik}^{\gamma(t)}, \boldsymbol{\beta})P(\alpha_{ik}). \tag{9}$$

This is obtained by

$$(\alpha_{ik,(n+1)}^{\gamma(T)}, \alpha_{ik,(n)}^{\gamma(T)}) = \min\left\{1, \exp(\gamma(T) \times (\log(P(\alpha_{ik,(n+1)}|\boldsymbol{\beta}, \mathbf{Y})) - \log(P(\alpha_{ik,(n)}|\boldsymbol{\beta}, \mathbf{Y})))\right\} \tag{10}$$

Steps 2 and 3 are repeated until a stopping criterion is met. As in simulated annealing, the stop criterion is either determined by fixing the number of temperature schedule values, or by terminating operation of the algorithm if the Markov chains are identical for a number of chains (e.g., van Laarhoven and Aarts 1989).

3 Application to the DINA Model

The DINA model (Junker and Sijtsma 2001) is first considered as an example. This model has the item parameters $\boldsymbol{\beta} = (\mathbf{s}, \mathbf{g})$, and then

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \boldsymbol{\beta}_j) = P(Y_{ij} \mid \boldsymbol{\alpha}_i, s_j, g_j) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})}, \tag{11}$$

where $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}{}^{q_j k}$, $s_j = P(Y_j = 0 \mid \eta_j = 1)$, $g_j = P(Y_j = 1 \mid \eta_j = 0)$.

The marginal likelihood of this model is available in closed form, and the corresponding MLE and its asymptotic variance can be derived algebraically. Therefore the estimation of maximum likelihood with the EM-algorithm has been commonly used for the estimation of model parameters (e.g., de la Torre 2009). As mentioned in the section of introduction, however, this approach has its own weaknesses such as the sensitivity of initial values and the possibility of local maxima. Furthermore, the main difficulty is encountered when the number of latent skill attributes is larger than $K = 5$. Following are the steps of the algorithm to update parameters at iteration n;

Step 1. Estimate the initial value of this algorithm: person parameter $\boldsymbol{\alpha}$.

This algorithm states with the estimates of $\boldsymbol{\alpha}$ as mentioned above.

Step 2. Determine the temperature $\gamma(T)$.

A constant temperature can be implemented, or can be updated as an additional parameter until it reaches a frozen value. In this study, a constant Temperature $\gamma(T)$ = 1, 5, 10, or 20 is proposed as the Temperature schedule.

Step 3. Updating $s_j for j = 1, 2, \ldots, J$

A candidate value $s_j^\star$ is drawn from the uniform distribution on the interval $(s^l - \delta, s^h + \delta)$, where s^l and s^h are, respectively, the lower bound and the higher bound of the slip parameters; $\delta = 0.1$ in the following analyses. Typically, s^l and $s^h \in (0, 0.5)$. Calculate

$$r_n = \exp(\gamma(T) \times (\log(L(\mathbf{Y}_j|\boldsymbol{\alpha}^{(n-1)}, s_j^\star, g_j^{(n-1)})) - \log(L(\mathbf{Y}_j|\boldsymbol{\alpha}^{(n-1)}, s_j^{(n-1)}, g_j^{(n-1)})))) \tag{12}$$

which is the acceptance ratio and gives the probability of accepting the proposed value. Let $s_j^{(n)} = s_j^\star$ with probability $\min(1, r_n)$, otherwise let $s_j^{(n)} = s_j^{(n-1)}$.

Step 4. Updating $g_j for j = 1, 2, \ldots, J$

A candidate value $g_j^\star$ is drawn from the same uniform distribution used for $s_j^\star$. Compute the acceptance probability,

$$r_n = \exp(\gamma(T) \times (\log(L(\mathbf{Y}_j|\boldsymbol{\alpha}^{(n-1)}, s_j^{(n-1)}, g_j^\star)) - \log(L(\mathbf{Y}_j|\boldsymbol{\alpha}^{(n-1)}, s_j^{(n-1)}, g_j^{(n-1)})))) \tag{13}$$

Let $g_j^{(n)} = g_j^\star$ with probability $\min(1, r_n)$, otherwise let $g_j^{(n)} = g_j^{(n-1)}$.

Step 5. Updating $\alpha_{ik} for i = 1, 2, \ldots, I, k = 1, 2, \ldots, K$

For $\alpha_{ik}^\star$ in $\boldsymbol{\alpha}_i$, a candidate value is drawn from the binomial distribution $(1, 0.5)$. Compute the acceptance probability,

$$r_n = \exp(\gamma(T) \times (\log(L(\mathbf{Y}_i|\alpha_{ik}^\star, \mathbf{s}^{(n-1)}, \mathbf{g}^{(n-1)})) - \log(L(\mathbf{Y}_i|\alpha_{ik}^{(n-1)}, \mathbf{s}^{(n-1)}, \mathbf{g}^{(n-1)})))) \tag{14}$$

Let $\alpha_{ik}^{(n)} = \alpha_{ik}^\star$ with probability $\min(1, r_n)$, otherwise let $\alpha_{ik}^{(n)} = \alpha_{ik}^{(n-1)}$.

Step 5 is repeated until no values are updated during an iteration. Note that flat prior distributions are used throughout. For a sufficiently long chain, the parameters are estimated to approximate the posterior mode.

3.1 Simulation Study

A simulation study was carried out to evaluate the performance of the proposed MCMC algorithm under various conditions. In each condition, an item response data set from the DINA model with $K = 7$ was generated. Four conditions were considered: two test lengths J (short = 25, long = 50) and two examinee sample sizes I (small = 250, large = 1000). A Q-matrix for $J = 25$ was randomly generated

Table 1 Correctly specified Q ($K = 7$)

Item	$K = 7$							Item	$K = 7$						
1	0	0	1	1	1	1	1	14	1	1	0	0	0	0	0
2	0	0	1	0	1	1	0	15	1	0	1	0	0	0	1
3	1	0	0	1	0	0	0	16	1	1	0	0	0	1	1
4	1	1	0	0	1	1	1	17	0	0	0	1	0	0	1
5	0	1	1	1	1	0	0	18	1	0	1	0	0	0	0
6	1	0	0	0	1	0	1	19	0	0	1	0	1	0	0
7	0	0	0	1	1	1	1	20	0	1	0	0	1	1	0
8	0	0	1	1	0	0	1	21	0	0	0	0	0	0	1
9	1	1	0	1	0	0	1	22	1	0	1	0	1	1	1
10	0	0	1	0	1	1	0	23	0	1	0	0	0	0	1
11	0	0	1	0	1	0	1	24	1	1	1	0	1	1	1
12	1	0	0	0	0	0	1	25	1	1	1	0	1	1	1
13	0	1	1	0	1	1	0								

from $2^K - 1$ possible $q-$vectors as presented in Table 1. The Q-matrix for $J = 50$ was obtained by duplicating the matrix two times.

The item parameters were generated from $s_j \sim Unif(0, 0.3)$, and $g_j \sim Unif(0, 0.3)$. The person parameters $\boldsymbol{\alpha}$ were sampled from the bivariate Normal distribution with mean vector $\mu = (0, 0, 0, 0, 0, 0, 0)$ and covariance matrix $\sum$ with all 1's on the diagonal and off-diagonal elements of 0.3. Binary traits were constructed as in Chiu et al. (2009),

$$\alpha_{ik} = \begin{cases} 1, & \text{if } \theta_{ik} \geq \boldsymbol{\Phi}^{-1} \frac{k}{K+1}; \\ 0, & \text{otherwise} \end{cases}$$

Markov chains were run with four different values of $\gamma(T)$: 1, 5, 10, and 20. The estimates of model parameters were obtained by estimating the modes of the draws based on the the criterion of Gelman and Rubin (1992) which was calculated by generating five parallel Markov chains. The criterion was satisfied for all item parameters. The convergence of the person parameter was estimated indirectly by evaluating the agreement rate between the true α and estimated $\hat{\alpha}$ in the simulation study because this parameter is dichotomous.

Table 2 reports the results of item parameter estimation. The estimation accuracy was calculated by $RMSE = \sqrt{\sum_{j=1}^{J} (\hat{\beta}_j - \beta_j)^2 / J}$. Furthermore the results were compared with the results from two different approaches: one was from marginal maximum likelihood estimation with the EM-algorithm (Robitzsch et al. 2015), and the other one was from the fully Bayesian MCMC model based on the algorithm proposed for the Hierarchical DINA model (de la Torre and Douglas 2004).

The RMSE from the proposed algorithm decreased as the sample size I increased when $J = 50$. Unlike the estimates from the fully Bayesian model and marginal

Table 2 RMSE of item parameters

Condition	Parameter	EM-A	Fully-B	$\gamma(T)$			
				20	10	5	1
$J = 25$, $I = 250$	Slip	0.114	0.058	0.081	0.087	0.094	0.126
	Guess	0.042	0.036	0.043	0.060	0.062	0.063
$J = 25$, $I = 1000$	Slip	0.064	0.048	0.086	0.088	0.085	0.089
	Guess	0.041	0.028	0.055	0.060	0.057	0.056
$J = 50$, $I = 250$	Slip	0.078	0.068	0.076	0.079	0.080	0.103
	Guess	0.041	0.030	0.027	0.028	0.028	0.029
$J = 50$, $I = 1000$	Slip	0.041	0.043	0.051	0.053	0.052	0.061
	Guess	0.015	0.019	0.019	0.020	0.022	0.020

maximum likelihood with the EM algorithm, the influence of the test length J was moderate. For this reason, this method worked property for the condition of sort test length $J = 25$ and small sample size $I = 250$. The theoretical convergence results in simulated annealing indicate that, as $\gamma(T)$ increases, the draw will converge to the joint maximum likelihood estimate. However, in this application, the RMSE slightly increased as the $\gamma(T)$ increased. This might be caused by local maxima. This problem could be fixed when the optimal temperature $\gamma(T)$ was determined in the simulated annealing with a slowly increasing temperature schedule.

Figure 1 shows the draws of the guessing parameter for the four runs of the algorithm with $\gamma(T) = 1, 5, 10$, and 20. The horizontal red lines show the true parameter value. The plots confirm that moderate increases in $\gamma(T)$ quickly reduce the variance of draws. The draws of the slip parameter for the same item are shown in the Fig. 2. Like the guessing parameter, the variance of the draws was reduced as the $\gamma(T)$ increased. However, $\gamma(T)$ seems to need to be increased further to reduce the variance.

The proportion of the times in which the true $\boldsymbol{\alpha}_i$ and the estimated $\hat{\boldsymbol{\alpha}}_i$ agreed was summarized for each condition in two different ways: one was the Component-wise Agreement Rate (CAR)= $(\sum_{i=1}\sum_{k=1}|\alpha_{i_k} = \hat{\alpha}_{i_k}|)/(I \times K)$, and the other one is the Vector-wise Agreement Rate (VAR) = $(\sum_{i=1}|\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i|)/I$. Likewise in Table 3, correct classification rates obtained from JMLE are similar to the rates of the other methods.

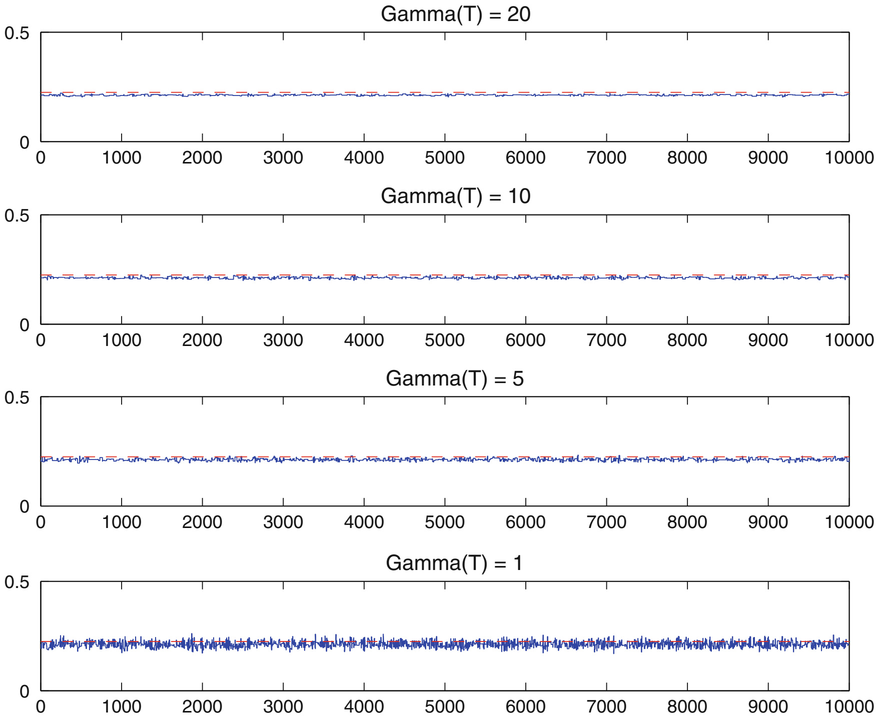


Fig. 1 Time plot of item 1 guess parameter in $J = 50$, $I = 1000$

3.2 Analysis of Fraction Subtraction Data with DINA Model

As an illustration of the model with real data, the DINA model with the proposed algorithm was fitted to the fraction subtraction data that includes the item responses to 20 items with 8 necessary attributes from 536 examinees. The data were originally collected and analyzed by Tatsuoka (1990) and have been analyzed in numerous studies. Here we use the Q matrix in Table 4 for the data that appeared in de la Torre and Douglas (2004). The specified attributes are (1) Convert a whole number to a fraction, (2) Separate a whole number from fraction, (3) Simplify before subtracting, (4) Find a common denominator, (5) Borrow from whole number part, (6) Column borrow to subtract the second numerator from the first, (7) Subtract numerators, and (8) Reduce answers to simplest form.

Unlike the simulation studies, the optimal $\gamma(T)$ was empirically determined by searching uniform simulated annealing schedule until reach the frozen value. The estimates of model parameters were obtained from the mode. The proportions of examinees I who mastered or not mastered each attribute are summarized in Table 5. The proportions were consistent for all three approaches (i.e., the overall difference of maters with Fully-B is 0.016 and with EM-A is 0.007).

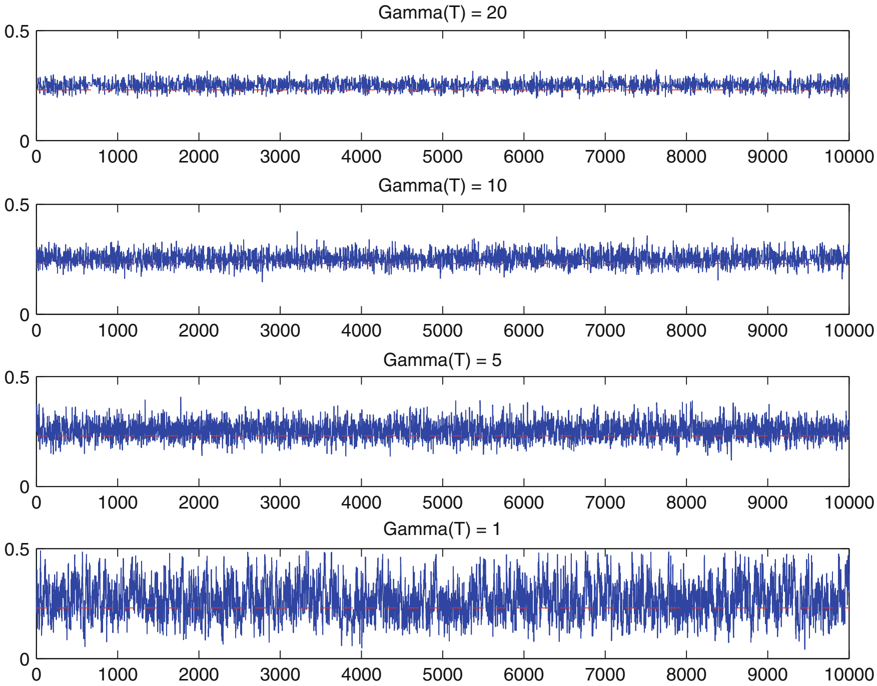


Fig. 2 Time plot of item 1 slip parameter in $J = 50$, $I = 1000$

Table 3 Agreement rates between $\hat{\alpha}$ and α

Condition	Parameter	EM-A	Fully-B	$\gamma(T)$			
				20	10	5	1
$J = 25$, $I = 250$	CAR	0.773	0.781	0.809	0.801	0.783	0.784
	VAR	0.324	0.328	0.323	0.316	0.304	0.300
$J = 25$, $I = 1000$	CAR	0.762	0.763	0.803	0.779	0.782	0.778
	VAR	0.286	0.293	0.305	0.294	0.301	0.296
$J = 50$, $I = 250$	CAR	0.866	0.864	0.855	0.829	0.837	0.829
	VAR	0.452	0.468	0.444	0.404	0.388	0.376
$J = 50$, $I = 1000$	CAR	0.844	0.802	0.820	0.799	0.798	0.790
	VAR	0.433	0.355	0.368	0.357	0.362	0.354

Table 4 Q for the fraction subtraction data

Item	$K = 8$								Item	$K = 8$							
1	0	0	0	1	0	1	1	0	11	0	1	0	0	1	0	1	0
2	0	0	0	1	0	0	1	0	12	0	0	0	0	0	0	1	1
3	0	0	0	1	0	0	1	0	13	0	1	0	1	1	0	1	0
4	0	1	1	0	1	0	1	0	14	0	1	0	0	0	0	1	0
5	0	1	0	1	0	0	1	1	15	1	0	0	0	0	0	1	0
6	0	0	0	0	0	0	1	0	16	0	1	0	0	0	0	1	0
7	1	1	0	0	0	0	1	0	17	0	1	0	0	1	0	1	0
8	0	0	0	0	0	0	1	0	18	0	1	0	0	1	1	1	0
9	0	1	0	0	0	0	0	0	19	1	1	1	0	1	0	1	0
10	0	1	0	0	1	0	1	1	20	0	1	1	0	1	0	1	0

Table 5 Attribute mastery or non-mastery rates for individual attributes

Attribute	Number of items	Mastery proportion			Non-mastery proportion		
		JMLE	Fully-B	EM-A	JMLE	Fully-B	EM-A
1	3	0.549	0.591	0.545	0.451	0.409	0.455
2	13	0.806	0.808	0.804	0.194	0.192	0.196
3	3	0.659	0.679	0.674	0.341	0.321	0.326
4	5	0.674	0.662	0.670	0.326	0.338	0.330
5	8	0.618	0.604	0.593	0.382	0.396	0.407
6	2	0.698	0.735	0.726	0.302	0.265	0.274
7	19	0.825	0.832	0.830	0.175	0.168	0.170
8	3	0.741	0.780	0.785	0.259	0.220	0.215
Mean		0.696	0.712	0.703	0.304	0.288	0.297

As shown in the Table 6, the estimates from JMLE are slightly different from the estimates of other methods (i.e., for guessing parameters, the difference with Fully-B is 0.001, and with the EM-A is 0.001; for slipping parameters, the difference with Fully-B is 0.032, and with the EM-A is 0.029).

4 Discussion

In this study, an MCMC algorithm is proposed for joint maximum likelihood estimation of parameters of various cognitive diagnosis models. This MCMC algorithm has the advantage of the standard MCMC algorithm and simulated annealing simultaneously. The significance of this approach is that it enables researchers to trim back model complexity by considering each α as an individual parameter to be estimated;

Table 6 Items parameter estimation with DINA model

Item	Guessing			Slipping		
	JMLE	Fully-B	EM-A	JMLE	Fully-B	EM-A
1	0.023	0.045	0.030	0.105	0.102	0.090
2	0.053	0.037	0.016	0.036	0.036	0.042
3	0.011	0.008	0.000	0.110	0.120	0.134
4	0.256	0.229	0.224	0.078	0.114	0.110
5	0.279	0.308	0.302	0.097	0.179	0.172
6	0.103	0.064	0.096	0.057	0.046	0.041
7	0.059	0.029	0.025	0.128	0.201	0.197
8	0.350	0.430	0.443	0.169	0.186	0.182
9	0.043	0.162	0.253	0.202	0.248	0.245
10	0.044	0.034	0.029	0.150	0.215	0.214
11	0.071	0.068	0.066	0.078	0.079	0.082
12	0.189	0.133	0.131	0.026	0.049	0.041
13	0.016	0.018	0.013	0.301	0.333	0.335
14	0.070	0.065	0.062	0.058	0.066	0.061
15	0.053	0.035	0.032	0.076	0.109	0.106
16	0.122	0.112	0.109	0.089	0.118	0.111
17	0.044	0.046	0.039	0.132	0.139	0.135
18	0.149	0.126	0.119	0.147	0.144	0.138
19	0.037	0.026	0.022	0.113	0.242	0.241
20	0.040	0.017	0.013	0.104	0.160	0.157

thus it is possible to estimate the item parameters and person parameters simultaneously.

The applications of the DINA model was provided as examples. As expected, as $\gamma(T)$ slightly increased, the variance of draws was reduced. The estimates of model parameters relatively were consistent regardless of the sizes of sample I and item J. It indicates that this approach is appropriate for the estimation of small sizes of sample I and item J. However, it is unreasonable to determine the performance of the algorithm by comparing the estimates of model parameters given each $\gamma(T) = 1, 5, 10, 20$ with their true values because as $\gamma(T)$ increases up to the optimal value, the draws will be closer to the true values (e.g., Jacquier et al. 2007).

Future research might include simulation using more attributes. In addition, the optimal temperature $\gamma(T)$ for each cognitive diagnosis model could be examined empirically. At the present time, however, the new MCMC algorithm appears to be a promising approach for joint maximum likelihood estimation of the parameters of cognitive diagnosis models.

References

Andrieu, C., & Doucet, A. (2000). Simulated annealing for maximum a posteriori parameter estimation of hidden markov models. *IEEE Transactions on Information Theory*, *46*, 994–1004.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Chauveau, D., & Diebolt, J. (1998). An automated stopping rule for MCMC convergence assessment. Retrieved from https://hal.inria.fr/inria-00073116

Chiu, C., & Douglas, J., (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*, 225–250.

Chiu, C., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–655.

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

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

Doucet, A., Godsill, S. J., & Robert, C. P. (2002). Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, *12*, 7784.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, *62*, 7–28.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement.*, *13*, 77–90.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457–472.

Ghosh, M., Ghosh, A., Chen, M., & Agresti, A. (2000). Noninformative priors for one parameter item response models. *Journal of Statistical Planning and Inference*, *88*, 99–115.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics*, *5*, 815–841.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hulin, C., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*, 249–260.

Jacquier, E., Johannes, M., & Polson, N. (2007). MCMC maximum likelihood for latent state models. *Journal of Econometrics*, *137*, 615640.

Johannes, A. M., Doucet, A., & Davy, M. (2008). Particle methods for maximum likelihood estimation in latent variable models. *Statistics and Computing*, *18*, 47–57.

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

Lim, Y. S., & Drasgow, F. (2017). Nonparametric calibration of item-by-attribute matrix. *Multivariate Behavioral Research*, *52*, 562–575.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*, 989–1020.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247–264.

Lord, F. M. (1975). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters,(Research Bulletin RB-75-33)*. Princeton, NJ: Educational Testing Service.

Neyman, J., & Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 132.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Robitzsch A., Kiefer, T., Geoge, A. C., & Uenlue, A. (2015). Cognitive Diagnosis Modeling: The R Package CDM.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), *New Horizons in Testing* (pp. 13–20). New York: Academic Press.

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

Van, Laarhoven, P. J. M., & Aarts, E. H. L. (1989). *Simulated annealing: Theory and application*. Amsterdam: Reidel.

Zomaya, A. Y., & Kazman, R. (2010). Simulated annealing techniques. In M. J. Atallah & M. Blanton (Eds.), *Algorithms and Theory of Computation Handbook*. Boca Raton, FL: CRC Press.

An Exploratory Discrete Factor Loading Method for Q-Matrix Specification in Cognitive Diagnostic Models

Wenyi Wang, Lihong Song and Shuliang Ding

Abstract The Q-matrix is usually unknown for many existing tests. If the Q-matrix is specified by subject matter experts but contains a large amount of misspecification, it will be difficult for the recovery of a high-quality Q-matrix through a validation method, because the performance of the validation method relies on the quality of a provisional Q-matrix. Under these two situations above, an exploratory technique is necessary. The purpose of this study is to explore a simple method for Q-matrix specification, called a discretized factor loading (DFL) method, in which exploratory factor analysis regarding latent attributes as latent factors is used to estimate a factor loading matrix after which a discretization process is employed on the factor loading matrix to obtain a binary Q-matrix. A series of simulation studies were conducted to investigate the performance of the DFL method under various conditions. The simulation results showed that the DFL method can provide a high-quality provisional Q-matrix.

Keywords Cognitive diagnosis · The Q-matrix · Exploratory factor analysis
The DINA model · The reduced RUM · The DINO model

W. Wang · S. Ding
School of Computer and Information Engineering, Jiangxi Normal University,
99 Ziyang Road, Nanchang, Jiangxi, People's Republic of China
e-mail: wenyiwang@jxnu.edu.cn

S. Ding
e-mail: ding06026@163.com

L. Song (✉)
Elementary Educational College, Jiangxi Normal University, 99 Ziyang Road,
Nanchang, Jiangxi, People's Republic of China
e-mail: viviansong1981@163.com

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_29

1 Introduction

In educational assessment, Cognitive Diagnostic Assessment (CDA) that combines psychometrics and cognitive science has received increased attention (Leighton and Gierl 2007; Rupp et al. 2010; Tatsuoka 2009). This approach potentially provides useful diagnostic information regarding students' strengths and weaknesses, and can facilitate individualized learning (Chang 2015; Chang and Wang 2016). Cognitive Diagnostic Models (CDMs) often utilize a Q-matrix (Embretson 1984; Tatsuoka 1990, 1995, 2009). The Q-matrix is an incidence matrix that shows the relationship between items and the underlying cognitive skills and attributes. The entries of the Q-matrix are 1 or 0, in which $q_{jk} = 1$ means that attribute k is involved in correctly answering item j, otherwise, $q_{jk} = 0$.

The Q-matrix plays an important role in establishing the relation between latent attribute patterns (or knowledge states) and ideal response patterns. The ideal response patterns are defined as latent responses of examinees without slipping and guessing. Meanwhile, a CDM entails developing a clear correspondence between examinees' observed item response patterns and the corresponding ideal response patterns. Thus, an inference of whether an individual has mastered some attributes or not can be drawn from an examinee's observed item response pattern.

To guarantee the validity of this inference, a correct specification of the Q-matrix is a fundamental step for CDA (Im 2007; Im and Corter 2011; McGlohen 2004; McGlohen and Chang 2008). The procedure for specifying the Q-matrix is usually an iterative process (Buck et al. 1998; Jang 2009): (a) the provisional Q-matrix is basically exploratory based on a current related theory, subject matter experts' judgment, and an item analysis; in addition (b) the modified Q-matrix is basically confirmatory based on statistical methods. The above two steps represent qualitative and quantitative methods respectively, and either of them alone is not enough to guarantee the correctness of a Q-matrix.

In order to improve the quality of a Q-matrix, researchers have proposed several quantitative methods for Q-matrix validation, such as the (sequential EM-based) δ method (de la Torre 2008) and its extension, the ς^2 method (de la Torre and Chiu 2010, 2016; Huo and de la Torre 2013), the γ method (Tu et al. 2012), the Bayesian approach (DeCarlo 2012), the data-driven approach (Liu et al. 2012, 2013), the nonparametric Q-matrix refinement method (Chiu 2013), and the stepwise reduction algorithm (Hartz 2002).

These validation methods often needed a high-quality provisional Q-matrix. For instance, if the provisional Q-matrix is unknown for an existing test, the validation methods can not be used. In addition, if the provisional Q-matrix is specified by subject matter experts but contains a large amount of misspecification, it will be difficult for the recovery of a high-quality Q-matrix through an validation method, because the performance of the validation methods relies on the precision of classification of attribute patterns resulting from the provisional Q-matrix (de la Torre 2008; Rupp and Templin 2008).

Under these two situations above, an exploratory technique is necessary. There has been a study about the adoption of principal components analysis as an exploratory technique for finding the Q-matrix, assuming that items measuring the same skill set will load on the same component (Close 2012). However, since a large number of attribute sets are expected to yield a large number of components, it is hard to determine the meaning of the components (Close 2012).

The purpose of this study is to explore a simple method for Q-matrix specification, called a discretized factor loading (DFL) method. An exploratory factor analysis (EFA) regarding latent attributes as latent factors is used to estimate a factor loading matrix after which a discretization process is employed on the factor loading matrix to obtain a binary Q-matrix. A series of simulation studies were conducted to investigate the performance of the DFL method under various conditions. Response data were simulated from the deterministic-inputs, noisy "and" gate (DINA) model (Haertel 1989; Junker and Sijtsma 2001), the reduced reparameterized unified model (rRUM; Hartz 2002), and the deterministic-inputs, noisy "or" gate (DINO) model (Templin and Henson 2006).

2 Method

2.1 Cognitive Diagnostic Models

Let X_{ij} be the response of examinee i to item j, i = 1, 2, …, N, j = 1, 2, …, J. Let $\boldsymbol{\alpha}_\mathbf{i} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ be the attribute pattern of examinee i and $\mathbf{q}_j = (q_{j1}, q_{j2}, \ldots, q_{jK})$ be the j-th row of the Q-matrix, where K is the number of attributes and the entries of both $\boldsymbol{\alpha}_\mathbf{i}$ and $\mathbf{q}_j$ only contains 0s and 1s.

The item response function for the DINA model is as follows:

$$P_j(\boldsymbol{\alpha}_\mathbf{i}) = P(X_{ij} = 1 \mid \boldsymbol{\alpha}_\mathbf{i}) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}}, \tag{1}$$

where $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ is a latent response variable or an ideal response (see Junker and Sijtsma 2001), and s_j and g_j are the slipping and guessing parameters of item j.

The item response function for the rRUM is as follows:

$$P_j(\boldsymbol{\alpha}_\mathbf{i}) = P(X_{ij} = 1 \mid \boldsymbol{\alpha}_\mathbf{i}) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{ik})q_{jk}}, \tag{2}$$

where π_j^* is the baseline parameter and r_{jk}^* is the penalty parameter. π_j^* is the probability of a correct response to item j given that an examinee has mastered all the required attributes for the item. The probability of a correct response to item j is proportional to r_{jk}^* when an examinee has not mastered attribute k.

The item response function for the DINO model is as follows:

$$P_j(\boldsymbol{\alpha}_{\mathbf{i}}) = P(X_{ij} = 1 \mid \boldsymbol{\alpha}_{\mathbf{i}}) = (1 - s_j)^{w_{ij}} g_j^{1 - w_{ij}}, \tag{3}$$

where $w_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}$ is an ideal latent response. As in the DINA model, s_j and g_j are the slipping and guessing parameters of item j.

2.2 An Exploratory Method: Discretized Factor Loading (DFL) Method

There has been a study about the adoption of principal components analysis for finding the Q-matrix, assuming that items measuring the same skill sets will load on the same component (Close 2012). However, since a large combination of attributes resulted in the total number of components increased, it is hard to determine the meaning of the components. Thus, for Q-matrix specification, we attempt to use an exploratory factor analysis (EFA) method, regarding the latent attributes as the latent factors. Four steps of the algorithm are as follows:

Step 1. Use the item responses of all examinees to estimate the tetrachoric correlation coefficient of each pair of items. Let a, b, c, and d denote the four cell counts of a 2 × 2 contingency table between a pair of items. An estimate of tetrachoric correlation is $r_{tet} = \cos\left(\pi/\left(1 + \sqrt{(ad)/(bc)}\right)\right)$ (Castellan 1966).

Step 2. Obtain the maximum likelihood estimate of the factor loading matrix.

Step 3. Perform Promax or Varimax rotation to maximize the variance of the factors, and the resulting loading matrix is denoted by $\Lambda = (a_{jk})$. Varimax rotation developed by Kaiser (1958) is an orthogonal rotation, which assumes no intercorrelations between components or attributes. Promax rotation relaxes the orthogonality constraint in order to gain simplicity of interpretation.

Step 4. Apply the discretization process on Λ to obtain a binary matrix. Each entry of the loading matrix is converted to a binary value by the use of threshold t. If $a_{jk} \geq t$ then $q_{jk} = 1$, otherwise $q_{jk} = 0$.

3 Simulation Study

A simulation study was conducted to investigate the performance of the DFL method under various conditions.

3.1 Simulation Design

To investigate whether the DFL method can work under certain conditions, five factors were included in the design of the simulation study. We considered five attributes. A total of 108 conditions were simulated (3 correlations × 3 sample sizes × 3 models × 2 Q-matrices × 2 item parameters). 30 replication data sets were enough for each condition. This is because the DFL method is an exploratory method to obtain a provisional Q-matrix, we only need to consider the uncertainly of the item responses.

(a) The source of the examinees' attribute patterns includes a discrete uniform distribution (Cheng 2009; Liu et al. 2013) and a multivariate normal threshold model (Chiu et al. 2009). For the former distribution, the test takers are generated assuming that every examinee has a 50% chance of mastering each attribute. In other words, for a 5-attribute test, the 32 attribute mastery patterns are equally likely in the population. For the latter distribution, the underlying continuous ability are drawn from a multivariate normal distribution (i.e., $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \boldsymbol{\rho})$), where $\boldsymbol{\rho}$ represents a correlation matrix with equal off-diagonal elements. The elements of $\boldsymbol{\rho}$ are either all 0.5 or all 0.75 (Henson and Douglas 2005), representing moderate or high correlation relationship, respectively. It would be better to use a design where some attributes are easier to master than others. Thus, the i-th individual's mastery for attribute k $(k = 1, 2, \ldots, K)$ was simulated, as in the study of Chiu et al. (2009):

$$\alpha_{ik} = \begin{cases} 1 & \mathit{if}\ \theta_{ik} \geq \Phi^{-1}(\frac{k}{K+1}) \\ 0 & \mathit{otherwise} \end{cases}, \tag{4}$$

where Φ^{-1} is the inverse of the normal distribution function Φ.

(b) The number of examinees is $N = 300$, 500, or 1,000.

(c) Three cognitive diagnostic models are considered, such as the DINA model, the rRUM, and the DINO model. The DINA model is a conjunctive/noncompensatory model; the rRUM is a partial compensatory model; the DINO model is a compensatory/disjunctive model.

(d) Two true Q-matrices are designed. The first Q-matrix was fixed as the reduced Q-matrix (Qr) with 31 items including all possible q-vectors. The second Q-matrix contains two identity matrices horizontally stacked and all possible items required two attributes, which is the same as Q_3 used in Liu et al. (2012). This Q-matrix, with an identity or an reachability ($\boldsymbol{R}$) matrix, was called a complete Q-matrix (Chiu et al. 2009) or a necessary and sufficient Q-matrix (Ding et al. 2010). Because the necessary and sufficient Q-matrix can distinguish all ideal item response patterns of attribute patterns, the correct classification rate of attribute patterns can be improved. Since each single factor or attribute is measured by different test items, the Q-matrix can be easily recovered by the DFL method.

(e) The quality of the items has two levels based on item parameters. Items with s, g ~ $U(0.05, 0.25)$ for the DINA or DINO model, or $\pi^* \sim U(0.8, 0.98)$ and $r^* \sim U(0.1, 0.6)$ for the rRUM were labeled high quality. Items with s, g ~ $U(0.05, 0.4)$ or $\pi^* \sim U(0.75, 0.95)$ and $r^* \sim U(0.2, 0.95)$ were labeled low quality. Item quality in this study was defined as the average of the discriminating powers of items in a test (Cui et al. 2012) or item parameters (Liu et al. 2016; Ma et al. 2016). In practice, item quality would be defined in terms of both discriminating power and coverage of the content specifications (Xing and Hambleton 2004). In general, for the DINA or DINO model, a high quality or "good" item will have small slipping and guessing parameters (Rupp et al. 2010), which means that the item discrimination powers are large (Cui et al. 2012). For the rRUM, a high quality or "good" item will have a high π^* and low r^* parameters (Rupp et al. 2010).

3.2 Methods and Evaluation Criteria

A computer program was written in Matlab 2008. At the Step 4 in the DFL method, we choose different thresholds to transform a continuous factor loading matrix into a discrete Q-matrix. Seven thresholds are used, including row/item mean, column/attribute mean, total mean, and four fixed thresholds (0.3, 0.2, 0.1, or 0.0):

$$q_{jk} = \begin{cases} 1 & if\, a_{jk} \geq \sum_{k=1}^{K} a_{jk}/K \\ 0 & otherwise \end{cases}, \tag{5}$$

$$q_{jk} = \begin{cases} 1 & if\, a_{jk} \geq \sum_{J=1}^{J} a_{jk}/J \\ 0 & otherwise \end{cases}, \tag{6}$$

$$q_{jk} = \begin{cases} 1 & if\, a_{ik} \geq \sum_{J=1}^{J} \sum_{J=1}^{J} a_{jk}/(JK) \\ 0 & otherwise \end{cases}, \tag{7}$$

$$q_{jk} = \begin{cases} 1 & if\, a_{jk} \geq t, t\, is\, fixed\, as\, 0.3, 0.2, 0.1, or\, 0 \\ 0 & otherwise \end{cases}. \tag{8}$$

The results reported in this study focused on the accuracy of an estimated Q-matrix, because it was directly related to the performance of the DFL method.

The correct recovery rate (CRR) equals the ratio of the number of correct q-entries in the estimated Q-matrix to the total number of q-entries (Chiu 2013). For each condition, the mean and standard deviation of the CRR values of the 30 replications were reported for each method.

For gaining insight into the performance of these methods in two different aspects, the under-specified and over-specified rates of q-entries were presented. The under-specified rate, denoted by USR, indicates the proportion of q-entries in which true $q_{jk} = 1$ was estimated as $q_{jk} = 0$. The over-specified rate, denoted by OSR indicates the proportion of q-entries in which true $q_{jk} = 0$ was estimated as $q_{jk} = 1$.

3.3 Results

For each simulated dataset, one kind of rotate criterion and one kind of discrete transformation were considered in the DFL method. Thus, the DFL method generated an estimated Q-matrix for each response data, each rotation and each discrete transformation.

On the whole, results show that the DFL method can explore a Q-matrix with high CRR. First, we consider the impact of the thresholds which should be set in the discretization process. Table 1 lists the average of CRR, USR, and OSR of the q-entries for the DFL method under two kinds of Q-matrix across all conditions. We found that the DFL method using the row/column mean for discretizing the factor loadings obtains the highest CRR for the reduced Q-matrix; while the DFL method using a fixed threshold (i.e. 0.3) obtains the highest CRR for Q_3. For Q_3, the difference of CRRs between using the row/column mean and the fixed threshold (i.e. 0.3) is very small. As expected, the USRs will decrease as the thresholds decrease and the OSRs will increase as the thresholds increase. Thus the row/column mean is a good choice for the threshold.

Table 1 The average of CRR, USR, and OSR of the q-entries for the DFL method under two kinds of Q-matrix across all conditions

	Qr			Q_3		
Threshold	CRR	USR	OSR	CRR	USR	OSR
Row mean	0.720	0.214	0.067	0.866	0.063	0.072
Column mean	0.729	0.194	0.077	0.873	0.056	0.071
Total mean	0.714	0.218	0.068	0.864	0.064	0.072
0.3	0.652	0.337	0.011	0.878	0.112	0.010
0.2	0.688	0.272	0.040	0.877	0.078	0.044
0.1	0.704	0.180	0.116	0.804	0.056	0.140
0.0	0.645	0.067	0.288	0.557	0.023	0.420

Table 2 The average of CRR, USR, and OSR of the q-entries for the DFL method under two rotations, three correlations, and two Q-matrices across all conditions

		Qr			Q_3		
Correlation	Rotation	CRR	USR	OSR	CRR	USR	OSR
0.00	Promax	0.790	0.161	0.048	0.939	0.025	0.036
	Varimax	0.785	0.147	0.068	0.920	0.024	0.056
0.50	Promax	0.659	0.253	0.088	0.771	0.094	0.135
	Varimax	0.622	0.224	0.154	0.691	0.085	0.224
0.75	Promax	0.660	0.252	0.088	0.768	0.094	0.138
	Varimax	0.613	0.218	0.169	0.671	0.083	0.246

Second, the impact of the rotations and the correlations is considered. Table 2 shows the average of CRR, USR, and OSR of the q-entries for the DFL method under two rotations, three correlations, and two Q-matrices across all conditions. The results suggest that the CRRs of the Promax rotation are higher than that of the Varimax rotation regardless of the correlation and the Q-matrix. As expected, the CRRs of the Promax rotation, on average, is 6.5% higher than that of the Varimax rotation, when the underlying abilities or attributes have moderate or high correlation relationship. However, the difference of CRRs between these two rotations is relatively small. From Table 2, we found that the Q-matrix was more precisely estimated under the discrete uniform distribution than under the realistic multivariate normal threshold model. One reason for this result is that some attribute patterns contained too few examinees under the multivariate normal threshold model to identify misspecified q-vectors, noticing that if $\rho = 0.5$ or $\rho = 0.75$ was positive, then an individual with a specific attribute was more likely to have mastered the second attribute. Since the Promax rotation performs better than the Varimax, we will next only focus on results obtained from the Promax rotation.

Third, the impact of the rotations and the correlations is considered. Table 3 shows the average of CRR, USR, and OSR of the q-entries for the DFL method under three models, two levels of item quality, and two Q-matrices across all

Table 3 The average of CRR, USR, and OSR of the q-entries for the DFL method under three models, two levels of item quality, and two Q-matrices across all conditions ($\rho = 0$, rotation = Promax, threshold = column mean)

		Qr			Q_3		
Model	Item quality	CRR	USR	OSR	CRR	USR	OSR
The DINA model	High	0.861	0.100	0.039	0.999	0.000	0.001
	Low	0.741	0.173	0.085	0.980	0.003	0.017
The rRUM	High	0.934	0.059	0.007	0.982	0.003	0.015
	Low	0.863	0.130	0.007	0.924	0.038	0.038
The DINO model	High	0.879	0.096	0.025	1.000	0.000	0.000
	Low	0.723	0.183	0.094	0.977	0.004	0.019

Table 4 The average of CRR, USR, and OSR of the q-entries for the DFL method different sample sizes (ρ = 0, rotation = Promax, threhold = column mean)

	Qr			Q_3		
Sample size	CRR	USR	OSR	CRR	USR	OSR
300	0.803	0.140	0.058	0.960	0.012	0.028
500	0.829	0.127	0.044	0.980	0.007	0.013
1000	0.869	0.104	0.027	0.991	0.006	0.004

conditions. The DINA and DINO models give comparable results in terms of CRR. The rRUM is more promising than the DINA and DINO models in terms of CRR.

Finally, the impact of sample size is considered. Table 4 shows the average of CRR, USR, and OSR of the q-entries for the DFL method under different sample sizes. As expected, when sample size increases, the CRRs increase and the USRs and OSRs decrease.

4 Real Data Application

The DFL method was applied to analyse the fraction subtraction data (de la Torre 2008; Tatsuoka 1990) and the reading comprehension data (Jang 2009). For the two real data sets, the original Q-matrices were shown in Table 7 of de la Torre (2008) and in Table 3 of Jang (2009). Table 5 shows the CRR, USR, and OSR of the q-entries between the original and estimated Q-matrix under the two real data sets. The CRR, USR, and OSR of the fraction subtraction Q-matrix are highly similar to those of the reduced Q-matrix from the simulation study. The DFL method performs similar, but slightly different between the reading comprehension data and the simulated Q_3. One possible reason is that only primary skills were included in the final Q-matrix of the reading comprehension test.

Table 5 The CRR, USR, and OSR of the q-entries for the DFL method under two real data sets (rotation = Promax)

	Q[a]			Q[b]		
Threshold	CRR	USR	OSR	CRR	USR	OSR
Row mean	0.667	0.280	0.053	0.742	0.054	0.204
Column mean	0.667	0.253	0.080	0.733	0.039	0.228
Total mean	0.707	0.253	0.040	0.751	0.051	0.198
0.3	0.560	0.360	0.080	0.841	0.129	0.030
0.2	0.707	0.253	0.040	0.841	0.105	0.054
0.1	0.693	0.187	0.120	0.784	0.072	0.144
0.0	0.627	0.120	0.253	0.538	0.021	0.441

[a] Q-matrix of the fraction subtraction data

[b] Q-matrix of the reading comprehension data

5 Conclusion

Since the validation methods rely on the provisional Q-matrix which is often unknown, the DFL method is introduced based on exploratory factor analysis and response data. Results indicate that this method can mine information from data to provide a high quality provisional Q-matrix in terms of CRR. The following listed some important findings.

(a) On the whole, the results show that the DFL method can explore a Q-matrix with high CRR.
(b) The Promax rotation performs better than the Varimax rotation.
(c) The row/column mean threshold is a good choice for discretizing continuous factor loading matrix and transform it into a discrete Q-matrix.
(d) When sample size increases, the CRRs increase and the USRs and OSRs decrease.

The contributions of this study are the following. First, the DFL method is easy to implement as it is based on EFA that has been one of the most widely used statistical procedures in psychological research. Second, the DFL method can be applied to obtain a high-quality Q-matrix for the DINA model, the rRUM, and the DINO model. Finally, the proposed DFL method can provide a provisional Q-matrix for any validation methods for cognitive test developers. It was concluded that this study provided an exploratory approach for assisting subject matter experts in specifying a Q-matrix.

Some future research directions are also pointed out. One limitation of this study is that the number of factors or attributes are known in advance. It is necessary to consider how to determine the number of factors or attributes for the DFL method. Maybe one possibility to obtain the number of factors in the EFA is to consider latent class models, or use model fit tools from EFA. There exist a lot of methods for Q-matrix validation. It would be interesting to propose a method to explore and validate a Q-matrix. The second limitation is that the EFA approach taken here assumes continuous latent attributes, while the CDA approach assumes dichotomous attributes. In addition, the results depend on the arbitrary choice for the thresholds. A sensitivity analysis should ideally be considered to check the robustness of the results.

Acknowledgements This research is supported by the National Natural Science Foundation of China (Grant No. 31500909, 31360237, and 31160203), the Key Project of National Education Science "Twelfth Five Year Plan" of Ministry of Education of China (Grant No. DHA150285), the National Social Science Fund of China (Grant No. 16BYY096), the Humanities and Social Sciences Research Foundation of Ministry of Education of China (Grant No. 12YJA740057), the National Natural Science Foundation of Jiangxi (Grant No. 20161BAB212044), the Education Science Foundation of Jiangxi (Grant No. 13YB032), the Social Science Foundation of Jiangxi (Grant No. 17JY10), and the Youth Growth Fund and the Doctoral Starting up Foundation of Jiangxi Normal University. The authors would like to thank the editor Dylan Molenaar for the valuable comments on our submitted manuscript.

References

Buck, G., VanEssen, T., Tatsuoka, K. K., Kostin, I., Lutz, D., & Phelps, M. (1998). *Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I verbal sentence completion section (RR-98-23).* Princeton, NJ: Educational Testing Services.

Castellan, N. J. (1966). On the estimation of the tetrachoric correlation coefficient. *Psychometrika, 31*(1), 67–73.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*(1), 1–20.

Chang, H.-H., & Wang, W. Y. (2016). "Internet Plus" measurement and evaluation: A new way for adaptive learning. *Journal of Jiangxi Normal University (Natural Science), 40*(5), 441–455.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619–632.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618.

Chiu, C.-Y., Douglas, J. A., & Li, X.-D. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633–665.

Close, C. N. (2012). *An exploratory technique for finding the Q-matrix for the DINA model in cognitive diagnostic assessment: Combining theory with data.* Unpublished doctoral dissertation, University of Minnesota, Educational Psychology.

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19–38.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

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

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447–468.

Ding, S. L., Yang, S. Q., & Wang, W. Y. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *Journal of Jiangxi Normal University (Natural Science), 34*(5), 490–495.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49* (2), 175–186.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301–321.

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

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*(4), 262–277.

Huo, Y., & de la Torre, J. (2013, April). *Data-driven Q-matrix specification for subsequent test forms*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Im, S. (2007). *Statistical consequences of attribute misspecfication of the rule space model.* Unpublished doctoral dissertation, Columbia University.

Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement, 71*(4), 712–731.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion model application to language assessment. *Language Testing, 26*(1), 31–73.

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

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187–200.

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

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*(7), 548–564.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli, 19*(5A), 1790–1817.

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2016). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement, 77*(2), 220–240.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*(3), 200–217.

McGlohen, M. K. (2004). *The application of cognitive diagnosis and computerized adaptive testing to a large-scale assessment.* Unpublished doctorial dissertation, University of Texas at Austin.

McGlohen, M. K., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*(3), 808–821.

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78–96.

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

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

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessments* (pp. 327–359). Hillsdale: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Taylor & Francis Group.

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

Tu, D.-B., Cai, Y., & Dai, H.-Q. (2012). A new method of Q-matrix validation based on DINA model. *Acta Psychologica Sinica, 44*(4), 558–568.

Xing, D., & Hambleton, R. K. (2004). Test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*(1), 5–21.

Identifiability of the Latent Attribute Space and Conditions of Q-Matrix Completeness for Attribute Hierarchy Models

Hans-Friedrich Köhn and Chia-Yi Chiu

Abstract Educational researchers have argued that a realistic view of the role of attributes in cognitively diagnostic modeling should account for the possibility that attributes are not isolated entities, but interdependent in their effect on test performance. Different approaches have been discussed in the literature; among them the proposition to impose a hierarchical structure so that mastery of one or more attributes is a prerequisite of mastering one or more other attributes. A hierarchical organization of attributes constrains the latent attribute space such that several proficiency classes, as they exist if attributes are not hierarchically organized, are no longer defined, because the corresponding attribute combinations cannot occur with the given attribute hierarchy. Hence, the identification of the latent attribute space is often difficult—especially, if the number of attributes is large. As an additional complication, constructing a complete Q-matrix may not at all be straightforward if the attributes underlying the test items are supposed to have a hierarchical structure. In this article, the conditions of identifiability of the latent space if attributes are hierarchically organized and the conditions of completeness of the Q-matrix are studied.

Keywords Cognitive diagnosis · Attribute hierarchy · Latent attribute space
Q-matrix · Completeness · DINA model

1 Introduction

Cognitive diagnosis (CD), a relatively recent development in educational measurement (DiBello et al. 2007; Haberman and von Davier 2007; Leighton and Gierl 2007; Nichols et al. 1995; Rupp et al. 2010) explicitly targets mastery of the instructional

H.-F. Köhn (✉)
University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
e-mail: hkoehn@illinois.edu

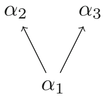
C.-Y. Chiu
Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA
e-mail: cychiu@gse.rutgers.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_30

content and seeks to provide immediate feedback to students about their strengths and weaknesses in terms of skills learned and skills needing study. CD terminology refers to skills, specific knowledge, aptitudes—any cognitive characteristic required to perform tasks—collectively as "attributes" (denoted by α_k, $k = 1, 2, \dots, K$) that an examinee may or may not possess. CD models—or "Diagnostic Classification Models" (DCMs), as they are called here—describe an examinee's ability as a composite of these attributes, each of which an examinee may or may not have mastered. Mastery/non-mastery of attributes is recorded as a binary string; different 0-1-combinations define attribute profiles of distinct proficiency classes (denoted by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_K)'$). The entire set of realizable attribute profiles $\mathcal{L}$, given a particular set of attributes, is called the latent attribute space (Tatsuoka 2009). Modeling educational testing data within the CD framework involves estimating the item parameters and assigning examinees to proficiency classes—that is, estimating their individual attribute profiles $\boldsymbol{\alpha}$.

Educational researchers have argued that a realistic view of the role of attributes in cognitively diagnostic modeling should account for the possibility that attributes are not isolated entities, but interdependent in their effect on test performance. Different approaches have been discussed in the literature. de la Torre and Douglas (2004) proposed a higher-order model linking the latent attribute space to an underlying multivariate normal distribution with possibly correlated dimensions. Haertel and Wiley (1994) and Leighton et al. (2004) (see also Leighton and Gierl 2007; Tatsuoka 2009; Templin and Bradshaw 2014) developed a different approach to account for potential relations/interdependencies among attributes by imposing a hierarchical structure so that mastery of one or more attributes is a prerequisite of mastering one or more other attributes. These DCMs are commonly referred to as Attribute Hierarchy Models (AHMs).

As an example, consider the divergent attribute hierarchy among α_1, α_2, and α_3—mastery of attribute α_1 is a prerequisite of mastering attributes α_2 and α_3:



Several complications can arise from imposing a hierarchy on the attributes. First, a hierarchical organization of attributes constrains the latent attribute space such that several proficiency classes, as they exist if attributes are not hierarchically organized, are no longer defined because the corresponding attribute combinations cannot occur with the given attribute hierarchy. Hence, the identification of the latent attribute space $\mathcal{L}$ is often difficult—especially, if the number of attributes is large. Second, constructing a complete Q-matrix may not at all be obvious. In this article, the conditions of identifiability of the latent space if attributes are hierarchically organized, and the conditions of completeness of the Q-matrix are studied.

2 Completeness of the Q-Matrix

The items of a test are also characterized by individual attribute profiles that determine which specific attributes are required to respond correctly to an item. The entire set of these item-attribute associations constitutes the Q-matrix of a test (Tatsuoka 1985). The Q-matrix must be known (or the data cannot be analyzed within the CD framework) and complete. A Q-matrix is said to be complete if it guarantees the identifiability of all realizable proficiency classes among examinees (Chiu et al. 2009; Köhn and Chiu 2017). An incomplete Q-matrix causes examinees to be assigned to proficiency classes to which they do not belong. Formally, a Q-matrix is complete if the equality of two expected item response vectors, $\mathbf{S}(\boldsymbol{\alpha})$ and $\mathbf{S}(\boldsymbol{\alpha}^*)$, implies that the underlying attribute profiles, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, are also identical: $\mathbf{S}(\boldsymbol{\alpha}) = \mathbf{S}(\boldsymbol{\alpha}^*) \Rightarrow \boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, where $\mathbf{S}(\boldsymbol{\alpha}) = E(\mathbf{Y} \mid \boldsymbol{\alpha})$, and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_J)'$ is the vector of observed item responses.

Take the **D**eterministic **I**nput **N**oisy “**A**ND” Gate (DINA) model (Haertel 1989; Junker and Sijtsma 2001; Macready and Dayton 1977) as an example. The item response function (IRF) of the DINA model is

$$P(Y_j = 1|\boldsymbol{\alpha}, s_j, g_j) = (1 - s_j)^{\eta_j} g_j^{(1-\eta_j)}$$

(for succinctness, the examinee index i is omitted if the context permits). Hence, the J entries in $\mathbf{S}(\boldsymbol{\alpha})$ are defined as

$$S_j(\boldsymbol{\alpha}) = E(Y_j \mid \boldsymbol{\alpha}) = \begin{cases} g_j & \text{if } \eta_j = 0 \\ 1 - s_j & \text{if } \eta_j = 1 \end{cases} \text{ with } 1 - s_j > g_j$$

where $\eta_j = 0, 1$ is the conjunction parameter indicating whether an examinee has mastered all attributes required for correctly responding to item j; s_j, g_j are item parameters formalizing the probabilities of “slipping”, $s_j = P(Y_j = 0|\eta_j = 1)$ (i.e., failing item j despite the ability to solve it—“having a bad day”) and “guessing”, $g_j = P(Y_{ij} = 1|\eta_{ij} = 0)$ (i.e., solving item j, but lacking the required attributes). Consider the two Q-matrices $\mathbf{Q}_{1:3}$ and $\mathbf{Q}_{4:6}$, with rows representing $J = 3$ items and columns $K = 3$ attributes:

$$\mathbf{Q}_{1:3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{Q}_{4:6} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Their completeness for the DINA model is evaluated by computing the expected item response profiles $\mathbf{S}(\boldsymbol{\alpha})$:

$\boldsymbol{\alpha}$	$\mathbf{Q}_{1:3}$			$\mathbf{Q}_{4:6}$		
	$\boldsymbol{q}_1 = (100)$	$\boldsymbol{q}_2 = (010)$	$\boldsymbol{q}_3 = (001)$	$\boldsymbol{q}_4 = (011)$	$\boldsymbol{q}_5 = (101)$	$\boldsymbol{q}_6 = (110)$
	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$	$S_3(\boldsymbol{\alpha})$	$S_4(\boldsymbol{\alpha})$	$S_5(\boldsymbol{\alpha})$	$S_6(\boldsymbol{\alpha})$
(000)	g_1	g_2	g_3	g_4	g_5	g_6
(100)	$1 - s_1$	g_2	g_3	g_4	g_5	g_6
(010)	g_1	$1 - s_2$	g_3	g_4	g_5	g_6
(001)	g_1	g_2	$1 - s_3$	g_4	g_5	g_6
(110)	$1 - s_1$	$1 - s_2$	g_3	g_4	g_5	$1 - s_6$
(101)	$1 - s_1$	g_2	$1 - s_3$	g_4	$1 - s_5$	g_6
(011)	g_1	$1 - s_2$	$1 - s_3$	$1 - s_4$	g_5	$1 - s_6$
(111)	$1 - s_1$	$1 - s_2$	$1 - s_3$	$1 - s_4$	$1 - s_5$	$1 - s_6$

$\mathbf{Q}_{1:3}$ is complete for the DINA model, whereas $\mathbf{Q}_{4:6}$ is not because, for example, $\boldsymbol{\alpha}_1 = (000) \neq \boldsymbol{\alpha}_2 = (100)$, but $\mathbf{S}(\boldsymbol{\alpha}_1) = \mathbf{S}(\boldsymbol{\alpha}_2)$.

For tests with a large number of items involving multiple attributes, completeness of the Q-matrix is often difficult to establish. As an additional complication, completeness is not an intrinsic property of the Q-matrix, but can only be assessed in reference to a specific DCM supposed to underlie the data. In the extreme, the Q-matrix of a particular test can be complete for one model, but incomplete for another.

How is Q-completeness affected if the attributes are hierarchically organized? Consider again the divergent attribute hierarchy described earlier. For the DINA model, none of the two Q-matrices, $\mathbf{Q}_{1:3}$ and $\mathbf{Q}_{4:6}$, is complete if the divergent hierarchy is imposed on α_1, α_2, and α_3, because Items 2, 3, and 4 are no longer admissible, as they do not include α_1 as a required attribute. But α_1 is a prerequisite for attributes α_2 and α_3—casually speaking, α_2 and α_3 "cannot be had" without α_1:

$\boldsymbol{\alpha}$	$\mathbf{Q}_{1:3}$			$\mathbf{Q}_{4:6}$		
	$\boldsymbol{q}_1 = (100)$	$\boldsymbol{q}_2 = (010)$	$\boldsymbol{q}_3 = (001)$	$\boldsymbol{q}_4 = (011)$	$\boldsymbol{q}_5 = (101)$	$\boldsymbol{q}_6 = (110)$
	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$	$S_3(\boldsymbol{\alpha})$	$S_4(\boldsymbol{\alpha})$	$S_5(\boldsymbol{\alpha})$	$S_6(\boldsymbol{\alpha})$
(000)	g_1	g_2	g_3	g_4	g_5	g_6
(100)	$1 - s_1$	g_2	g_3	g_4	g_5	g_6
(010)	g_1	$1 - s_2$	g_3	g_4	g_5	g_6
(001)	g_1	g_2	$1 - s_3$	g_4	g_5	g_6
(110)	$1 - s_1$	$1 - s_2$	g_3	g_4	g_5	$1 - s_6$
(101)	$1 - s_1$	g_2	$1 - s_3$	g_4	$1 - s_5$	g_6
(011)	g_1	$1 - s_2$	$1 - s_3$	$1 - s_4$	g_5	$1 - s_6$
(111)	$1 - s_1$	$1 - s_2$	$1 - s_3$	$1 - s_4$	$1 - s_5$	$1 - s_6$

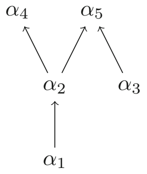
3 The Latent Attribute Space $\mathcal{L}$

If the K binary attributes underlying a test do not have a hierarchical structure, then the latent attribute space $\mathcal{L}$, the set of all realizable attribute profiles (characterizing distinct proficiency classes), contains 2^K elements (Tatsuoka 2009). The case of $K =$ 3 attributes, α_1, α_2, and α_3, may serve as a simple example; then, $\mid \mathcal{L} \mid = 2^K = 8$:

No.	$\boldsymbol{\alpha} = (\alpha_1\ \alpha_2\ \alpha_3)$
1	(0 0 0)
2	(1 0 0)
3	(0 1 0)
4	(0 0 1)
5	(1 1 0)
6	(1 0 1)
7	(0 1 1)
8	(1 1 1)

But if the three attributes have, say a divergent hierarchy as shown earlier—that is, attribute α_1 is a prerequisite of mastering attributes α_2 and α_3—then, $\mathcal{L}$ consists of only five proficiency classes: (000), (100), (110), (101) and (111), because the proficiency classes (010), (001), and (011) are no longer defined.

Here is a more complex example of an attribute hierarchy involving $K = 5$ attributes; they are organized in the hierarchy displayed by the following tree graph:



Without a hierarchy imposed on the attributes, the latent attribute space $\mathcal{L}$ would contain $2^5 = 32$ proficiency classes; but due to the complex prerequisite structure, most of these theoretically realizable proficiency classes are not defined. For example, the single-attribute profiles $(01000) = \mathbf{e}_2$, $(00010) = \mathbf{e}_4$, and $(00001) = \mathbf{e}_5$ are no longer defined and must be replaced by (11000), (11010), and (11101), respectively, so that the prerequisite relations defining the hierarchy among attributes are satisfied. In fact, out of the original 32 proficiency classes, only ten proficiency classes are defined, and $\mathcal{L}$ is reduced to

No.	$\boldsymbol{\alpha} = (\alpha_1\ \alpha_2\ \alpha_3\ \alpha_4\ \alpha_5)$
1	0 0 0 0 0
2	0 0 1 0 0
3	1 0 0 0 0
4	1 0 1 0 0
5	1 1 0 0 0
6	1 1 1 0 0
7	1 1 0 1 0
8	1 1 1 1 0
9	1 1 1 0 1
10	1 1 1 1 1

4 Attribute Hierachies: A Lattice-Theoretic Approach

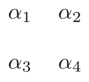
The last two sections described in greater detail the challenges posed by AHMs:

(a) How to identify the latent attribute space $\mathcal{L}$ for attribute hierarchies with a complex prerequisite structure?
(b) How to identify a complete Q-matrix for attribute hierarchies with a complex prerequisite structure?

Extant approaches are ad hoc and merely descriptive. Alternatively, a general approach based on lattice theory is developed here that accommodates attribute hierarchy models as well as DCMs with no attribute hierarchy. The following definitions are needed:

(1) The K attributes $\alpha_k \in \{0, 1\}$ are Boolean variables.
(2) The 2^K vectors $\boldsymbol{\alpha}$ are called Boolean vectors.
(3) The order relation $\leq$ for two binary K-dimensional attribute vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ is defined such that $\boldsymbol{\alpha}^* \leq \boldsymbol{\alpha}$ if and only if $\alpha_k^* \leq \alpha_k \;\; \forall k$. The relation $\leq$ is reflexive, antisymmetric, and transitive; hence, it is a partial order.
(4) $\mathcal{L}$ is called a partially ordered set (poset) written $\langle \mathcal{L}, \leq \rangle$.
(5) Consider Boolean vectors $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ and $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_K^*)$. Their infimum is defined as the conjunction: $\inf\{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*\} = \boldsymbol{\alpha} \wedge \boldsymbol{\alpha}^* = \boldsymbol{\alpha} \cdot \boldsymbol{\alpha}^* = (\alpha_1\alpha_1^*, \ldots, \alpha_K\alpha_K^*)$ (recall $0 \wedge 1 = 0$). Their supremum is defined as the disjunction: $\sup\{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*\} = \boldsymbol{\alpha} \vee \boldsymbol{\alpha}^* = \boldsymbol{\alpha} + \boldsymbol{\alpha}^* = (\alpha_1 + \alpha_1^*, \ldots, \alpha_K + \alpha_K^*)$ (recall $0 \vee 1 = 1$).
(6) Boolean operators $\wedge$ and $\vee$ are idempotent (e.g., $1 \vee 1 = 1$). Thus, the infimum and supremum of the singletons $\{\boldsymbol{\alpha}_m\}$ are just $\inf\{\boldsymbol{\alpha}_m\} = \sup\{\boldsymbol{\alpha}_m\} = \boldsymbol{\alpha}_m$—for example: $\inf\{\boldsymbol{\alpha}_1\} = \sup\{\boldsymbol{\alpha}_1\} = (00, \ldots, 0)$, $\inf\{\boldsymbol{\alpha}_M\} = \sup\{\boldsymbol{\alpha}_M\} = (11, \ldots, 1)$.
(7) $\mathcal{L}$ as a lattice: $\langle \mathcal{L}, \leq \rangle$ is called a lattice if each of its (finite) subsets has an infimum and a supremum.
(8) A lattice is called complete if it has universal bounds—that is, a least element 0 and a largest element I. For $\mathcal{L}$, $O = \mathbf{0}_K = \boldsymbol{\alpha}_1 = (00, \ldots, 0)$ and $I = \mathbf{1}_K = \boldsymbol{\alpha}_M = (11, \ldots, 1)$.

Consider first the case where attributes have no hierarchy; for example, the $K = 4$ attributes shown in the following graph (the absence of directed edges indicates that attributes lack a hierarchy):



The latent attribute space $\mathcal{L}$ is defined by $2^K = 16 = M$ realizable proficiency classes:

No.	$\boldsymbol{\alpha} = (\alpha_1\ \alpha_2\ \alpha_3\ \alpha_4)$
1	(0 0 0 0)
2	(1 0 0 0)
3	(0 1 0 0)
4	(0 0 1 0)
5	(0 0 0 1)
5	(1 1 0 0)
6	(1 0 1 0)
7	(1 0 0 1)
8	(0 1 1 0)
9	(0 1 0 1)
10	(0 0 1 1)
11	(1 1 1 0)
12	(1 1 0 1)
13	(1 0 1 1)
14	(0 1 1 1)
15	(1 1 1 0)
16	(1 1 1 1)

The lattice $\mathcal{L}$ can be displayed as a Hasse diagram that has the following properties:

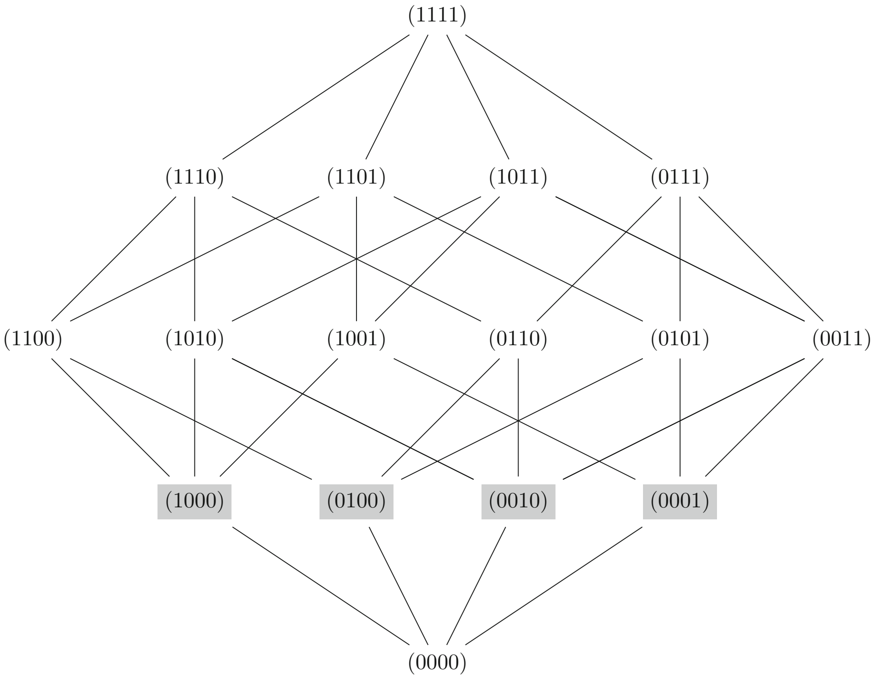
(a) The proficiency classes are vertically ordered and connected by an edge if they are in the relation $\leq$.

(b) Because order relations are transitive, any relation between proficiency classes can be deduced by following the edges upward.

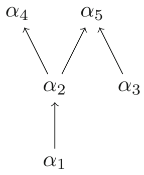
(c) The infimum and supremum of every subset of $\mathcal{L}$ are also in $\mathcal{L}$.

(d) All realizable proficiency classes can be obtained through the Boolean operations $\wedge$ and $\vee$ performed on the four single-attribute profiles $(1000) = \mathbf{e}_1$, $(0100) = \mathbf{e}_2$, $(0010) = \mathbf{e}_3$, and $(0001) = \mathbf{e}_4$.

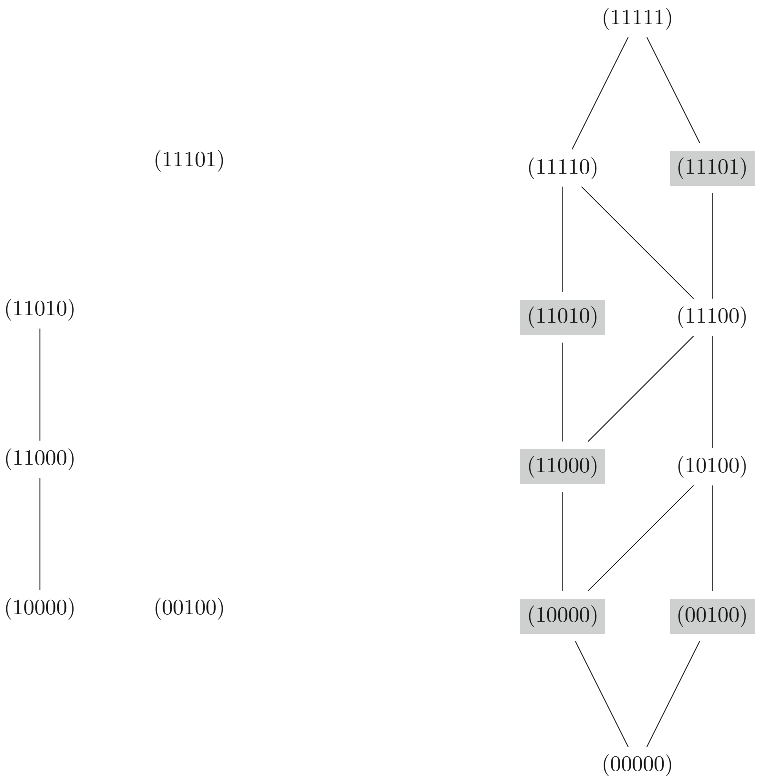
(e) The vectors $\mathbf{e}_1, \ldots, \mathbf{e}_4$ are called basic attribute vectors; they can be derived from the graph of the four attributes directly by inspection; they are underlain by grey boxes in the Hasse diagram below.



Second, to study the case of attributes having a hierarchical structure, consider again the earlier example of a hierarchy involving $K = 5$ attributes, the tree graph of which is repeated here as a convenience:



From the tree graph, the five basic attribute vectors, (10000), (11000), (00100), (11010), and (11101), are obtained by inspection and arranged as an incomplete lattice (left panel below). The lattice of the entire latent attribute space $\mathcal{L}$ is reconstructed from these attribute vectors; for example, (10000) $\vee$ (00100) = (10100), or (11000) $\vee$ (00100) = (11100); note that (10000) $\wedge$ (00100) = (11000) $\wedge$ (00100) = (00100) $\wedge$ (11010) = (00000) (right panel below).



5 Key Results and Discussion

(1) A set of attribute profiles called "basic attribute vectors" can be derived by inspection from the tree graph of any attribute hierarchy.
(2) These basic attribute vectors are a subset of the latent attribute space $\mathcal{L}$.
(3) $\mathcal{L}$ is a lattice; hence, the latent attribute space can be reconstructed in its entirety from the basic attribute vectors using the operations $\wedge$ and $\vee$.
(4) Any Q-matrix that contains the unique $K \times K$ submatrix formed by the basic attribute vectors is complete for the DINA model, given any attribute hierarchy.

To demonstrate Claim (4), consider again the last example, with a hierarchy involving $K = 5$ attributes. The Q-matrix derived from the basic attribute vectors is

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

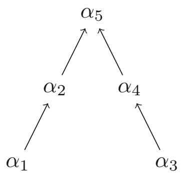
Computing the expected item response vectors $\mathbf{S}(\boldsymbol{\alpha})$ confirms its completeness:

$\boldsymbol{\alpha}$	**Q**				
	$\boldsymbol{q}_1 = (10000)$	$\boldsymbol{q}_2 = (11000)$	$\boldsymbol{q}_3 = (00100)$	$\boldsymbol{q}_4 = (11010)$	$\boldsymbol{q}_5 = (11101)$
	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$	$S_3(\boldsymbol{\alpha})$	$S_4(\boldsymbol{\alpha})$	$S_5(\boldsymbol{\alpha})$
(00000)	g_1	g_2	g_3	g_4	g_5
(00100)	g_1	g_2	$1 - s_3$	g_4	g_5
(10000)	$1 - s_1$	g_2	g_3	g_4	g_5
(11000)	$1 - s_1$	$1 - s_2$	g_3	g_4	g_5
(10100)	$1 - s_1$	g_2	$1 - s_3$	g_4	g_5
(11100)	$1 - s_1$	$1 - s_2$	$1 - s_3$	g_4	g_5
(11010)	$1 - s_1$	$1 - s_2$	g_3	g_4	g_5
(11110)	$1 - s_1$	$1 - s_2$	$1 - s_3$	g_4	g_5
(11101)	$1 - s_1$	$1 - s_2$	$1 - s_3$	$1 - s_4$	$1 - s_5$
(11111)	$1 - s_1$	$1 - s_2$	$1 - s_3$	$1 - s_4$	$1 - s_5$

In conclusion, one might ask why not extend the results of this study to general DCMs? (General DCMs have received considerable attention as a framework for expressing the specific functional relation between attribute mastery and the probability of a correct item response of diverse DCMs in a unified mathematical form and parameterization; von Davier 2005, 2008; Henson et al. 2009; de la Torre 2011). First, the procedure for identifying the latent attribute space suggested by the theoretical results of this study can also be applied to general DCMs. Second, however, the results on Q-completeness for the DINA model when attributes have a hierarchical structure do not apply to general DCMs as well. The complex parameterization of general DCMs causes anomalies in Q-completeness that cannot be accounted for at present. As an illustration, consider a relatively simple instance of a general DCM, de la Torre's (2011) Additive Cognitive Diagnosis Model (A-CDM). The IRF of the A-CDM is defined as a linear combination of K attributes

$$P(Y_j = 1 \mid \boldsymbol{\alpha}) = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_{ik}$$

where q_{jk} indicates whether mastery of attribute α_k is required for item j. (Additional constraints on the coefficients β_{jk}—not described here—guarantee $0 \leq P(Y_j = 1 \mid \boldsymbol{\alpha}) \leq 1$.) Suppose a test involves $K = 5$ attributes that have a convergent hierarchy, as shown in the following tree graph:



For the DINA model, the five basic vectors that form the unique complete Q-matrix can be derived by inspection from the tree graph (see matrix **Q** below).

$$\mathbf{Q} = \begin{pmatrix} 1\,0\,0\,0\,0 \\ 1\,1\,0\,0\,0 \\ 0\,0\,1\,0\,0 \\ 0\,0\,1\,1\,0 \\ 1\,1\,1\,1\,1 \end{pmatrix} \quad \mathbf{Q}_{1:2} = \begin{pmatrix} 1\,1\,0\,0\,0 \\ 1\,1\,1\,1\,1 \end{pmatrix} \quad \mathbf{Q}_{3:5} = \begin{pmatrix} 0\,0\,1\,0\,0 \\ 1\,1\,1\,0\,0 \\ 1\,1\,1\,1\,1 \end{pmatrix}$$

Q is still complete for the A-CDM, as can be shown by computing the vectors of expected item responses $\mathbf{S}(\boldsymbol{\alpha})$ (for the A-CDM, the expected response to item j is $S_j(\boldsymbol{\alpha}) = P(Y_j \mid \boldsymbol{\alpha}) = \beta_{j0} + \sum_{k=1}^{5} \beta_{jk}\alpha_k q_{jk}$). However, **Q** is no longer uniquely complete for the A-CDM, as the example of the two Q-matrices $\mathbf{Q}_{1:2}$ and $\mathbf{Q}_{3:5}$ shows. The vectors of expected item responses for the A-CDM corresponding to $\mathbf{Q}_{1:2}$ and $\mathbf{Q}_{3:5}$ are:

$\boldsymbol{\alpha}$	$\mathbf{Q}_{1:2}$	
	$\mathbf{q}_1 = (11000)$	$\mathbf{q}_2 = (11111)$
	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$
(00000)	β_{10}	β_{20}
(10000)	$\beta_{10} + \beta_{11}$	$\beta_{20} + \beta_{21}$
(11000)	$\beta_{10} + \beta_{11} + \beta_{12}$	$\beta_{20} + \beta_{21} + \beta_{22}$
(00100)	β_{10}	$\beta_{20} + \beta_{23}$
(10100)	$\beta_{10} + \beta_{11}$	$\beta_{20} + \beta_{21} + \beta_{23}$
(11100)	$\beta_{10} + \beta_{11} + \beta_{12}$	$\beta_{20} + \beta_{21} + \beta_{22} + \beta_{23}$
(00110)	β_{10}	$\beta_{20} + \beta_{23} + \beta_{24}$
(10110)	$\beta_{10} + \beta_{11}$	$\beta_{20} + \beta_{21} + \beta_{23} + \beta_{24}$
(11110)	$\beta_{10} + \beta_{11} + \beta_{12}$	$\beta_{20} + \beta_{21} + \beta_{22} + \beta_{23} + \beta_{24}$
(11111)	$\beta_{10} + \beta_{11} + \beta_{12}$	$\beta_{20} + \beta_{21} + \beta_{22} + \beta_{23} + \beta_{24} + \beta_{25}$

$\boldsymbol{\alpha}$	$\mathbf{Q}_{3:5}$		
	$\mathbf{q}_3 = (00100)$	$\mathbf{q}_4 = (11100)$	$\mathbf{q}_5 = (11111)$
	$S_3(\boldsymbol{\alpha})$	$S_4(\boldsymbol{\alpha})$	$S_5(\boldsymbol{\alpha})$
(00000)	β_{30}	β_{40}	β_{50}
(10000)	β_{30}	$\beta_{40} + \beta_{41}$	$\beta_{50} + \beta_{51}$
(11000)	β_{30}	$\beta_{40} + \beta_{41} + \beta_{42}$	$\beta_{50} + \beta_{51} + \beta_{52}$
(00100)	$\beta_{30} + \beta_{33}$	$\beta_{40} + \beta_{43}$	$\beta_{50} + \beta_{53}$
(10100)	$\beta_{30} + \beta_{33}$	$\beta_{40} + \beta_{41} + \beta_{43}$	$\beta_{50} + \beta_{51} + \beta_{53}$
(11100)	$\beta_{30} + \beta_{33}$	$\beta_{40} + \beta_{41} + \beta_{42} + \beta_{43}$	$\beta_{50} + \beta_{51} + \beta_{52} + \beta_{53}$
(00110)	$\beta_{30} + \beta_{33}$	$\beta_{40} + \beta_{43}$	$\beta_{50} + \beta_{53} + \beta_{54}$
(10110)	$\beta_{30} + \beta_{33}$	$\beta_{40} + \beta_{41} + \beta_{43}$	$\beta_{50} + \beta_{51} + \beta_{53} + \beta_{54}$
(11110)	$\beta_{30} + \beta_{33}$	$\beta_{40} + \beta_{41} + \beta_{42} + \beta_{43}$	$\beta_{50} + \beta_{51} + \beta_{52} + \beta_{53} + \beta_{54}$
(11111)	$\beta_{30} + \beta_{33}$	$\beta_{40} + \beta_{41} + \beta_{42} + \beta_{43}$	$\beta_{50} + \beta_{51} + \beta_{52} + \beta_{53} + \beta_{54} + \beta_{55}$

Remarkably, $\mathbf{Q}_{1:2}$ is complete although it contains only two items. Two items, however, do not automatically guarantee completeness for the A-CDM, as the columns $\mathbf{q}_4 = (11100)$ and $\mathbf{q}_5 = (11111)$ of $\mathbf{Q}_{3:5}$ demonstrate. Completeness depends on the specific item parameter values and is not guaranteed because, for example, $\beta_{41} = \beta_{43}$ and $\beta_{51} = \beta_{53}$ cannot be ruled out so that for $\boldsymbol{\alpha} = (10000)$ and $\boldsymbol{\alpha} = (00100)$, $\boldsymbol{S}_{4:5}(10000) = \boldsymbol{S}_{4:5}(00100)$ is possible. Thus, a Q-matrix formed just by $\mathbf{q}_4$ and $\mathbf{q}_5$ cannot be guaranteed to be complete. The inconclusiveness of $\mathbf{q}_4$ and $\mathbf{q}_5$ can be resolved by adding item $\mathbf{q}_3 = (00100)$—indeed, $\mathbf{Q}_{3:5}$ is complete. (Note that the alternative additions (00110) and (11111), instead of $\mathbf{q}_3$, would also make $\mathbf{Q}_{4:5}$ guaranteed complete). In summary, this small-scale example of the A-CDM shows that for a given attribute hierarchy there are at least three different ways to construct a complete Q-matrix for a general DCM. At present, it seems not possible to identify common rules of Q-completeness for general DCMs that can be used with any attribute hierarchy.

References

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665.

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

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

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

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skill diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 1031–1038). Amsterdam: Elsevier.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.

Haertel, E. H., & Wiley, D. E. (1994). Representation of ability structure: Implications for testing. In N. Fredriksen, R. Mislevy, & I. Bejar (Eds.), *Testing theory for a new generation of tests* (pp. 359–384). Hillsdale, NJ: Erlbaum.

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

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

Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika*, *82*, 112–132.

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

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

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *33*, 379–416.

Nichols, P. D., Chipman, S. E., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Mahwah, NJ: Erlbaum.

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

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconception in the pattern classification approach. *Journal of Educational and Behavioral Statistics*, *12*, 55–73.

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

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*, 317–339.

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

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–301.

Different Expressions of a Knowledge State and Their Applications

Shuliang Ding, Fen Luo, Wenyi Wang, Jianhua Xiong,
Heiqiong Duan and Lihong Song

Abstract Based on the Augment algorithm, any column of Q matrix can be expressed as a Boolean union of some columns of reachability matrix R, but the expression is not unique. There are two different expressions for a column of the reduced Q matrix, say x, a redundant expression of x and a concise expression of x. When a test length is short, the redundant expression of a knowledge state can be used to simplify the proof of an important property of the reachability matrix R in the design of cognitive diagnostic test, and provides a novel method to specify Q matrix. This specification method can be employed to deal with the polytomous Q matrix.

Keywords The Augment algorithm · Redundant expression · Concise expression · Specification of Q matrix · Polytomous Q matrix

S. Ding (✉) · F. Luo · W. Wang · J. Xiong
School of Computer and Information Engineering, Jiangxi Normal University,
99 Ziyang Road, Nanchang, Jiangxi, People's Republic of China
e-mail: ding06026@163.com

F. Luo
e-mail: luofen312@163.com

W. Wang
e-mail: hicosdor@aliyun.com; wenyiwang@jxnu.edu.cn

J. Xiong
e-mail: xiongjianhua1212@qq.com

H. Duan
School of Foreign Languages, Nanchang Hangkong University,
696 South Fenghe Road, Nanchang, Jiangxi, People's Republic of China
e-mail: englishduan2011@163.com

L. Song
Elementary Educational College, Jiangxi Normal University, 99 Ziyang Road,
Nanchang, Jiangxi, People's Republic of China
e-mail: viviansong1981@163.com

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_31

1 Some Concepts and Symbols

For Boolean matrices, any nonzero knowledge state is a column of a reduced Q matrix Q_r (Tatsuoka 1995, 2009) and it can be expressed as a Boolean union of the columns of the reachability matrix R based on the Augment algorithm (Ding et al. 2008, 2016) or on the incremental augment algorithm (Yang et al. 2010). There may be several expression forms (Ding et al. 2017). If x is a column of reduced Q matrix Q_r, let S_x = { r | (r is a column of R) and (r $\leq$ x)}, then the Boolean union of all elements in the set of S_x is called as the redundant expression of x, where r $\leq$ x means that every element of x-r is non-negative. Let U_x be a subset of S_x and any two different elements in the set U_x have no prerequisite relationship, i.e., they are not comparable. At this time, the Boolean union of all elements in U_x is called as the concise expression of x (Ding et al. 2017). Suppose that a, b are any two elements in S_x and a < b, a is deleted from S_x, then the set U_x can be obtained through comparing any pair elements in S_x and deleting the smaller element from S_x. The set S_x is called as the set of redundant expression of x (SREx) in this paper. The elements in S_x and U_x are called vectors or knowledge states hereafter.

The following Lemma 1 and Theorem 1 have been proved (Ding et al. 2017).

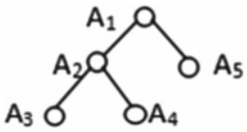
Lemma 1 *The number of the redundant expression of x is equal to the sum of the elements in x.*

Example 1 R and Q_p are reachability matrixes and potential Q-matrix (the reduced Q-matrix) corresponding to a divergent structure in Fig. 1, respectively.

$$Q_p = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix} = [r_1\, r_2\, r_3\, r_4\, r_5\, q_6\, q_7\, q_8\, q_9\, q_{10}]$$

where, q_7 = $r_3 \vee r_4$ (the concise expression of q_7) = $r_1 \vee r_2 \vee r_3 \vee r_4$ (the redundant expression of q_7) and $q_7 = (11110)^T$ contains four 1 and one 0. The number of the combination components in the redundant expression of q_7 is 4. It means that there are 4 elements in S_x and x = q_7.

Fig. 1 Divergent structure



$$R = \begin{bmatrix} 11111 \\ 01110 \\ 00100 \\ 00010 \\ 00001 \end{bmatrix} = [r_1\, r_2\, r_3\, r_4 r_5]$$

Theorem 1 *Suppose that $\alpha \in Q_p$ and $\boldsymbol{\alpha} = \vee_{j=1}^{h} r_{i_j}$ is the redundant expression of α, then $\boldsymbol{\alpha} = \vee_{j=1}^{h} r_{i_j} = \sum_{j=1}^{h} \mathrm{e}_{i_j}$, $\sum_{j=1}^{h} \mathrm{e}_{i_j} = \vee_{j=1}^{h} \mathrm{e}_{i_j}$, where $(e_1, e_2, \ldots, e_k)$ is the column partition of the identity matrix.*

Notice that in Theorem 1, the subscripts of the combinational components in the redundant expression of x are equal to column indices of corresponding to the columns in the identity matrix such that $\mathrm{x} = \vee_{j=1}^{h} \mathrm{e}_{i_j}$; that is, x is expressed by the linear combination of columns of the identity matrix.

Example 2 (Cont. Ex.1). $q_7 == \mathrm{r}_1 \vee \mathrm{r}_2 \vee \mathrm{r}_3 \vee \mathrm{r}_4$ (the redundant expression of q_7) $= \mathrm{e}_1 + \mathrm{e}_2 + \mathrm{e}_3 + \mathrm{e}_4$.

An interesting relation between the redundant expression of x and its concise expression.

Property 1 *The two kinds of expressions of x are equal, if and only if the set U_x contains one element, and at this time, the concise expression of x must be a column of the identity matrix.*

Proof If $\mathrm{U_x}$ contains two different elements, say a and b, and a and b do not compare to each other, so $\mathrm{c} = \mathrm{a} \vee \mathrm{b}$ does not equal to a, nor to b, and c belongs to $\mathrm{S_x}$. So two sets $\mathrm{U_x}$ and $\mathrm{S_x}$ are not equal. If $\mathrm{U_x} = \mathrm{S_x}$, then $\mathrm{U_x}$ contains one vector, say u, u must be a column of the identity matrix. If it is not true, suppose that u is a vector of the reachability matrix R and u contains at least two non-zero elements. There is another column of R, and $\mathrm{v} < \mathrm{u}$, and u, v belongs to $\mathrm{S_x}$. This is contrary to the statement $\mathrm{U_x} = \mathrm{S_x}$. If u is augmented from some columns of R, it is obvious that $\mathrm{U_x}$ is not equal to $\mathrm{S_x}$. If $\mathrm{U_x}$ only contains a column of the identity matrix, then $\mathrm{U_x} = \mathrm{S_x}$ by the definition of $\mathrm{U_x}$.

2 Applications

2.1 To Simplify the Proof of a Theorem

The definitions of redundant and concise expression may be applied to generate personalized learning paths for different learning style learners and personalized remedy route. And the definitions can be applied to prove the fact that under some conditions, the reachability matrix R or the equivalent class of R plays an important role in the design of cognitive diagnostic testing. Some other interesting applications of these definitions are discussed as following.

Theorem 2 (Ding et al. 2010, 2011) *Supposed a 0–1 scoring rubric is adopted and the attributes are non-compensatory. Let $\alpha \circ Q$ be the expected examinee response vector of knowledge state (attribute mastery pattern) α for a test Q-matrix Q. If R is the test Q-matrix Q (i.e. R = Q), then for any knowledge state α satisfying α $R = \alpha^T$. Otherwise, if R is a submatrix of the test Q-matrix Q, and α_1, α_2 are different knowledge states, then $\alpha_1 \circ Q \neq \alpha_2 \circ Q$.*

Proof Take the redundant expression of α, $\alpha = \vee_{j=1}^{h} r_{i_j}$, then for any $r_{i_j} \in \{r_{i1}, r_{i2}, \ldots, r_{ih}\}$, $r_{i_j} \leq \alpha$, so $\alpha \circ r_{i_j} = 1$, from Theorem 1 and $\alpha \circ R = (\alpha \circ r_1, \alpha \circ r_2, \ldots, \alpha \circ r_k) = \sum_{j=1}^{h} e_{i_j}^{T} = \alpha^T$.

The result is old (Ding et al. 2010, 2011, 2017), but the proof is new.

2.2 Specify Q Matrix Under Ideal Response Situation

Theorem 1 may be applied to specify unknown elements in a new item if all of the columns in the reachability matrix R are specified correctly under ideal response conditions. Namely, there are no slipping nor guessing in the observable response patterns when examinees take a test. To specify the elements in Q matrix, please follow the steps of this method:

Step 1. Suppose the items corresponding to the columns of R and the new item which will be answered by examinees.

Step 2. Choose examinees whose responses on the new item are correct, and collect their responses to the items corresponding to all columns in the reachability matrix, say $y_1, y_2, \ldots, y_n$.

Step 3. Calculate the hierarchical consistency index (Cui and Leighton 2009) based on the attribute hierarchy (HCI) and the responses $y_1, y_2, \ldots, y_n$, Delete the responses with lower HCI (say smaller than 0.9) from the set $\{y_1, y_2, \ldots, y_n\}$. Denote the remaining be $z_1, z_2, \ldots, z_t$.

Step 4. Let z be equal to the Boolean conjunction of $z_1, z_2, \ldots, z_t$, then x = z.

If $y_1, y_2, \ldots, y_n$ are n ideal response patterns, then the algorithm listed as above (the step 3 can be omitted) can be proven based on Theorem 1. Even if n = 1, x still equals to the first K components of y_1.

Example 3 (Cont.Ex.1.). If $x = q_6$, then if i is in {6, 8, 9, 10}, $q_i \circ x = 1$, i = 6, 8, 9, 10, then $x = q_6 \Lambda q_8 \Lambda q_9 \Lambda q_{10}$ and $q_6 \leq q_i, i = 6, 8, 9, 10,$ so $x = q_6$.

In fact, we have the following theorem.

Suppose that b ∈ R or $b \in Q_p$ denote that b is a column of R or b is a column of Q_p, respectively.

Let $x \in Q_p$ be a unknown vector.

For any $b \in Q_p$, let

$$S_b(x) = \{x | (r \in R) \wedge (r \leq b) \wedge (x \leq b)\}$$

$S_b(x)$ represents the set of components in the redundant expression of b which satisfies $x \leq b$, then

Theorem 3 $\cap_{b\in Q_p} S_b(x) = \{r|(r \in R)\wedge(r\le x)\}$ and $x = \vee r_{r\in\{r|(r\in R)\wedge(r\le x)\}}$

Proof If $x\le b$, then $S_x(x)\subseteq S_b(x)$

$$\therefore S_x(x)\subseteq \cap_{b\in Q_p} S_b(x)$$

Notice that x $\in$ Q_p and $x\le x$, so

$$\cap_{b\in Q_p} S_b(x)\subseteq S_x(x)$$

and $(x\le x)$ is true, so

$$S_x(x)=\{r|(r\in R)\wedge(r\le x)\wedge(x\le x)\}$$

From the definition of the redundant expression of x, then

$$x = \bigvee_{r\in S_x(x)} r$$

The condition of Theorem 3 is rigorous because it requires all responses to x being ideal responses. We know that the observed responses are not satisfying.

3 Generalizing the Results to Polytomous Q Matrix

A Q-matrix is called as a polytomous Q-matrix if each of its element is a non-negative integer. Some modifications are made. For example, the Boolean union is replaced by wise-element MAX-operator. Then some analogues for the polytomous Q matrix (e.g., Chen and de la Torre 2013; Sun et al. 2013) are given. For example, the sum of all elements in a knowledge state, say x, is equal to the number of the vectors in the redundant expression of x. And if the scoring rubric format is changed, the importance of the quasi-reachability matrix (Ding et al. 2016) in a design of cognitive diagnostic testing is proved by using a polytomous Q matrix.

Theorem 4 *Suppose α is a column in a polytomous Q matrix and its redundant expression is $\alpha = \vee_{t=1}^{h} r_{it}$, then the sum of all elements in α equals to h, which is the number of the combinational components of α.*

For the polytomous Q matrix, it is interesting that under ideal response situation, the analogous to Theorem 4 can be obtained. Suppose that x is a column of the polytomous potential Q matrix and its elements are unknown, the test Q matrix, Q_t, is a pile of the quasi-reachability matrix, denoting as R_p and x. That is to say, Q_t = (R_p | x), and if x <= α, then α ∘ x = 1.

4 Discussion

Unfortunately, there are some slippages in the observed response patterns. Because the quality of items is not satisfying, the HCI must be calculated for deleting some observed response patterns, and the ORPs with HCI approaching to 1 are chosen to calculate their Boolean conjunction.

In ideal response patterns, researchers can identify new items' attribute vectors perfectly even if their sample size is small. For the ordinary observable response patterns, accurate identification result is near to 0.5 after Monte Carlo simulations. This fact calls some theory or method to deal with the random errors in the observable response patterns.

Acknowledgements This research is supported by the National Natural Science Foundation of China (Grant No. 31360237, 31500909, and 31160203), the National Social Science Fund of China (Grant No. 16BYY096, 13BYY087), the Humanities and Social Sciences Research Foundation of Ministry of Education of China (Grant No. 13YJC880060 and 12YJA740057), the National Natural Science Foundation of Jiangxi Province (Grant No. 20161BAB212044), Jiangxi Education Science Foundation (Grant No. 13YB032), the Science and Technology Research Foundation of Education Department of Jiangxi Province (Grant No. GJJ13207, GJJ150356, GJJ160282), the China Scholarship Council (CSC No. 201509470001), and the Youth Growth Fund and the Doctoral Starting up Foundation of Jiangxi Normal University.

References

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419–437.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(4), 429–449.

Ding, S.-L., Luo, F., Cai, Y., Lin, H.-J., & Wang, X.-B. (2008). Complement to Tatsuoka's Q matrix theory. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 417–424). Tokyo: Universal Academy Press.

Ding, S.-L., Luo, F., Wang, W.-Y., & Xiong, J.-H. (2016). Dichotomous and polytomous Q matrix theory. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J.A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research*. *Springer Proceedings in Mathematics & Statistics*, *167* (277–288). https://doi.org/10.1007/978-3-319-38759-8_21.

Ding, S.-L., Wang, W.-Y., Luo, F., Xiong, J.-H., & Meng, Y.-R. (2017). Irreplaceability of a reachability matrix. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology* (pp. 229–238). Cham, Switzerland: Springer.

Ding, S.-L., Wang, W.-Y., & Yang, S.-Q. (2011). The design of cognitive diagnostic test blueprints. *Journal of Psychological Science, 34*(2), 258–265.

Ding, S.-L., Yang, S.-Q., & Wang, W.-Y. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *Journal of Jiangxi Normal University (Natural Science Edition), 34*(5), 490–494.

Sun, J., Xin, T., Zhang, S., & Jimmy, D. L. T. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement, 37*(7), 503–521.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern classification approach. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessments*. Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Taylor & Francis Group.

Yang, S.-Q., Ding, S.-L., & Ding, Q.-L. (2010). Incremental augment algorithm based on reduced Q matrix. *Transaction of Nanjing University of Aeronautics & Astronautics, 27,* 183–189.

Accuracy and Reliability of Autoregressive Parameter Estimates: A Comparison Between Person-Specific and Multilevel Modeling Approaches

Siwei Liu

Abstract This simulation study compares the person-specific (PS) and multilevel modeling (MLM) approaches in the accuracy and reliability of autoregressive (AR) parameter estimates when data are generated from a first-order AR model and the functional form of the analytic model is correctly specified. Influences of a variety of factors on accuracy and reliability are examined, including time series length, sample size, the distribution of the AR coefficients, and the variability of the AR coefficients. Neither sample size nor distribution has an effect on accuracy or reliability. MLM generally has better accuracy than PS at both the population level and the individual level. However, in MLM, individuals who deviate farther from the sample mean are modeled less accurately than individuals who are closer to the sample mean. The two approaches do not differ in the reliability of the AR estimates. For both approaches, higher variability in the AR coefficients is associated with higher reliability. Implications on modeling practices are discussed.

Keywords Autoregressive model • Person-specific • Multilevel modeling
Accuracy • Reliability

1 Introduction

With increasing popularity of intensive longitudinal data in psychology, time series models have gained enormous attention in recent years. One widely used and general model for time series analysis is the autoregressive (AR) model, which, in the univariate case, describes the temporal dependency of one variable on itself in the form of lagged regression. For instance, the first-order AR model (i.e., maximal lag = 1) is often used in emotion research to examine the emotional regulatory ability of individuals (Hamaker and Grasman 2015; Jongerling et al. 2015; Kuppens

S. Liu (✉)
University of California at Davis, 1318 Hart Hall, 301 Shields Avenue,
Davis, CA 95616, USA
e-mail: sweliu@ucdavis.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_32

et al. 2010). In this model, the AR coefficient represents the extent to which a person's affect at the current time point depends on his/her own affect at the previous time point. Hence, a larger coefficient is often interpreted as an index of emotional inertia, an indicator of psychological maladjustment (Hamaker and Grasman 2015; Jongerling et al. 2015; Kuppens et al. 2010). In developmental research, the AR coefficients have been used to represent stability of household income during the early years of development. Higher stability can predict greater educational attainment later in life, which conforms to the life history theory (Li et al. in prep; Nettle et al. 2013). In other areas of psychology, the AR model has been used to study substance use (Rovine and Walls 2006; Zheng et al. 2013), stress reactivity (Liu et al. 2013), and brain connectivity (Ding et al. 2006; Liu and Molenaar 2016), just to name a few.

Given the popularity of the AR model, the ability of different modeling approaches to recover the underlying AR mechanisms in time series data becomes an important issue. Currently in psychology, AR models are typically estimated using one of two approaches. One is the person-specific (PS) modeling approach, where an AR model is fitted to one individual's data at a time, and inferences at the population level are drawn based on the empirical distributions of the estimated person-specific AR parameters (Bollen and Curran 2006). The other approach is multilevel modeling (MLM), where an AR model is fitted to data from a sample of individuals simultaneously assuming a common AR pattern (e.g., first-order), but individuals are allowed to vary in the magnitude of their AR coefficients. The differences between the individual AR coefficients and the sample mean are known as *random effects*. They are usually assumed to be normally distributed with mean zero, and can be estimated using Empirical Bayes (EB) methods (Verbeke and Molenberghs 2000b).

Conceptually, the PS and MLM approaches have several crucial differences. With MLM, researchers have to assume that all individuals' data can be described by the same functional form, such as a first-order AR model. In contrast, the PS approach allows individuals to have idiographic AR patterns, such as AR models with different numbers of lags. Hence, PS may be particularly suitable for modeling highly heterogeneous dynamic processes. On the other hand, multilevel models are estimated by pooling information across individuals, whereas with PS, only one individual's information is used for each model. Therefore, MLM may be more suitable for making population level inferences when limited information is available per person, such as when the number of measurement occasions is small.

In a previous study (Liu 2017), I simulated data to compare the two approaches in accurately recovering the AR parameters at both the population level and the individual level. I investigated the influences of *sample heterogeneity*, *time series length* (*T*), *sample size* (*N*), and the *distribution* of the AR coefficients on the accuracy of AR parameter estimates. I found that when the sample was relatively homogeneous, MLM generally outperformed PS at both levels, regardless of *T*, *N*, and *distribution*. When the sample is heterogeneous, such that different individuals are characterized by AR processes with different numbers of lags, the relative performance of the two approaches depends heavily on *T*, with PS more sensitive to the number of measurement occasions. These findings provided important implications for research design and model selection with intensive longitudinal data.

The current study is a follow-up investigation of Liu (2017), and it extends the previous study in two ways. First, I aim to examine the influences of an additional factor—*variability* of AR coefficients (hereafter referred to as $\sigma_{a_1}^2$)—on the performance of the two approaches. This investigation would help elucidate how the two approaches compare given various amounts of individual differences in the magnitude of the AR coefficients. Specifically, I hypothesize that $\sigma_{a_1}^2$ would not affect the performance of PS, but would influence the EB estimates of MLM. Because EB estimates are posterior estimates based on a prior normal distribution, they are known to be biased towards the sample mean (i.e., the "shrinkage effect"; Verbeke and Molenberghs 2000b). Hence, I speculate that smaller *variability* in the true AR coefficients will be associated with higher accuracy in the EB estimates. The second extension of the current study is in the outcome measures. Because the individual level AR estimates are often used as predictors in further analyses (Kuppens et al. 2012; Li et al. in prep), I aim to examine the reliability of these estimates in addition to their accuracy. Similar to the accuracy measures (to be introduced later), I will examine how the reliability is affected by *analytic approach* (PS vs. MLM), *T*, *N*, the *distribution* of the AR coefficients, and $\sigma_{a_1}^2$.

2 Method

2.1 Simulation

Data are simulated according to a first-order AR model. To control for the influence of the intercept on the estimation of the AR parameters, the following model with no intercept is used, where $i = 1, \ldots, N$ represents individual, and $t = 1, \ldots, T$ represents time point:

$$y_{it} = a_{1i} y_{i(t-1)} + \varepsilon_{it}. \tag{1}$$

The population mean of the AR coefficients a_{1i} is fixed to 0.30, and its variance, $\sigma_{a_1}^2$, is fixed to 0.01 in this simulation. The variance of ε_{it} is set to 1 for all individuals. A three-way factorial design is used to examine the effects of *T*, *N*, and the *distribution* of the AR coefficients. Both *T* and *N* have three levels, 20, 50, and 100. The factor *distribution* also has three levels, with a_{1i} generated from a normal distribution, a uniform distribution, or a symmetric bimodal distribution consisting of two equal-variance normal distributions.[1] For each condition, 1000 time series

[1] I also simulated data following a highly skewed distribution, which was not used in Liu (2017). It did not show any large effect on the results. Hence, it is excluded in this paper to facilitate comparison between the two studies.

are generated. Importantly, the design of the current simulation is identical to that in Liu (2017), with the only exception that in the previous study, $\sigma^2_{a_1}$ was fixed to 0.10, ten times the value used here. Hence, although $\sigma^2_{a_1}$ is not a factor manipulated in the current simulation, a comparison between this study and Liu (2017) will provide insights to the influence of AR coefficients variability.

2.2 Analysis and Outcome Measures

To compare results between PS and MLM with the correct model specification, each data set is analyzed with both methods assuming a zero-intercept first-order AR pattern. The PS models are estimated using the *ar* function in R (R Core Team 2015) with ordinary least square (OLS) estimation, and the multilevel models are estimated using the *nlme* package in R (Pinheiro et al. 2016).

Three outcome measures are considered. *Population level accuracy* is assessed by comparing the estimated population mean, $\hat{\mu}_{a_1}$, to the true population mean, 0.30. *Individual level accuracy* is evaluated using the *mean square error* (MSE) of the individual AR coefficients:

$$MSE_{a_1} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{a}_{1i} - a_{1i}\right)^2 \tag{2}$$

where N represents sample size, a_{1i} is the true AR coefficient for individual i, and $\hat{a}_{1i}$ is the estimated AR coefficient for the same individual. Hence, a smaller MSE_{a_1} indicates higher accuracy, averaged across individuals. Finally, *reliability* is evaluated using the square of the Pearson correlation between the estimated AR coefficients ($\hat{a}_{1i}$) and the true values (a_{1i}).

Repeated-measures analysis of variance (RM-ANOVA; Myers 1979) is used with data simulated in the current study, in which *analytic approach* (PS vs. MLM) is the within-subject factor, and T, N, and the *distribution* are the between-subjects factors. Because the p-values in these models are influenced by the number of replications, which is arbitrary in a simulation study, the importance of an effect will be evaluated instead using the effect size measure, η^2, which indicates the proportion of variance explained by an effect. In the following, only results with at least a medium effect size ($\eta^2 \geq 0.06$) will be reported (Cohen 1988). For each outcome measure, similarity and differences between results from the current study and the previous study (Liu 2017) are highlighted to assess the influence of $\sigma^2_{a_1}$.

3 Results

3.1 Population Level Accuracy

For population level accuracy, the results resemble those in the previous study in which the AR coefficients had the same mean but a larger variance (Liu 2017). Specifically, none of the effects involving N or *distribution* has an $\eta^2 \geq 0.06$. However, there is a large main effect of *analytic approach* ($\eta^2 = 0.83$), and an interaction effect between *analytic approach* and T ($\eta^2 = 0.64$). Table 1 shows the average estimates of the population mean, the standard deviations of the estimated means, the average standard errors, and the coverage rates of the true value in the 95% confidence intervals for the two approaches, broken down by T. The MLM estimates are almost identical to the true value, 0.30, regardless of T. In contrast, the PS estimates are negatively biased, especially with short time series data. When there are only 20 measurement occasions, PS, on average, produces a relative bias (i.e., the ratio of bias over the true value) of 9%. This number reduces to 4% with 50 measurement occasions, and 2% with 100 measurement occasions. The average standard errors from PS tend to be smaller than those from MLM, although the differences are tiny. Accordingly, with 100 measurement occasions, both PS and MLM have good coverage rates of the true value in their 95% confidence intervals. With fewer measurement occasions, MLM still has satisfactory coverage rates, but PS does not. Importantly, the estimated population means and coverage rates reported here are almost identical to those found in Liu (2017), suggesting that the variability of the AR coefficients does not affect the accuracy of the population level estimates. This is consistent with the MLM literature that population level inferences are generally robust (Verbeke and Molenberghs 2000a).

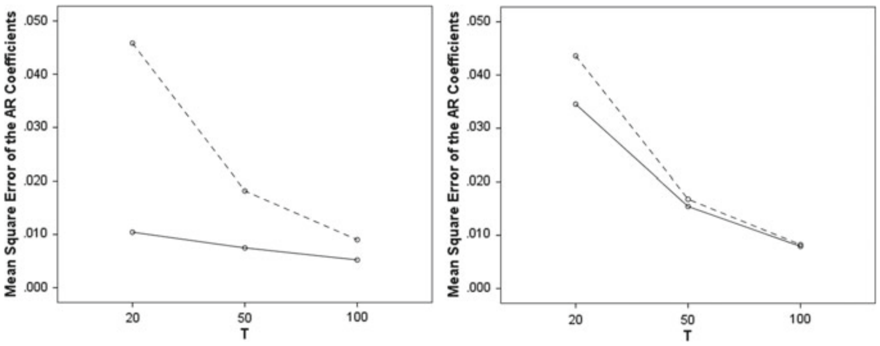
Table 1 Population level accuracy in the AR(1) coefficients by the *analytic approach* and *time series length*, averaged over the factors *sample sizes* and *distribution*. MLM = multilevel modeling; PS = person-specific

Time series length	Estimate (standard deviation)		Standard error		95% Confidence interval coverage rate	
	MLM	PS	MLM	PS	MLM	PS
20	0.2994 (0.0385)	0.2734 (0.0368)	0.0375	0.0356	0.95	0.86
50	0.2998 (0.0261)	0.2885 (0.0254)	0.0260	0.0254	0.96	0.93
100	0.3000 (0.0208)	0.2943 (0.0205)	0.0213	0.0210	0.95	0.95

3.2 Individual Level Accuracy

Results for the accuracy of the individual level estimates are also similar to those in the previous study (Liu 2017). There are main effects of *analytic approach* ($\eta^2 = 0.87$) and T ($\eta^2 = 0.83$), as well as an interaction effect between the two ($\eta^2 = 0.82$). As shown in Fig. 1a, MLM (solid line) generally produces smaller MSE_{a_1} than PS (dotted line), and the difference is more dramatic when T is small. This interaction pattern is consistent with results from the previous study with a larger $\sigma^2_{a_1}$, which is showed in Fig. 1b. However, a comparison between the two figures reveals an impact of $\sigma^2_{a_1}$ on the accuracy of MLM. Specifically, a smaller $\sigma^2_{a_1}$ is associated with smaller MSE_{a_1} from MLM. In other words, when individuals are more similar to one another in the magnitude of their AR coefficients, MLM produces more accurate EB estimates, averaged across individuals.

This finding is consistent with the well-known "shrinkage effect" in the Bayesian literature, which refers to the phenomenon that individual EB estimates are shrunken towards the prior average profile. In the current context, the prior distribution for the random effects is a normal distribution with mean zero. Hence, individuals whose AR coefficients deviate from the sample mean are shrunken towards the sample mean. In addition, the strength of this effect depends on the amount of deviation, with larger deviations associated with greater shrinkage. Hence, the average MSEs from MLM in Fig. 1a are smaller than those in Fig. 1b. To further illustrate this, I compute the correlation between the individual MSE (Eq. 2) and the absolute difference between an individual's true AR coefficient and the sample mean. As shown in Fig. 2, there are indeed positive associations between the two for the MLM approach. Because a larger T gives a heavier weight to the empirical data in comparison to the prior distribution, the shrinkage effect, as



(A) $\sigma^2_{a_1} = 0.01$ (B) $\sigma^2_{a_1} = 0.10$

Fig. 1 Mean square error of the AR coefficients by the *analytic approach* and *time series length*, averaged over the factors *sample size* and *distribution*. Solid line = multilevel modeling. Dotted line = person-specific. T = time series length

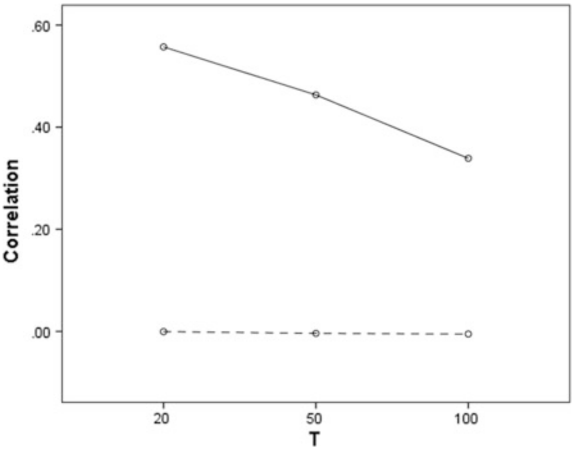


Fig. 2 Correlations between individual MSE and the absolute difference between the individual AR coefficients and the sample mean, broken down by the factors *analytic approach* and *time series length*. Solid line = multilevel modeling. Dotted line = person-specific. T = time series length

indicated by the strength of the above correlation, decreases with increasing T. However, even with 100 measurement occasions, the average correlation is above 0.30, indicating at least a medium effect size. In contrast, no such association is present for the PS approach.

3.3 Reliability of the Autoregressive Coefficients

Lastly, I compare the two approaches on the reliability of the individual AR coefficients by examining the squared correlation between the estimated and true values. When $\sigma^2_{a_1} = 0.01$, the only effect that reaches a medium effect size is the main effect of T ($\eta^2 = 0.62$). With $T = 20$, the average squared correlation between the true and estimated AR coefficients is 0.18. This value increases to 0.35 when $T = 50$, and 0.52 when $T = 100$. Notably, the reliability is much higher when $\sigma^2_{a_1} = 0.10$. Specifically, the average squared correlations are 0.67 for $T = 20$, 0.85 for $T = 50$, and 0.92 for $T = 100$.

4 Discussions

This study compares the PS and MLM approaches in the accuracy and reliability of the AR parameter estimates when data are generated based on a first-order AR model, which is correctly specified in the analysis (i.e., no model misspecification). Various factors that may affect their relative performance are examined, including time series length (T), sample size (N), and the *distribution* of the AR coefficients. In addition, a comparison of results from the current study and a previous study with similar design (Liu 2017) provides insights to the influence of the AR coefficients variability ($\sigma^2_{a_1}$).

Consistent with findings in the previous study (Liu 2017), MLM in general produces more accurate AR parameter estimates than PS, both at the population and individual levels. Whereas variability of the AR coefficients has no impact on the PS approach, smaller variability is associated with higher individual level accuracy for MLM. In other words, in terms of individual level accuracy, MLM benefits from the scenario where individuals are more similar to one another. It is important to note, however, that with the MLM approach, the accuracy for a specific individual also depends on how similar that individual is to the average profile in the sample. Specifically, higher accuracy is to be expected for individuals whose true values are closer to the sample mean. For individuals who deviate farther from the mean, the discrepancy between the estimated AR coefficients and the true values may be large with small numbers of measurement occasions. In terms of reliability, no large difference is found between PS and MLM. However, both approaches are affected by $\sigma^2_{a_1}$ and T, such that higher variability and longer time series length are associated with higher reliability.

These results provide several practical implications. Specifically, when choosing between the PS and MLM approaches, researchers need to consider the purpose of their research and the characteristics of the data. For example, if researchers are concerned with the accuracy at the population level, such as in examinations of the stability in a variable over time, or when evaluating the effect of a treatment on emotional regulation by comparing the AR coefficients across different groups, MLM is preferred over PS because it produces less bias as well as higher coverage rates of the true values. However, if the goal is to extract the individual AR coefficients and use them as predictors in further analyses, having high reliability in these estimates is most critical. In this case, researchers can choose either approach because they produce similar reliability. However, caution should be used if the variability in the AR coefficients is small. With $\sigma^2_{a_1} = 0.01$, for example, it is generally not a good idea to treat the AR estimates as predictors because they are likely to contain a large amount of estimation error, which may lead to bias in the next step. In contrast, with $\sigma^2_{a_1} = 0.1$, such modeling procedure may be acceptable, especially when the number of measurement occasions is large. Lastly, if the goal of research is to identify individuals whose AR coefficients exceed a certain threshold so that they are eligible for a treatment or intervention, neither approach seems ideal. Although MLM, on average, has higher individual level accuracy than PS, its performance declines as individuals deviate farther and farther away from the sample mean. In other words, it performs worst when accuracy is needed the most. Future research needs to examine whether this effect may be alleviated by assuming a prior distribution with flatter tails in the EB estimation.

It should be noted that the recommendations above are provided based on the assumption that researchers are working with a homogeneous sample, where all individuals can be characterized by an AR pattern with the same number of lags, and the analytic model is correctly specified to match the underlying data generating mechanisms. In reality, the sample may be heterogeneous, in which case the MLM approach may be less ideal (Liu 2017). In addition, the results presented here should be interpreted taking into account the limitations of the study. Like all simulation research, the current study is limited in the factors considered and the range of parameters used to simulate the data. For instance, although a variety of factors are included in the simulation, the population mean of the AR parameters and the residual variance are fixed to be constants. It has not been studied whether and how these factors may affect the performance of the two approaches. In addition, in this simulation I do not include a measurement model. However, the reliability of the variable is likely to affect the performance of both approaches, another factor that needs to be investigated in the future.

Despite these limitations, the current study extends and complements the previous study, which was the first to directly compare PS and MLM, the two most commonly used approaches for analyzing intensive longitudinal data in psychology. Together, they provide unique contributions to the literature by simultaneously considering model performance at both the population level and individual level. This comes in time as psychology as a whole is moving towards more intensive data collection, allowing individuals to be "brought back" to scientific psychology (Molenaar 2004; Molenaar and Campbell 2009).

References

Bollen, K. A., & Curran, P. J. (2006). Unconditional latent curve model. In K. A. Bollen & P. J. Curran (Eds.), *Latent curve models: A structural equation perspective* (pp. 16–57). Hoboken, NJ: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Ding, M., Chen, Y., & Bressler, S. L. (2006). Granger causality: Basic theory and application to neuroscience. In B. Schelter, M. Winterhalder, & J. Timmer (Eds.), *Handbook of time series analysis: Recent theoretical developments and applications* (pp. 437–460). Weinheim, Germany: Wiley-VCH.

Hamaker, E. L., & Grasman, R. P. P. P. (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology, 5,* 1–15. https://doi.org/10.3389/fpsyg.2014.01492.

Jongerling, J., Laurenceau, J.-P., & Hamaker, E. L. (2015). A multilevel AR(1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research, 50*(3), 334–349. https://doi.org/10.1080/00273171.2014.1003772.

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science, 21*(7), 984–991. https://doi.org/10.1177/0956797610372634.

Kuppens, P., Sheeber, L. B., Yap, M. B. H., Whittle, S., Simmons, J. G., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. *Emotion, 12*(2), 283–289. https://doi.org/10.1037/a0025046.

Li, Z., Belsky, J., & Liu, S. (in prep). Modeling income unpredictability via multilevel autoregressive model: Impact of time-series length on reliability of parameter estimates and prediction power.

Liu, S. (2017). Person-specific versus multilevel autoregressive models: Accuracy in parameter estimates at the population and individual levels. *British Journal of Mathematical and Statistical Psychology, 70,* 480–498. https://doi.org/10.1111/bmsp.12096.

Liu, S., & Molenaar, P. C. M. (2016). Testing for Granger causality in the frequency domain: A phase resampling method. *Multivariate Behavioral Research, 51*(1), 53–66. https://doi.org/10.1080/00273171.2015.1100528.

Liu, S., Rovine, M. J., Klein, L. C., & Almeida, D. M. (2013). Synchrony of diurnal cortisol pattern in couples. *Journal of Family Psychology, 27*(4), 579–588. https://doi.org/10.1037/a0033735.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective, 2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science, 18*(2), 112–117.

Myers, J. L. (1979). *Fundamentals of research design*. New York, NY: Allyn and Bacon.

Nettle, D., Frankenhuis, W. E., & Rickard, I. J. (2013). The evolution of predictive adaptive responses in human life history. *Proceedings of the Royal Society B: Biological Sciences, 280* (1766). https://doi.org/10.1098/rspb.2013.1343.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2016). nlme: Linear and nonlinear mixed effects models (Version 3.1-126). Retrieved from http://CRAN.R-project.org/package=nlme.

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Rovine, M. J., & Walls, T. A. (2006). Multilevel autoregressive modeling of interindividual differences in the stability of a process. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 124–147). New York: Oxford University Press.

Verbeke, G., & Molenberghs, G. (2000a). Estimation of the marginal model. In G. Verbeke & G. Molenberghs (Eds.), *Linear mixed models for longitudinal data* (pp. 41–54). New York, NY: Springer.

Verbeke, G., & Molenberghs, G. (2000b). Inference for the random effects. In G. Verbeke & G. Molenberghs (Eds.), *Linear mixed models for longitudinal data* (pp. 77–92). New York, NY: Springer.

Zheng, Y., Wiebe, R. P., Cleveland, H. H., Molenaar, P. C. M., & Harris, K. S. (2013). An idiographic examination of day-to-day patterns of substance use craving, negative affect, and tobacco use among young adults in recovery. *Multivariate Behavioral Research, 48*(2), 241–266. https://doi.org/10.1080/00273171.2013.763012.

A Two-Factor State Theory

John Tisak, Guido Alessandri and Marie S. Tisak

Abstract When studying longitudinal phenomena, the notions of traits and states can be a useful classification. Specifically, traits represent basic human characteristics that have a permanency or enduring property, while on the other hand, states are environmental or ephemeral that are more time specific. Admittedly, research often focuses on traits and the relationships of these traits to other important variables. Moving in a different direction, this contribution focuses on the more ephemeral aspects of longitudinal variables, that is, states. A very practical justification for this direction is model fit indices. A probably more important rationale for expanding the state model is to obtain a more accurate reflection of the situation under study. To establish a common foundation, a longitudinal factor analytic model and a latent curve model are presented. Next, a statistical model of the ephemeral effects or state, which is analogous to Spearman's Two-Factor Theory is given. Lastly, a substantive illustration demonstrates the worthwhileness of this Two-Factor State Theory.

Keywords Traits · States · Longitudinal factor analysis · Latent curve analysis

J. Tisak (✉)
Department of Psychology, Bowling Green State University,
Bowling Green, OH 43403, USA
e-mail: jtisak@bgsu.edu

G. Alessandri
Sapienza – Università di Roma, Rome, Italy
e-mail: guido.alessandri@uniroma.it

M. S. Tisak
Bowling Green State University, Bowling Green, USA
e-mail: mtisak@bgsu.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_33

1 A Two-Factor State Theory

1.1 Introduction

When studying longitudinal phenomena, the notions of traits and states (Lord and Novick 1968, pp. 27–28) can be a useful classification. Specifically, traits represent basic human characteristics that have a permanency or enduring property, while on the other hand, states are environmental or ephemeral that are more time specific (Tisak and Tisak 2000). Admittedly, research often focuses on traits and the relationships of these traits to other important variables. In addition, one might be interested in the decomposition of observed measure variance into trait, state, and error variances for any psychological variable (Alessandri et al. 2012).

Moving in a different direction, this contribution focuses on the more ephemeral aspects of longitudinal variables, that is, states. A very practical justification for this direction is model fit indices. Concretely, the modeling of states may improve the acceptability of one's statistical model or more precisely one's Structural Equations Model (SEM) without the inclusion of "nuisance" parameters. A probably more important rationale for expanding the state model is to obtain a more accurate reflection of the situation under study. Parenthetically, this circumstance is analogous to the dichotomy between common and specific factors in classical factor analysis (Thurstone 1947). To improve model fit, one could include additional common factors, however, this approach might lead to theoretical unimportant factors, which could reflect undesirable or nuisance features of the items, such as sentence length.

To establish a common foundation, a longitudinal factor analytic model and a latent curve model are presented. Since these models are well established, the exposition will be terse. However, to facilitate an understanding, the common notation used in LISREL (Jöreskog and Sörbom 1996) is used. Next, a statistical model of the ephemeral effects or state, which is both conceptually and structurally analogous to Spearman's Two-Factor Theory (Spearman 1904) is given. Concretely, in Spearmen's Two-Factor Theory, there is a general or *g*-factor that is common to all the items, and there are specific factors, which are unique to each item. Analogously, in the proposed Two-Factor State Theory, there is a temporal/general or *t*-factor that is present at each time point and that impacts each of the factors or saliences. Additionally, there are temporal-specific effects, which are unique to each factor or salience at each time point. Lastly, a substantive illustration demonstrates the worthwhileness of this Two-Factor State Theory.

1.2 A Basic Longitudinal Factor Analytic Model (FAM)

In this and the following two sections, three related longitudinal models are presented. The first is a longitudinal factor model (FAM), which is a standard measurement model with measured variates, latent variables or factors, and measurement errors. The second or latent curve model (LCA) restricts each of the longitudinal factors to have a specific structure and includes temporal effects at this second-level. Finally, the third latent curve model with state structure (LCA-S) permits the usually uncorrelated state effects to be correlated.

Initially, consider a basic longitudinal factor analytic model (Tisak and Meredith 1989):

$$\mathbf{y}^{(k)} = \boldsymbol{\tau}_{\mathbf{y}} + \Lambda_{\mathbf{y}}\boldsymbol{\eta}^{(k)} + \boldsymbol{\epsilon}^{(k)}, \tag{1}$$

where $k = 1, 2, \ldots, g$ indicates the populations or groups. $\mathbf{y}^{(k)}$ is an observed random vector of size mp (m is the number of measurement periods and p is the number of variables). The unobserved random vectors, $\boldsymbol{\eta}^{(k)}$ and $\boldsymbol{\epsilon}^{(k)}$, are of size mr and mp, respectively. Here r indicates the number of factors at each time point. The intercepts, $\boldsymbol{\tau}_{\mathbf{y}}$, and slopes, $\Lambda_{\mathbf{y}}$, have dimensions $mp \times 1$ and $mp \times mr$, respectively. Notice that both the intercepts and slopes exhibit the property of stationarity (invariance across time) and invariance across populations.

Concretely,

$$\boldsymbol{\tau}_{\mathbf{y}} = [1_m \otimes \boldsymbol{\tau}] \text{ and } \Lambda_{\mathbf{y}} = [\mathrm{I}_m \otimes \boldsymbol{\lambda}],$$

where 1_m and I_m are a unit vector and identity matrix of size m; $\otimes$ is the kronecker product.

For this first-order model, the Means and Covariance Structure (MACS) are

$$\boldsymbol{\mu}_{\mathbf{y}}^{(k)} = \boldsymbol{\tau}_{\mathbf{y}} + \Lambda_{\mathbf{y}}\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(k)} \text{ and } \Sigma_{\mathbf{y}}^{(k)} = \Lambda_{\mathbf{y}}\Sigma_{\boldsymbol{\eta}}^{(k)}\Lambda_{\mathbf{y}}^{'} + \Theta_{\boldsymbol{\epsilon}}^{(k)}, \tag{2}$$

where $\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(k)}$ is a $mr \times 1$ mean vector and $\Sigma_{\boldsymbol{\eta}}^{(k)}$ is a $mr \times mr$ covariance matrix of the first-order factors. Lastly, $\Theta_{\boldsymbol{\epsilon}}^{(k)}$ is a $mp \times mp$ covariance matrix of the unique factors. More specifically, in this longitudinal situation it has the following form.

$$\Theta_{\boldsymbol{\epsilon}}^{(k)} = \begin{bmatrix} \Theta_{11}^{(k)} & \cdots & \Theta_{1m}^{(k)} \\ \vdots & \ddots & \vdots \\ \Theta_{m1}^{(k)} & \cdots & \Theta_{mm}^{(k)} \end{bmatrix},$$

where $\Theta_{tt'}^{(k)}$ is a diagonal matrix of uniqueness of size p with $t = 1, 2, \ldots, m$.

1.3 A Basic Latent Curve Model (LCM)

Next, consider a basic latent curve model (Meredith and Tisak 1990) that contains both traits and states:

$$\boldsymbol{\eta}^{(k)} = \boldsymbol{\alpha} + \Gamma\boldsymbol{\xi}^{(k)} + \boldsymbol{\zeta}^{(k)}, \tag{3}$$

where $\boldsymbol{\alpha}$ is a vector of temporal effects that impacts everyone in the same fashion; $\boldsymbol{\zeta}^{(k)}$ are individual temporal or state influences; and $\boldsymbol{\xi}^{(k)}$ is a set of individual saliences that determines how individuals change across time (these are the trait aspect of the model). Parenthetically in the parlance of latent curve analysis, salience is the weighting or individual change; it is analogous to a common factor in factor analysis.

The set of basis curves, Γ, describe general change across time. In general, they have the following form:

$$\Gamma = \begin{bmatrix} \gamma_{11} & & \gamma_{1r} \\ \gamma_{21} & \cdots & \gamma_{2r} \\ \gamma_{31} & & \gamma_{3r} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mr} \end{bmatrix}.$$

Notice that the elements, γ_{tj}, ($t = 1, 2, \ldots, m$; $j = 1, 2, \ldots, r$) can be fixed or parameters to be estimated, and if they are to be estimated, then identification constraints will be needed.

For this second-order model, the Means and Covariance Structure (MACS) are

$$\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(k)} = \boldsymbol{\alpha} + \Gamma\boldsymbol{\kappa}^{(k)} \text{ and } \Sigma_{\boldsymbol{\eta}}^{(k)} = \Gamma\Phi^{(k)}\Gamma' + \Psi^{(k)}, \tag{4}$$

where $\boldsymbol{\kappa}^{(k)}$ are the means for latent factors or salience weights and the covariance matrices of trait and state factors are given by $\Phi^{(k)}$ and $\Psi^{(k)}$, respectively. Further, $\Phi^{(k)}$ is usually a symmetrical matrix and $\Psi^{(k)}$ is usually assumed to be a diagonal matrix, that is, temporal or state variables are unrelated.

1.4 A Latent Curve Model with State Structure (LCM-S)

Clearly from (4) an additional structure could be imposed on either the trait, $\Gamma\Phi^{(k)}\Gamma'$, or on the state, $\Psi^{(k)}$, aspects of the model (Tisak et al. 2017), however, for this project the diagonal covariances of the state or temporal factors, $\boldsymbol{\zeta}^{(k)}$, will be generalized to include correlated state factors. Concretely, the state factors associated with each trait factor will be allowed to correlate across time:

$$\Psi_{\zeta}^{(k)} = \begin{bmatrix} \Psi_{11}^{(k)} & \cdots & \Psi_{1m}^{(k)} \\ \vdots & \ddots & \vdots \\ \Psi_{m1}^{(k)} & \cdots & \Psi_{mm}^{(k)} \end{bmatrix}, \tag{5}$$

where $\Psi_{tt'}^{(k)}$ is a diagonal matrix size r ($t = 1, 2, \ldots, m$). Note that this pattern is analogous to Spearman's Two-Factor Theory (Spearman 1904). Concretely, for each time of measurement, there will be a general state factor, which influences all the factors in the second-order model, and specific state factors, which are uncorrelated.

2 A Substantive Illustration of the Impact of State Variables in the Development of Positive Orientation

2.1 Introduction

The positive psychology movement (Seligman and Csikszentmihalyi 2000) has generated interest in the positive features of individual functioning. These findings have lead Caprara and colleagues (Caprara et al. 2010) to address what is common to self-esteem, life satisfaction, and optimism. In particular, they identified a common latent factor named positive orientation (POS). Additionally, in a longitudinal study (Alessandri et al. 2012), it was reported how POS relates to three additional constructs: (1) the quality of affective experiences (Watson et al. 1988); (2) the quality of social interactions (Hartup 1993); and (3) psychological resilience (Block and Kremen 1996). Given this longitudinal study of positive orientation, positive and negative affects, quality of social experiences, and psychological resilience across three time periods, this contribution generalizes the latent curve model with uncorrelated temporal effects to one that includes the suggested two-factor model on the state or temporal effects.

2.2 Method

2.2.1 Participants

As part of a longitudinal study the participants were male ($N = 45$) and female ($N = 81$) adolescents, who had complete data, from Genzano, Italy, a residential community near Rome. Notice that the original sample at Time 1 had 228 observations and that the attrition was mainly due to relocation from the area. For additional information on the attrition, see Alessandri et al. (2012). The first assessment (T1) was in 2000 at the age of 16; the second (T2) was in 2002 at the age of 18; and the third (T3) was in 2004 at the age of 20.

2.2.2 Measures

1. *Self-esteem.* Assessed by the 10 items of the Self-Esteem Scale (RSGE) of Rosenberg (1965). Coefficient alpha's at T1, T2, and T3 were respectively, 0.80, 0.81, and 0.83.
2. *Life satisfaction.* Assessed by the five items of the Satisfaction with Life Scale (Diener et al. 1985). Coefficient alpha's at T1, T2, and T3 were respectively, 0.90, 0.91, and 0.93.
3. *Optimism.* Assessed by the 10 items of the Life Orientation Test (SWLS) of Scheier et al. (1994). Coefficient alpha's at T1, T2, and T3 were respectively, 0.79, 0.83, and 0.81.
4. *Positive affectivity.* The Positive and Negative Affect Schedule (PANAS-P) of Watson et al. (1988). For positive affectivity, there were 10 items. Coefficient alpha's at T1, T2, and T3 were respectively, 0.81, 0.78, and 0.83.
5. *Negative affectivity.* The Positive and Negative Affect Schedule (PANAS-N) of Watson et al. (1988). For negative affectivity, there were 10 items. Coefficient alpha's at T1, T2, and T3 were respectively, 0.87, 0.80, and 0.81.
6. *Perceived quality of interpersonal relationships.* Assessed by the nine items of the Quality of Friendships Questionnaire (QDA) of Capaldi and Patterson (1989). Coefficient alpha's at T1, T2, and T3 were respectively, 0.81, 0.79, and 0.73.
7. *Psycholological resilence.* Assessed by the 14 items of the Ego Resiliency Scale (ER89) of Block and Kremen (1996). Coefficient alpha's at T1, T2, and T3 were respectively, 0.73, 0.74, and 0.73.

Note that because of the small samples, the measures were aggregated into scales and the first three scales formed the construct of positive orientation. As described in the next section, these aggregations will lead to a "measurement model", which contains both measured variable (without errors) and latent variables with measurement errors.

2.3 Statistical Analyses

2.3.1 Program and Model Fit

For estimating the hypothesized model, we utilized LISREL 8.80 (Jöreskog and Sörbom 1996). To evaluate the fit of the models, chi-square and restricted chi-square tests were used. Additionally, the root mean square of approximate (RMSEA) of Steiger and Lind (1980) was used. Browne and Cudeck's (1993) guidelines are that RMSEA < 0.05 is a close fit, and RMSEA < 0.08 is a reasonable or near fit, but RMSEA > 0.10 is a poor fit.

2.3.2 Structural Equation Models (SEM)

Two major Structural Equation Models (SEM) were evaluated:

1. A Longitudinal Factor Analytic Model (FAM) with the corresponding modeling equation (1) and MACS (2). Specifically, on this first-order model both invariance and stationarity conditions were imposed on the intercepts and slopes. There were seven measured variables (scales) that were assessed at the three time points; hence, there were 21 variables. The first-order construct of positive orientation was obtained from self-esteem, life satisfaction, and optimism. The remaining four variables were treated without measurement error, that is unstructured except for the invariant and stationary intercepts and slopes. The covariance matrix of the unique factors was zero, except for self-esteem, life satisfaction, and optimism. Each of these variables were allowed to covary with themselves across time, and they had positive variances. This model reduced the 21 measured variables to 15 latent variables.
2. A Latent Curve Model (LCM) with the corresponding modeling equation (3) and MACS (4). For each of the five longitudinal variables (positive orientation, positive affectivity, negative affectivity, quality of relationships, ego resiliency), the temporal effects, $\boldsymbol{\alpha}$, were set to zero, and a single latent curve was used. Concretely, the 15 × 5 matrix of basis curves (Γ) is

$$\Gamma = \begin{bmatrix} 1 & 0 & & 0 \\ \gamma_{21} & 0 & & 0 \\ \gamma_{31} & 0 & & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & \gamma_{22} & & 0 \\ 0 & \gamma_{32} & & 0 \\ & \vdots & \ddots & \vdots \\ 0 & 0 & & 1 \\ 0 & 0 & \cdots & \gamma_{25} \\ 0 & 0 & & \gamma_{35} \end{bmatrix}.$$

Further, the covariance matrix of the temporal (state) variables, $\Psi^{(k)}$, was as usual constrained to be a diagonal matrix. Lastly, one of the major interests in this study was the covariance matrix of the latent factor, $\Phi^{(k)}$, because it gives the relationships among the individual saliences.

Based on the findings of the two previous models, a modified third model was explored. This model with correlated temporal or state variables demonstrates the point of the manuscript.

Table 1 A summary of the fit indices for the different models assessed

Model	Df	Chi-Square	RMSEA	p-value
FAM	194	255.521	0.0358	0.00201
LCM-S	344	462.302	0.0459	0.00002
LCM	404	659.149	0.0790	0.00000

2.4 Results

The simple modeling fitting results for the two models (FAM and LCM) are given in the first and third rows of Table 1. Notice that FAM has a very acceptable RMSEA of 0.0358, while LCM only has a marginally acceptable RMSEA of 0.0790. Further using a restricted chi-square test ($\chi^2_R(210) = 403.628$, $p = 0.00000$), the reduced LCM was significantly different from the general FAM.

Given these results, how should one proceed? One could report the Latent Curve Model, or one could try to generalize it by modifying the number of basis curves (the trait aspect) or by modifying the covariance matrix of the temporal or state factors. Concretely, since $\Sigma^{(k)}_{\boldsymbol{\eta}} = \Gamma\Phi^{(k)}\Gamma' + \Psi^{(k)}$, $\Gamma\Phi^{(k)}\Gamma'$ and $\Psi^{(k)}$ represents the trait and the state aspect of the model.

If one changes the trait aspect, that is, the number of basis curves, there are numerous combinations, which could lead to adding "nuisance parameters" to the model. Hence, one avenue to explore is to add structure to the previously diagonal matrix, $\Psi^{(k)}$. Using the Two-Factor State Theory, $\Psi^{(k)}$, has the form depicted in (5). The fit indices of this model, which includes correlated temporal effects, are given in the second row of Table 1 (LCM-S Model). Notice that LCM-S has a very respectable RMSEA of 0.0459. Further, when one compares the more general LCM-S to the more specific LCM, the restricted chi-square (LCM versus LCM-S) equals $\chi^2_R(60) = 196.847$, $p = 0.00000$. Thus, correlated temporal factors should not be ignored. In conclusion, there is a model (LCA-S) between the general (FAM) and the reduced (LCM) models, which is an improvement in terms of the fit indices over the (LCM).

Notice that the degrees for freedom for these models represent the difference between the observed data means (21) and the data covariances (231) for the two genders for a total of 504 and the number of parameters estimated in each model. To illustrate, the FAM has 310 parameters that are estimated; so the degrees of freedom for this model is 504 – 310 or 194.

3 Discussion

Earlier it was pointed out that the notions of traits and states can be an important and useful classification in longitudinal studies. These two entities are expressed in (3) and the corresponding means and covariance structure in (4). Concretely, traits may be expressed by the formulation: $\Gamma\boldsymbol{\xi}^{(k)}$, and states or temporal effects may be represented in the random vector, $\boldsymbol{\zeta}^{(k)}$.

Psychological constructs, like positive orientation, are not directly observable. Instead, constructs are latent entities introduced to explain the recurrent organization of an individual's internal states, such as feelings and emotions, as well to be used as causes of human behaviors (Borsboom et al. 2003). Researchers often studies those constructs at different timescales, depending on whether they are interested, for example, in the longitudinal development of individual's traits or aptitudes, or in the online tracking of individuals daily functioning. Whatever the timing of the study, psychological constructs usually reveal both trait and state variance, that researchers need to isolate and separately investigate. Whereas constructs characterized by trait variance only are rare, pure state-like constructs are often the exception. This contribution moves in a different direction, that is, states. Fit indices are a very practical justification for enhancing the structure on states. However, a more important justification for a more developed state model is an increased accuracy of the situation under study.

Additionally, whereas temporal consistency is one of the more distinctive characteristics of traits, states may often reveal a significant degree of continuity. Carry-over effects, denoting the tendency for a previous state to spill over time into the following state are often observed and often expected on a theoretical basis. For example, emotional states often display a high temporal continuity, denoting a tendency of the emotional dynamics to slow down until a state called emotional inertia (Kuppens et al. 2010).

To account for a significant continuity of states, researchers need tools able to allow the modeling of temporal variances in psychological attributes, such as the Latent Curve Model with state structure. Introduced as an expansion of the Latent Curves-Latent State-Trait modeling framework, the LCA-S is sensible to the continuity of states, allowing their inclusion in the model as covariances among subsequent states. Results presented in this paper point to this model as an interesting alternative to the general (FAM) and the reduced (LCM) models, which ensure a gain in terms of fit indices over the (LCM).

A longitudinal study on the development of position orientation illustrates the importance of states in one's statistical model. In Table 1, the Latent Curve Model (LCM) has a questionable mode fit index (RMSEA) of 0.0790. Further note that this LCM has a minimum formulation on the temporal effects, that is, these effects exist, but they do not correlate. On the other hand, if one generalizes the LCM to include structured states (LCM-S), the index of RMSEA is greatly improved. Lastly, when one compares the more general LCM-S to the more specific LCM, the restricted chi-square (LCM versus LCM-S) is significant. Thus, correlated temporal factors are statistically significant.

We surmise that in many situations, the LCA-S model may represent a more realistic alternative to simple LCM models. For example, we expect the LCA-S model to be of great value in allowing the modeling of intensive short-term studies (such as daily studies, weekly studies) whereas a significant continuity in trait variance can be expected. Of course, the results presented in this paper are preliminary, and more work is recommended to examine the stability of the LCA-S model, under different empirical data conditions, and different variance/covariance structures.

Moreover, it is likely that the benefit introduced by the use of the LCA-S model are directly correlated with the length of the temporal lag, being probably higher for lags introducing more temporal variance in psychological constructs. In conclusion, we recommend to routinely consider the LCA-S as an alternative to simple LCM models, most of all, in all those conditions where including more common factors lacks theoretical justification and thus risks overfitting the model without any practical contribution to the understanding of the phenomenon under study.

References

Alessandri, G., Caprara, G. V., & Tisak, J. (2012). A unified latent curve, latent state-trait analysis of the developmental trajectories and correlates of positive orientation. *Multivariate Behavioral Research, 47,* 3451–3468.

Block, J. H., & Kremen, A. M. (1996). IQ and ego resiliency: Conceptual and empirical connections and separateness. *Journal of Personality and Social Psychology, 70,* 349–361.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Capaldi, D. M., & Patterson, G. R. (1989). *Psychometric properties of fourteen latent constructs from the Oregon Youth Study*. New York, NY: Springer.

Caprara, G. V., Steca, P., Alessandri, G., Abela, J. R. Z., & McWinnie, C. M. (2010). Positive orientation: Explorations of what is common to life satisfaction, self-esteem, and optimism. *Epidemiologia e Psichiatria Sociale, 18,* 63–71.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49,* 71–75.

Hartup, W. W. (1993). Adolescents and their friends. In B. Laursen (Ed.), *New directions for child development: No. 60. Close friendships in adolescence* (pp. 2–22). San Francisco, CA: Jossey-Bass.

Jöreskog & Sörbom. (1996). *LISREL 8*. Chicago, IL: Scientific Software International.

Kuppens, P., Allen, N. B., & Sheeber, L. (2010). Emotional inertia and psychological maladjustment. *Psychological Science, 2,* 984–991.

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

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55,* 107–122.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67,* 1063–1087.

Seligman, M. E. P., & Csikszentmaihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist, 55,* 5–14.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Steiger, J. H., & Lind, J. C. (1980). Statistically-based tests for the number of common factors. *Multivariate Behavioral Research, 25,* 173–180.

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Tisak, J., Alessandri, G., & Tisak, M. S. (July, 2017). *Several approaches to the modeling of ephemeral effects.* Paper presented at the 82nd Annual Meeting of the Psychometric Society, Zürich, Switzerland.

Tisak, J., & Meredith, W. (1989). Exploratory longitudinal factor analysis in multiple populations. *Psychometrika, 54,* 261–281.

Tisak, J., & Tisak, M. S. (2000). Permanency and ephemerality of psychological measures with application to organizational commitment. *Psychological Methods, 5,* 175–198.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology, 54,* 1063–1070.

SPARK: A New Clustering Algorithm for Obtaining Sparse and Interpretable Centroids

Naoto Yamashita and Kohei Adachi

Abstract k-means clustering is one of the popular procedures for multivariate analysis in which observations are classified into a reduced number of clusters. The resulting centroid matrix is refereed to capture variables which characterize clusters, but between-clusters contrasts in the centroid matrix are not always clear and thus difficult to interpret. In this research, we address the problem in interpretation and propose a new procedure of k-means clustering which produces a sparse and thus interpretable centroid matrix. The proposed procedure is called SPARK. In SPARK, the sparseness of the centroid matrix is constrained and therefore it contains a number of exact zero elements. Because of this, the contrasts between-clusters are highlighted and it allows us to interpret clusters easier in comparison with the standard k-means clustering. A sparsity selection procedure for determining the optimal sparsity of the centroid with reduced computational load is also proposed. Behaviors of the proposed procedure are evaluated by two real data examples, and the results indicate that SPARK performs well for dealing with real world problems.

Keywords k-means clustering · Sparse estimation · Interpretability

1 Introduction

k-means clustering, known as a non-hierarchical clustering procedure, is widely used for extracting the homogeneity of observations, by assigning them into a small number of clusters. Let $\mathbf{X}$ be an n-obserbations × p-variables matrix, and the k-means clustering is formulated as a minization of the least squares loss function defined as

$$f(\mathbf{M}, \mathbf{Y}) = \sum_{i,l} m_{il} ||\mathbf{x}'_{(i)} - \mathbf{y}_l||^2 = ||\mathbf{X} - \mathbf{M}\mathbf{Y}'||^2, \tag{1}$$

N. Yamashita (✉) · K. Adachi
Graduate School of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita, Osaka, Japan
e-mail: nyamashita@hus.osaka-u.ac.jp; nyamashita.hus@osaka-u.ac.jp

K. Adachi
e-mail: adachi@hus.osaka-u.ac.jp

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_34

where $\mathbf{M} = \{m_{il}\}$ is an n-observations × p-variables membership matrix and $\mathbf{Y} = \{y_{jl}\}$ is a p-variables × k-clusters centroid matrix. $\mathbf{x}_{(i)}$ and $\mathbf{y}_l$ denote the ith row vector and the lth column vector of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

The centroid matrix is refereed for interpreting what variables characterize the clusters, and the within- and between-clusters contrasts in the centroid matrix are of help for the interpretation. These contrasts, however, are not always clearly observed and therefore the interpretation is difficult, as exemplified in Sect. 4. A typical strategy to discriminate the clusters is to replace the elements close to zero in the centroid matrix with zeros, by a certain threshold. It is not recommended, however, in that the threshold totally depends on users' decision, and it can spoil the reliability of the interpretation and the following decisions.

In this article, considering the above problem in interpretability of the resulting centroid matrix, we propose a new algorithm for clustering which produces an easily interpreted centroid matrix. We call this algorithm SPARK (abbreviation of Sparse k-means). In SPARK, the resulting centroid matrix is sparse in that it contains a number of entries exactly equal to zero. The contrasts of the clusters are therefore emphasized, without any subjective threshold, which facilitates the easier and more coherent interpretation than the existing procedures. Such a centroid matrix is obtained by minimizing (1) subject to the constraint that $\mathbf{Y}$ has a specific number of zero elements, namely,

$$Sp(\mathbf{Y}) = r \tag{2}$$

where $Sp(\mathbf{Y})$ is the number of zero in $\mathbf{Y}$. The positive integer r is specified beforehand.

1.1 Related Procedure

Sun et al. (2012) proposed regularized k-means clustering for obtaining such sparse centroid matrix, which is similar to the proposed method. It is formulated as a minimization of (1) subject to the row-wise constraint on $\mathbf{Y}$

$$||\mathbf{y}_{(j)}|| \le \lambda_j \quad (j = 1, \ldots, p) \tag{3}$$

where $||\mathbf{y}_{(j)}||$ is an L_1-norm of the $\mathbf{Y}$'s jth row vector $\mathbf{y}_{(j)}$ and a tuning parameters λ_j $(j = 1, \ldots, p)$ control the resulting sparsity of $\mathbf{Y}$. It therefore contains a number of zero elements, since the L_1-norm of rows of $\mathbf{Y}$ is constrained to be less than $\lambda_1, \ldots, \lambda_p$. This minimization is equivalent to the minimization of the following function;

$$f(\mathbf{M}, \mathbf{Y}) + \sum_{j}^{p} \lambda_j ||\mathbf{y}_{(j)}||. \tag{4}$$

We call this approach as a *penalty approach*, in that it adds the penalty function $\sum_j^p \lambda_j ||\mathbf{y}_{(j)}||$ to the original loss function (1). The tuning parameters take any positive integer, which are commonly determined by cross-validation. Similar approaches can be found in Witten and Tibshirani (2010) and Hastie et al. (2015). Penalty approach is originally proposed for avoiding over-fitting in clustering. Generalizability, however, does not always results in the easier interpretability of the clusters, which we focus on in this article. The proposed procedure directly controls the number of zero elements r in the centroid matrix within a restricted range, without introducing tuning parameters as in the penalty approaches. Within- and between-contrasts in the centroid matrix are therefore highlighted, and it allows users to find what variables manifest the clusters easily. It should be noted that controlling r cannot consider all possible values of $\lambda_1, \dots, \lambda_p$. For interpretation of clusters, however, inspecting all possible λs is not necessary, and sparseness of $\mathbf{Y}$ can be determined by how many elements in $\mathbf{Y}$ are zero and ignorable.

2 Algorithm

The proposed procedure SPARK is formulated as the following constrained minimization problem;

$$\min_{\mathbf{M},\mathbf{Y}} f(\mathbf{M}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{M}\mathbf{Y}'||^2 \tag{5}$$

subject to the sparsity constraint (2) and the membership constraint is imposed on $\mathbf{M}$ such that

$$m_{il} \in \{0, 1\} \text{ and } \sum_l m_{il} = 1. \tag{6}$$

The parameter matrices are alternately and iteratively updated in the M-step and Y-step, respectively, starting from multiple sets of initial values in order to avoid accepting a local minimum as the final solution. In these steps, the current parameter matrix is replaced by the one minimizing (1) keeping the other parameter matrix fixed. The update formulae used in the M-step and Y-step are presented as follows.

M-step The minimization of $f(\mathbf{M}, \mathbf{Y})$ with fixed $\mathbf{Y}$ subject to (6) is achieved by the k-means algorithm with the fixed centroid (MacQueen 1967). Therefore, the optimal $\mathbf{M} = \{m_{il}\}$ is obtained by

$$m_{il} = \begin{cases} 1 & (l = \arg\min\limits_l f(\mathbf{M}, \mathbf{Y})) \\ 0 & (\text{otherwise}) \end{cases}, \tag{7}$$

for $i = 1, \dots, n$.

Y-step Using the matrix $\mathbf{C} = \mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}$, (1) is rewritten as

$$\begin{aligned} f(\mathbf{M},\mathbf{Y}) &= ||\mathbf{X}-\mathbf{M}\mathbf{Y}'||^2 \\ &= ||\mathbf{X}-\mathbf{M}\mathbf{C}'+\mathbf{M}\mathbf{C}'-\mathbf{M}\mathbf{Y}'||^2 \\ &= ||\mathbf{X}-\mathbf{M}\mathbf{C}'||^2+||\mathbf{D}^{1/2}(\mathbf{C}-\mathbf{Y})||^2 \\ &\qquad -\mathrm{tr}(\mathbf{X}-\mathbf{M}\mathbf{C}')'(\mathbf{M}\mathbf{C}'-\mathbf{M}\mathbf{Y}'). \end{aligned} \tag{8}$$

where $\mathbf{D} = \mathrm{diag}\{d_{11}, \ldots, d_{ll}, \ldots, d_{kk}\}$ denotes the $k \times k$ diagonal matrix whose lth diagonal element is equal to the number of the observations classified into the lth cluster ($l = 1, \ldots, k$). The third term is proved to be zero as follows;

$$\begin{aligned} &\mathrm{tr}(\mathbf{X}-\mathbf{M}\mathbf{C}')'(\mathbf{M}\mathbf{C}'-\mathbf{M}\mathbf{Y}') \\ &= \mathrm{tr}\mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X} - \mathrm{tr}\mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X} - \mathrm{tr}\mathbf{X}'\mathbf{M}'\mathbf{M} + \mathrm{tr}\mathbf{X}'\mathbf{M}'\mathbf{M} \\ &= 0. \end{aligned} \tag{9}$$

Therefore, minimizing the second term in (8), $g(\mathbf{Y}) = ||\mathbf{D}^{1/2}(\mathbf{C}-\mathbf{Y})||^2$, is equivalent to the minimization of $f(\mathbf{M}, \mathbf{Y})$ with respect to $\mathbf{Y}$. Further, $g(\mathbf{Y})$ is rewritten as

$$g(\mathbf{Y}) = \sum_{(j,l)\in Z} d_{ll}^{1/2} c_{jl}^2 + \sum_{(j,l)\in Z^{\perp}} d_{ll}^{1/2} (c_{jl} - y_{jl})^2 \geq \sum_{(j,l)\in Z} d_{ll}^{1/2} c_{jl}^2 \tag{10}$$

where the Z denotes r pairs of indices (j, l)s indicating the locations of y_{jl}s to be zero. The last equality holds when the second term in (10) is equal to zero, that is, when y_{jl} with $(j, l) \in Z^{\perp}$ is taken equal to the corresponding c_{jl}. In addition, the limit $\sum_{(j,l)\in Z} d_{ll}^{1/2} c_{jl}^2$ is minimal when Z is composed of the indices of the r smallest c_{jl}^2s among all squared elements in $\mathbf{C}$. Therefore, $\mathbf{Y}$ that minimizes $g(\mathbf{Y})$ is obtained as

$$y_{jl} = \begin{cases} 0 & (iff\, c_{jl}^2 \leq c_{[r]}^2) \\ c_{jl} & (otherwise) \end{cases} \tag{11}$$

for $l = 1, \ldots, k$ and $j = 1, \ldots, p$, where $c_{[r]}^2$ denotes the rth smallest value among all $c_{jl}^2 s$. The update formulae (7) and (11) are used in the M-step and Y-step, respectively, and it is guaranteed that function value of $f(\mathbf{M}, \mathbf{Y})$ monotonically decreases in each of these steps. As presented in this section, $\mathbf{M}$ and $\mathbf{Y}$ are alternately updated until the convergence is reached. In the following real data examples, we used 100 different initial values for $\mathbf{M}$ and $\mathbf{Y}$.

3 Sparsity Selection Based on Information Criteria

In the proposed procedure, the number of zeros in $\mathbf{Y}$ has to be specified as a positive integer r in (2). In this article, the minimum and maximum of r, r_{min}, r_{max}, are defined as

$$r_{min} = 1, \;\; r_{max} = p \times (k - 1) \tag{12}$$

considering that $\mathbf{Y}$ has p non-zero elements when $\mathbf{Y}$ has a perfect cluster structure; each variable is associated with only one cluster. Selecting the number of zero elements in $\mathbf{Y}$ can be considered as a model selection problem, since this selection partially specifies the model part of $\mathbf{MY}'$ fitted to $\mathbf{X}$. In this respect, the information criterion such as AIC and BIC is suitable for specifying r, which controls how sparse the model is to be fitted to the data. In this section, we propose two criteria in order to select the "best" r among the interval $[r_{min}, r_{max}]$.

Here, let $\mathbf{E} = \{e_{ij}\}$ be the matrix of errors defined as $\mathbf{E} = \mathbf{X} - \mathbf{MY}'$. Under the assumption that $\mathbf{X}$ is generated by $\mathbf{X} = \mathbf{MY}' + \mathbf{E}$ with e_{ij} distributed independently and identically according to $N(0, \sigma^2)$ for all is and js with a specific error variance σ^2, it can be shown that the least squares estimation and maximum likelihood estimation in SPARK are equivalent. The log-likelihood function to be maximized in the ML estimation is

$$l(\mathbf{M}, \mathbf{Y}) = -\frac{np}{2} \log ||\mathbf{X} - \mathbf{MY}'||^2 \tag{13}$$

including $f(\mathbf{M}, \mathbf{Y})$ to be minimized in the least square estimation. With an arbitrary r, the maximum of $l(\mathbf{M}, \mathbf{Y})$ is attained as

$$l(\mathbf{M}, \mathbf{Y}) \leq -\frac{np}{2} \log f_{min}(r). \tag{14}$$

where $f_{min}(r)$ denotes the attained function value of (1). By (14), the information criteria $AIC(r)$ and $BIC(r)$ with the specific r are obtained by

$$AIC(r) = np \times \log f_{min}(r) + 2v(r) \tag{15}$$

$$BIC(r) = np \times \log f_{min}(r) + \log(np) \times v(r) \tag{16}$$

where $v(r)$ denotes the number of parameter to be estimated with a certain r;

$$v(r) = n + kp - r, \tag{17}$$

Therefore, r can be determined by $r = \underset{r_{min} \leq r \leq r_{max}}{\arg\min}\, AIC(r)$ or $BIC(r)$ in terms of minimizing the model selection criteria. This approach is considered to be computationally inefficient, however, as of 100 run of SPARK are required, in order to avoid a local minimum, for each of all possible rs. When $\mathbf{X}$ is of a large size, ($\mathbf{X}$ contains many observations and variables) the resulting centroid matrix is also of a large size, and thus higher computational cost is required for each run.

In order to find such r with lower computational cost, we propose the following algorithm.

Step 1. Set $S_{initial}$ and $S_{decrease}$ to an integer within the range [0, 1]. Set $r_t = S_{initial} \times r_{max}$

Step 2. Repeat Step 3 to Step 4 while $S > 1$.

Step 3. (*Forward search*) Repeat (a) to (c).

(a) Set $r = r_t$ and compute

$$\Delta AIC(r) = AIC(r+1) - AIC(r) \tag{18}$$

or

$$\Delta BIC(r) = BIC(r+1) - BIC(r) \tag{19}$$

(b) If $\Delta AIC(r)$ or $\Delta BIC(r)$ is smaller than 0, set $r_t = r_t + S$ and go back to 2. Otherwise proceed to (c).

(c) Set $S = S \times S_{decrease}$ and proceed to the *backward search.*

Step 4. (*Backward search*) Repeat (a) to (c).

(a) Set $r = r_t$ and compute $\Delta AIC(r)$ or $\Delta BIC(r)$.

(b) If $\Delta AIC(r)$ or $\Delta BIC(r)$ is greater than 0, set $r_t = r_t - S$ and go back to 4. Otherwise proceed to (c).

(c) Set $S = S \times S_{decrease}$ and proceed to the *forward search.*

Step 5. If the previous step is *Forward search*, repeat *barkward search* with $S =$ 1 until $\Delta AIC(r)$ or $\Delta BIC(r)$ is positive; otherwise repeat *Forward search* $\Delta AIC(r)$ or $\Delta BIC(r)$ is negative.

The above algorithm seeks r which minimizes $AIC(r)$ or $BIC(r)$ within the range $[r_{min}, r_{max}]$ by repeating the forward and backward search and reducing the step size S at each step of the iteration, starting from the initial step size $r_{max} \times S_{initial}$. The rate of decrement of the step size is controlled by $S_{decrease}$. The total computational cost is therefore dramatically reduced compared with applying SPARK for computing $AIC(r)$ or $BIC(r)$ for all rs. In the following simulation and the real data examples, we set $S_{initial} = 0.9$ and $S_{decrease} = 0.7$ which is empirically confirmed to be well-performed.

4 Real Data Examples

In this section, we demonstrate that SPARK extracts the sparse centroids underlying the dataset and facilitates interpretation of the centroid, with keeping the correctness of classification.

4.1 Example 1: Fisher's Iris Data

In the first example, SPARK was applied to Fisher's Iris data, where 150 samples, which are originally sampled from three species, were measured with respect to four variables. In order to find the optimal sparsity, the sparsity selection procedure based on BIC was used. It suggested that $r = 2$ was the best, and we also applied the standard k-means clustering to Iris data for comparison.

The estimated centroids are shown in Table 1 as a heatmap. As found in Table 1, the contrast between the first (C1) and the second (C2) clusters can be seen in Sepal.Length and Sepal.Width. In addition, C2 is different from the rest of clusters with respect to Sepal.Width The contingency table of two partitions, the species of samples and the estimated membership, for SPARK and the one for k-means, are shown in Table 2. It can be seen that the estimated memberships correspond to the species, in that $(49 + 37 + 42)/150 = 85.3\%$ of the observations are correctly classified, while $(50 + 39 + 36)/150 = 89.2\%$ in the k-means. These results indicate that SPARK appropriately produces sparser and thus easy-to-interpret centroid matrix in comparison with the exiting method, keeping the accuracy of classification.

Table 1 Estimated centroid matrices by SPARK for Fisher's iris dataset with $r = 2$ and k-means clustering

		Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
SPARK	C1	1.065		0.966	0.999
	C2		−0.928	0.322	0.236
	C3	−1.011	0.850	−1.301	−1.251
k-means	C1	1.132	0.088	0.993	1.014
	C2	0.050	−0.880	0.347	0.281
	C3	−1.011	0.850	−1.301	−1.251

Table 2 Contingency table for species versus the estimated partitions by SPARK and k-means

	SPARK			*k*-means		
	C1	C2	C3	C1	C2	C3
Setosa	49	0	0	50	0	0
Versicolor	1	37	8	0	39	14
Virginica	0	13	42	0	11	36

4.2 Example 2: Vicon Physical Action Dataset

The second example is Vicon Physical Action Dataset (Lichman 2013). A subject's walking was recorded by the 3-axis motion sensors attached to the subject's right and left wrists, elbows, knees, and ankles. The activity was recorded for approximately 8000 ms with the frequency of 20 Hz. Therefore we have 24 (*x*-/*y*-/*z*-axis sensors of right and left wrists, elbows, knees and ankles) × 173 (time elapsed) data matrix. *k*-means clustering is applied to the data matrix and the resulting centroid matrix is shown in Fig. 1 as a heatmap. The number of clusters is set to 5 which explains 75% of the total variance of the dataset.

We can interpret the estimated five clusters by referring the 173 × 5 centroid matrix as follows. For example, the first (C1) and the second (C2) clusters are well discriminated against the others; the first cluster is characterized by the lower output value in the middle phase of records (around 2000–6000 ms) and the higher value in the latter phase (around 6000–8500 ms), while this variation in the sensor outputs is shifted for 2000 ms earlier in the second cluster. The third (C3), fourth (C4) and fifth (C5) clusters are, however, hard to be discriminated mutually, in that the time evolutions of values are similar to each other especially in the early phase.

Before applying SPARK to the dataset, the sparsity selection procedures were applied. The AIC- and BIC-based procedures suggested that $r = 332$ and $r = 461$ were the best, respectively. We therefore determined to set $r = 461$ in order to obtain the sparser centroid matrix. This means that approximately 53.3% of the all elements of the centroid matrix were estimated as zero. The number of clusters was set at 5, as in the example of the *k*-means clustering in Sect. 1.

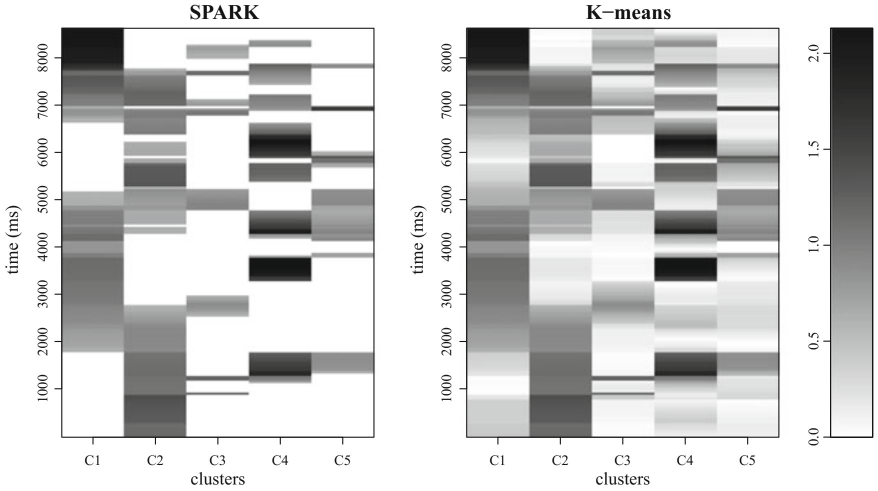


Fig. 1 Estimated centroid matrix by SPARK with $r = 461$ and *k*-means (absolute transformed) for Vicon Physical Action Dataset

Table 3 Estimated membership of 24 sensors; x/y/z-axis senror on the right (R) and left (L) wrist, elbow, knee and anckle

	wrist		elbow		knee		ankle	
	L	R	L	R	L	R	L	R
C1		z		z				
C2						z		z
C3	y/z	y	y/z	y	y	y	y	y
C4	x	x	x	x	x	x	x	x
C5					z		z	

The resulting centroid matrix is represented as a heatmap in Fig. 1. The elements estimated as zero are colored in white. It can be seen that, compared with the standard *k*-means clustering, the estimated centroid is sparse enough and the contrasts between clusters are clearer than in the *k*-means clustering solution. Based on the sparse centroid, each cluster can be interpreted as follows; the sensors classified into the first cluster show the lower values from approximately 2000–5000 ms and the higher values from 6500 ms to the end of recording, and this variation of sensor outputs is earlier by 1500 ms in the second cluster. The third cluster is characterized by the lower values around 6000 ms, which makes the cluster different from the other clusters. In the fourth cluster, the lower values and the higher values alternately appear except in the early phase of recording, while the sensor outputs are almost stable in the fifth cluster.

The centroids obtained by *k*-means are less sparse than the centroids for SPARK and the characteristics of clusters are unclear. As a measure of interpretability, Lorenzo-Seva (2003) proposed the index of simplicity called LS index in the context of factor analysis. The LS index ranges from 0 (least simple) to 1 (most simple) and the values LS index for the centroid matrices were 0.313 in the *k*-means and 0.590 in the SPARK, which indicates the sparsely estimated centroids are more simple and thus more interpretable compared with the existing method.

The sensor classified into each cluster are shown in Table 3. The first cluster is composed of the z-axis sensors on the right arm, while those on the left arm are classified into the third cluster. It indicates that the subject’s horizontal movement in the left and right arms are expressed in the first and the second clusters. The third cluster is composed of 10 sensors, the y/z-axis sensors on the left arm and the y-axis sensors on the leg. The x-axis sensor on all parts are classified into the fourth cluster, and refereeing the sparse centroids in Fig. 1 therefore indicate that the clear difference between the x-axis and the y-axis movement is observed around 6000 ms.

5 Concluding Remarks

In this article, we proposed a new procedure of clustering called SPARK, which produces a sparse centroid matrix The interpretation of the centroid matrix is easier compared with the ordinal k-means clustering by the sparsity constraint imposed on the centroid matrix. It is also possible to obtain such sparse centroid by adding a penalty term to the loss function of k-means clustering, as proposed by some authors. These procedures mainly aims to improve the robustness of clustering through the sparse estimation of centroid matrix. In SPARK, on the other hand, we rather focus on the interpretability of the resulting centroid matrix than robustness. The sparseness of the centroid matrix is therefore controlled by the number of zero elements in the centroid matrix, which is closely related to its interpretation. The results of the two real data examples indicate that the estimated sparse centroids surely facilitates to capture the characteristic of the clusters.

References

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity*. CRC press.

Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml

Lorenzo-Seva, U. (2003). A factor simplicity index. *Psychometrika*, *68*(1), 49–60.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297.

Sun, W., Wang, J., & Fang, Y. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, *6*, 148–167.

Witten, D., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, *105*(490), 713–726.